Set-Valued and Variational Analysis
https://doi.org/10.1007/s11228-021-00591-3

Variational Analysis Perspective on Linear Convergence of Some First Order Methods for Nonsmooth Convex Optimization Problems

Jane J. Ye[1] **· Xiaoming Yuan**[2] **· Shangzhi Zeng**[2] **· Jin Zhang**[3]

Received: 27 April 2020 / Accepted: 11 May 2021 / Published online: 29 June 2021

Abstract

We study linear convergence of some first-order methods such as the proximal gradient method (PGM), the proximal alternating linearized minimization (PALM) algorithm and the randomized block coordinate proximal gradient method (R-BCPGM) for minimizing the sum of a smooth convex function and a nonsmooth convex function. We introduce a new analytic framework based on the error bound/calmness/metric subregularity/bounded metric subregularity. This variational analysis perspective enables us to provide some concrete sufficient conditions for linear convergence and applicable approaches for calculating linear convergence rates of these first-order methods for a class of structured convex problems. In particular, for the LASSO, the fused LASSO and the group LASSO, these conditions are satisfied automatically, and the modulus for the calmness/metric subregularity is computable. Consequently, the linear convergence of the first-order methods for these important applications is automatically guaranteed and the convergence rates can be calculated. The new perspective enables us to improve some existing results and obtain novel results unknown in the literature. Particularly, we improve the result on the linear convergence of the PGM and PALM for the structured convex problem with a computable error bound estimation. Also for the R-BCPGM for the structured convex problem, we prove that the linear convergence is ensured when the nonsmooth part of the objective function is the group LASSO regularizer.

Keywords Metric subregularity · Calmness · Proximal gradient method · Proximal alternating linearized minimization · Randomized block coordinate proximal gradient method · Linear convergence · Variational analysis · Machine learning · Statistics

Mathematics Subject Classification 2010 90C25 · 90C52 · 49J52 · 49J53

The research was partially supported by NSERC, the general research fund from the Research Grants Council of Hong Kong 12302318, National Science Foundation of China 11971220, Natural Science Foundation of Guangdong Province 2019A1515011152

✉ Xiaoming Yuan
xmyuan@hku.hk

Extended author information available on the last page of the article.

1 Introduction

In recent years, there has been a revived interest in studying convex optimization in the form

$$\min_x \; F(x) := f(x) + g(x). \tag{1}$$

These kinds of optimization problems may originate from data fitting models in machine learning, signal processing, and statistics where f is a loss function and g is a regularizer. Throughout the paper, our results are given in n-dimensional Euclidean space under the following blanket assumptions.

Assumption 1 *$f, g : \mathbb{R}^n \to (-\infty, \infty]$ are two proper lower semi-continuous (lsc) convex functions.*

(i) *Function f has an effective domain $\mathrm{dom} f := \{x \mid f(x) < \infty\}$ assumed to be open and is continuously differentiable with Lipschitz continuous gradient on a closed set $\Omega \supseteq \mathrm{dom} f \cap \mathrm{dom} g$, and g is continuous on $\mathrm{dom} g$;*
(ii) *the Lipschitz constant of the gradient $\nabla f(x)$ is $L > 0$ and the Lipschitz constant of the $\nabla_i f(x) := \nabla_{x_i} f(x)$ is $L_i > 0$;*
(iii) *problem (1) has a nonempty solution set denoted by $\mathcal{X} := \arg\min_{x \in \mathbb{R}^n} F(x)$ with the optimal value F^*.*

As the size of the problem (1) in applications increases, first order methods such as the proximal gradient method (PGM) (see e.g., [42, 44]) have received more attention. Denote the proximal operator associated with g by

$$\mathrm{Prox}_g^{\gamma}(a) := \arg\min_{x \in \mathbb{R}^n} \left\{ g(x) + \frac{1}{2\gamma} \|x - a\|^2 \right\},$$

where $\gamma > 0$. The PGM for solving problem (1) has the following iterative scheme:

Algorithm 1 Proximal gradient method.

1: Choose $x^0 \in \mathbb{R}^n$
2: **for** $k = 0, 1, 2, \cdots$ **do**

$$x^{k+1} = \mathrm{Prox}_g^{\gamma}\left(x^k - \gamma \nabla f(x^k)\right).$$

3: **end for**

When g is an indicator function, the PGM reduces to the projected gradient method (see, e.g., [44, 47]); when $f \equiv 0$, it reduces to the proximal point algorithm (see, e.g., [35]) and when $g \equiv 0$ it reduces to the standard gradient descent method. It is known that for problem (1), the PGM converges at least with the sublinear rate of $O(1/k)$ where k is the number of iterations; see e.g., [7, 42, 59]. However, it has been observed numerically that very often for problem (1) with some structures, the PGM converges at a faster rate than that suggested by the theory; see, e.g., [1, 63]. In particular, when f is strongly convex and g is convex, [43, 55] have proved the global linear convergence of the PGM with respect to the sequence of objective function values.

In many big data applications arising from, e.g., network control [37], or machine learning [5, 8, 12], the regularizer g in problems (1) may have block separable structures, i.e.,

$g(x) := \sum_{i=1}^{N} g_i(x_i)$ with $x_i \in \mathbb{R}^{m_i}$, $g_i : \mathbb{R}^{m_i} \to (-\infty, \infty]$ and $n = \sum_{i=1}^{N} m_i$; see, e.g., [38]. In this setting, (1) can be specified as

$$\min_x \; F(x) := f(x_1, x_2, \ldots, x_N) + \sum_{i=1}^{N} g_i(x_i). \tag{2}$$

Numerous experiments have demonstrated that the block coordinate descent schemes are very powerful for solving huge scale instances of model (2). The coordinate descent algorithm is based on the idea that the minimization of a multivariable function can be achieved by minimizing it along one direction at a time, i.e., solving univariate (or at least much simpler) optimization problems in a loop. The reasoning behind this is that coordinate updates for problems with a large number of variables are much simpler than computing a full update, requiring less memory and computing power. Coordinate descent methods can be divided into two main categories: deterministic and random methods.

The simplest case of a deterministic coordinate descent algorithm is the proximal alternating linearized minimization (PALM) algorithm, where the (block) coordinates to be updated at each iteration are chosen in a cyclic fashion. The PALM for solving (2) reads as:

Algorithm 2 Proximal alternating linearized minimization.

1: Choose $x^0 \in \mathbb{R}^n$
2: **for** $k = 0, 1, 2, \cdots$ **do**
3: **for** $i \in \{1, 2, \ldots, N\}$ **do**

$x_i^{k+1} = \operatorname*{argmin}_{x_i \in \mathbb{R}^{m_i}} \left\{ \langle \nabla_i f(x^{k,i-1}), x_i - x_i^k \rangle + \frac{c_i^k}{2} \|x_i - x_i^k\|^2 + g_i(x_i) \right\}$,

where $x^{k,i} := (x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^{k+1}, x_{i+1}^k, \ldots, x_N^k)$ for all $i = 1, \ldots, N$, $x^{k,0} = x^k$, $c_i^k \geq L_i$, and $\sup_{k,i}\{c_i^k\} < \infty$.

4: **end for**
5: **end for**

The PALM algorithm was introduced in [10] for a class of composite optimization problems in the general nonconvex and nonsmooth setting. Without imposing more assumptions or special structures on (2), a global sublinear rate of convergence of PALM for convex problems in the form of (2) was obtained in [25, 54]. Very recently, a globally linear convergence of PALM for problem (2) with a strongly convex objective function was obtained in [30]. Note that PALM is also called the block coordinate proximal gradient algorithm in [25] and the cyclic block coordinate descent-type method in [30].

Unlike its deterministic counterpart PALM where the (block) coordinates which are to be updated at each iteration are chosen in a cyclic fashion, in the randomized block coordinate proximal gradient method (R-BCPGM), the (block) coordinates are chosen randomly based on some probability distribution. In this paper, we prove the linear convergence for the R-BCPGM with the uniform probability distribution described as follows.

The random coordinate descent method for smooth convex problems was initiated by [41]. [48] extended it to the nonsmooth case, where the R-BCPGM was shown to obtain an ε-accurate solution with probability at least $1 - \rho$ in at most $O((N/\varepsilon)log(1/\rho))$ iterations. [40] applied the R-BCPGM for linearly constrained convex problems and showed its expected-value type linear convergence under the smoothness and strong convexity. Note that the R-BCPGM is also called the randomized block-coordinate descent method in [48]

Algorithm 3 Randomized block coordinate proximal gradient method.

1: Choose $x^0 \in \mathbb{R}^n$
2: **for** $k = 0, 1, 2, \cdots$ **do**
Generate with the uniform probability distribution a random index i_k from $\{1, 2, \ldots, N\}$

$$x^{k+1} = \arg\min_x \left\{ \langle \nabla_{i_k} f(x^k), x_{i_k} - x^k_{i_k} \rangle + \frac{c^k_{i_k}}{2} \|x - x^k\|^2 + g_{i_k}(x_{i_k}) \right\}, \tag{3}$$

where $c^k_{i_k} \geq L_i$ and $\sup_k \{c^k_{i_k}\} < \infty$
3: **end for**

and the coordinate-wise proximal-gradient method in [26]. We refer to [16] for a complete survey of the R-BCPGM.

The classical method of proving linear convergence of the aforementioned first order methods requires the strong convexity of the objective function. Surprisingly, many practical applications do not have strongly convex objective but may still have linear rate of convergence; see, e.g., [1, 63].

A new line of analysis, that circumvents these difficulties, was developed using the error bound property that relates the distance of a point to the solution set $\mathcal{X}$ to a certain optimality residual function. The error bound property is in general weaker than the strong convexity assumption and hence can be satisfied by some problems that have a non-strongly convex objective function. For convex optimization problems (1) including (2), the use of error bound conditions for fast convergence rate dates back to [33, 34]. For problem (1) with g equal to an indicator function, Luo and Tseng [34] are among the first to establish the linear convergence of feasible descent methods which include the PGM as a special case, under a so-called Luo-Tseng (local) error bound condition, i.e., for any $\xi \geq \inf_{x\in\mathbb{R}^n} F(x)$, there exist constant $\kappa > 0$ and $\epsilon > 0$, such that

$$\begin{aligned} \operatorname{dist}(x, \mathcal{X}) \leq \kappa \left\| x - \operatorname{Prox}^{\gamma}_{g}(x - \gamma\nabla f(x)) \right\|, \\ \text{whenever } F(x) \leq \xi, \ \left\| x - \operatorname{Prox}^{\gamma}_{g}(x - \gamma\nabla f(x)) \right\| \leq \epsilon, \end{aligned} \tag{4}$$

where $\operatorname{dist}(c, \mathcal{C})$ denotes the distance of a point c to a set $\mathcal{C}$. Since the above condition is abstract, it is important to identify concrete sufficient conditions under which the Luo-Tseng error bound condition holds. Moreover, it would be useful to know some scenarios where the Luo-Tseng error bound condition holds automatically.

Unfortunately, there are only a few cases where the Luo-Tseng error bound condition holds automatically. First of all, if f is strongly convex, then the Luo-Tseng error bound condition holds automatically; see [60, Theorem 4]. If f is not strongly convex but satisfy the following structured assumption:[1] $f(x) = h(Ax) + \langle q, x\rangle$ where A is some given $m \times n$ matrix, q is some given vector in $\mathbb{R}^n$, and $h : \mathbb{R}^m \to (-\infty, \infty]$ is a strongly convex continuously differentiable function, then the Luo-Tseng error bound condition holds automatically provided that g either has a polyhedral epigraph ([60, Lemma 7]) or is the group LASSO regularizer ([59, Theorem 2]).

Under the strong convexity assumption of h, it is known that the affine mapping $x \to Ax$ is invariant over the solution set and hence the solution set $\mathcal{X}$ can be rewritten as

$$\mathcal{X} = \{x | Ax = \bar{y}, \quad -\bar{\zeta} \in \partial g(x)\}, \tag{5}$$

[1]The exact definition of this structured assumption will be given in (Assumption 2) in Section 4.

where ∂g denotes the subgradient of g, $\bar{y}$ is a constant and

$$\bar{\zeta} := A^T \nabla h(\bar{y}) + q. \tag{6}$$

In the recent paper [72], under the structured assumption on f (Assumption 2) and the compactness assumption of the solution set $\mathcal{X}$, the authors show that if the perturbed solution map

$$\Gamma(p_1, p_2) := \{x | p_1 = Ax - \bar{y}, \quad p_2 \in \bar{\zeta} + \partial g(x)\}, \tag{7}$$

is calm at $(0, 0, \bar{x})$ for any $\bar{x} \in \mathcal{X}$, then the Luo-Tseng error bound condition (4) holds. Under this framework, it is shown that a number of existing error bound results in [34, 59, 60, 69] can be recovered in a unified manner.

We say that $\partial F(x) = \nabla f(x) + \partial g(x)$ is metrically subregular at $(\bar{x}, 0)$ for $\bar{x} \in \mathcal{X}$ if

$$\exists\, \kappa, \epsilon > 0, \quad \text{dist}\,(x, \mathcal{X}) \le \kappa \text{dist}(0, \nabla f(x) + \partial g(x)),\ \forall x \in \mathbb{B}_\epsilon(\bar{x}), \tag{8}$$

where $\mathbb{B}_\epsilon(\bar{x})$ denotes the open ball around $\bar{x}$ with modulus $\epsilon > 0$. Very recently, for problems in the form (1) with both f and g possibly nonconvex, [62] proves the linear convergence of the PGM to a proximal stationary point under the metric subregularity and a local proper separation condition. The result for the case where g is convex improves the result of [59] in that only the metric subregularity (8) is required which is in general weaker than the Luo-Tseng error bound condition (4).

The concept of the metric subregularity of $\partial F(x)$ at $(\bar{x}, 0)$ is equivalent to the calmness at $(0, \bar{x})$ of the set-valued map

$$\mathcal{S}(p) := \left\{x \;\middle|\; p \in \nabla f(x) + \partial g(x)\right\},$$

which is the canonically perturbed solution set represented by its first order condition, i.e., $\mathcal{S}(0) = \mathcal{X} = \left\{x \;\middle|\; 0 \in \nabla f(x) + \partial g(x)\right\}$. The calmness for a set-valued map is a fundamental concept in variational analysis; see e.g., [22, 23]. Although the terminology of "calmness" was coined by Rockafellar and Wets in [53], it was first introduced in Ye and Ye [64, Definition 2.8] as the pseudo upper-Lipschitz continuity taking into account that the calmness is weaker than both the pseudo-Lipschitz continuity of Aubin [4] and the upper-Lipschitz continuity of Robinson [49]. Therefore the calmness condition can be verified by either the polyhedral multifunction theory of Robinson [51, Proposition 1] or by the Mordukhovich criteria based on the limiting normal cone [36]. More recent sufficient conditions for calmness include the quasinormality and pseudonormality (see e.g. [20, Theorem 5.2]). Moreover, recently based on the concept of directional limiting normal cones (see e.g., [18]), some verifiable sufficient conditions for calmness have been established; see, e.g. [21, Theorem 1] and [65]. In fact, by the equivalence result (see Proposition 2), (8) is equivalent to

$$\exists\, \kappa, \epsilon > 0, \quad \text{dist}(x, \mathcal{X}) \le \kappa \left\| x - \text{Prox}_g^\gamma\, (x - \gamma \nabla f(x)) \right\|,\ \forall x \in \mathbb{B}_\epsilon(\bar{x}). \tag{9}$$

Condition (8) or equivalently (9) is point-based i.e. the error estimate is only required to hold for all points near the reference point $\bar{x}$, while the Luo-Tseng error bound condition (4) is not. Hence the Luo-Tseng error bound condition (4) is in general stronger than its point-based counterpart (9). Various results on the linear convergence of PGM are also obtained in the literature under different kinds of regularity conditions imposed on the subdifferential mapping of the objective function in (1); see, e.g., [9, 14, 29, 68] with the references and discussions therein.

In this paper, for the structured convex problem, we will utilize the solution characterization (5) and its perturbed map (7) to derive more sufficient conditions for the metric regularity/calmness condition and identify cases where the condition holds automatically. We show that the calmness of $\mathcal{S}(p)$ at $(0, \bar{x})$ for $\bar{x} \in \mathcal{X}$ is equivalent to the calmness of

$\Gamma(p_1, p_2)$ at $(0, 0, \bar{x})$. Using the weaker condition (8) allows us to obtain the linear convergence under the structured assumption on the function f (Assumption 2) without the compactness assumption on the solution set $\mathcal{X}$ (see [72, Assumption 2]). Moreover, by rewriting $\Gamma(p_1, p_2)$ as an intersection of the two set-valued maps

$$\Gamma_1(p_1) := \{x | p_1 = Ax - \bar{y}\}, \;\; \Gamma_2(p_2) := \{x | p_2 \in \bar{\zeta} + \partial g(x)\}, \tag{10}$$

we propose to use the calm intersection theorem of Klatte and Kummer [27] instead of using the boundedly linear regularity as suggested in [72, Theorem 2] to verify the calmness of $\Gamma(p_1, p_2)$. The calm intersection theorem takes advantage of nice properties possessed by $\Gamma_1(p_1)$ which represents a solution map of a perturbed linear system. Using the calm intersection theorem enhances our understanding of error bound conditions for algorithmic convergence by Ronbinson's multifunction theory, and thereby allows us to derive desired calmness conditions (see Lemma 3). More importantly, the idea behind the calm intersection theorem inspires us to come up with ways of calculating the modulus for the calmness/metric subregularity for a wide range of application problems, see, e.g. Section 5.

In contrast to the PGM, the essential difficulties for establishing the linear convergence of the R-BCPGM are associated with the randomization. For the sequence generated by the R-BCPGM applied to (2), unfortunately, in general one can hardly expect the sequential convergence of the generated sequence, see, e.g. [16, 38, 41, 48]. As a consequence, the aforementioned calmness condition/metric subregularity fails to serve as an appropriate error bound estimation in that it usually measures the distance from the iterative points to the solution set for those points near the limiting point. For this reason, [38] established the expected-value type linear convergence of a parallel version of the R-BCPGM by using a generalized type of error bound property, while [26] proved the expected-value type linear convergence of the R-BCPGM under the global Kurdyka–Łojasiewicz (KL) condition with exponent $1/2$ which is equivalent to that the global metric subregularity (see Proposition 2).

Based on a recently developed concept of bounded metric subregularity introduced in [71], we say that $\partial F(x) = \nabla f(x) + \partial g(x)$ is bounded metrically subregular at $(\bar{x}, 0)$ for $\bar{x} \in \mathcal{X}$ if for any bounded set $V \ni \bar{x}$,

$$\exists\, \kappa, \epsilon > 0, \text{ s.t.} \quad \text{dist}\,(x, \mathcal{X}) \le \kappa \text{dist}(0, \nabla f(x) + \partial g(x)), \;\forall x \in V. \tag{11}$$

Note that the modulus κ may depend on the bounded set V and V is arbitrary chosen. Hence the bounded metric subregularity of $\partial F(x)$ is a weaker concept than the global metric subregularity but stronger than the point-based metric subregularity condition (8). A very useful observation is that a polyhedral multifunction is bounded metrically subregular but not globally metric subregular. In this paper, for the first time we introduce the bounded metric subregularity condition to the study of the R-BCPGM linear convergence. For the R-BCPGM, we show that the expected-value type linear convergence holds under the bounded metric subregularity. We extend the calm intersection theorem of Klatte and Kummer [27] to the bounded calmness intersection theorem. Under the structured convexity assumption on f, using the bounded calmness intersection theorem to the perturbed map (7) we obtain some concrete sufficient condition for bounded metric subregularity/calmness.

We now summarize our contributions as follows:

- For the PGM and PALM, based on the recent result in [62] we obtain the linear convergence under the calmness of $\mathcal{S}(p)$ at $(0, \bar{x})$ (or equivalently the metric subregularity of $\partial F(x)$ at $(\bar{x}, 0)$). For the structured convex optimization problem where f has the aforementioned convex structure, we use the calm intersection theorem to the set-valued map Γ and identify three scenarios under which Γ is calm and hence the linear convergence holds. Scenario one is the case where ∂g is a polyhedral multifunction (e.g., g is a polyhedral convex regularizer); scenario two is the case where ∂g is a metrically subregular and $\Gamma_2(0)$ is a convex polyhedral set (e.g. g is the group LASSO regularizer); scenario three is the case where ∂g is metrically subregular and the set-valued map $\widehat{\Gamma}(p_1) := \Gamma_1(p_1) \cap \Gamma_2(0)$ is calm (e.g. g is the indicateor function of a ball). Compared to the related literature, e.g. [29, 33, 39, 59–61, 69, 72, 73], our approach leads to some new verifiable sufficient conditions for calmness as well as two improvements, i.e., the compactness assumption of the solution set $\mathcal{X}$ is no longer required and, the calmness modulus for the LASSO, the fused LASSO (see, e.g., [58]), the OSCAR (see, e.g., [11]), the group LASSO (see, e.g., [17]), the ball constrained problem and etc. is practically computable. Therefore, the linear convergence rate of the PGM and PALM for solving a class of structured convex optimization problems can be explicitly characterized. To the best of our knowledge, these challenging tasks have never been accomplished in the literature before.

- For the R-BCPGM, we show that the expected-value type linear convergence holds under the bounded metric subregularity. For the structured convex optimization problem where f has the aforementioned convex structure, we prove the bounded calm intersection theorem and apply it to the set-valued map Γ and identify two scenarios under which Γ is boundedly calm and hence the expected-valued type linear convergence holds. Scenario one is the case where ∂g is a polyhedral multifunction (e.g., g is a polyhedral convex regularizer); scenario two is the case where ∂g is bounded metrically subregular and $\Gamma_2(0)$ is a convex polyhedral set (e.g. g is the group LASSO regularizer). It follows that the required bounded metric subregularity holds automatically for some commonly used structured convex optimization problems with separable regularizers studied in [38, 61]. In particular, when g is the group LASSO regularizer, the required bounded metric subregularity holds automatically and hence the linear convergence of the R-BCPGM for solving the structured convex optimization problem with the group LASSO model follows. Note that the group LASSO regularizer is a nonpolyhedral regularizer. To our knowledge, these kinds of results have never been given in the literature before.

In Table 1, we summarize our contributions to the theory of linear convergence for structured convex optimization problems by comparing some existing results in the literature. Note that in all results we do not need the compactness assumption on the solution set $\mathcal{X}$, which slightly improves [59, 69, 72, 73] as the group LASSO usually induces a unbounded solution set. Moreover, we introduce an applicable approach to estimate the error bound modulus, and hence the convergence rate, which significantly improves [29, 33, 39, 59–61, 69, 72, 73]; see Section 4 for details.

It is worth to mention that, [39, 61] have also characterized the linear convergence rates of the PGM and PALM when g is an indicator function of a convex polyhedral set. However, the technique in [39, 61] relies heavily on the explicit expression of the polyhedral set indicated by g, which therefore restricts its extension to applications such as the LASSO, the fused LASSO, the group LASSO and etc. We also observe that [72] imposes the compactness assumption on $\mathcal{X}$ when g is polyhedral convex, while literature prior to [72] did not require this condition, see, e.g., [33, 39, 60, 61]. Note that this compactness assumption is restrictive when g is an indicator function of a convex polyhedral set.

Table 1 Linear rate convergence results

Algorithms/Regularizers	polyhedral convex	group LASSO	ball constraint
PGM	[33, 39, 60, 61, 72] & ★	[59, 69, 72, 73] & ★	[29] &★
PALM	[39, 61] & ★	[59] & ★	-
R-BCPGM	[26, 38, 61] & *	✓	-

✓: the linear convergence result has been established in this paper for the first time.

★: the linear convergence results listed have been improved in this paper.

*: the linear convergence result has been recovered in this paper.

2 Preliminaries

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 one norm and the infinity norm are denoted by $\|x\|_1$ and $\|x\|_\infty$, respectively. For any matrix $A \in \mathbb{R}^{m\times n}$, $\|A\| := \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}$ and $\tilde{\sigma}_{\min}(A)$ denotes the smallest nonzero singular value of A. $\mathbb{B}_r(x)$ and $\overline{\mathbb{B}}_r(x)$ denote the open and the closed Euclidean norm ball around x with modulus $r > 0$, respectively. The open and closed unit ball centred at the origin are denoted by $\mathbb{B}$ and $\overline{\mathbb{B}}$, respectively. For a given subset $\mathcal{C} \subseteq \mathbb{R}^n$, bd $\mathcal{C}$ denotes its boundary, $\text{dist}(c, \mathcal{C}) := \inf\{\|c - c'\| \,\big|\, c' \in \mathcal{C}\}$ denotes the distance from a point c in the same space to $\mathcal{C}$, and $\delta_{\mathcal{C}}(x) := \begin{cases} 0 & \text{if } x \in \mathcal{C} \\ \infty & \text{if } x \notin \mathcal{C} \end{cases}$ denotes the indicator function of $\mathcal{C}$. Let $\Phi : \mathbb{R}^n \rightrightarrows \mathbb{R}^q$ be a set-valued map (multifunction), its graph is defined by $\text{gph}\,(\Phi) := \{(x, \upsilon) \in \mathbb{R}^n \times \mathbb{R}^q \mid \upsilon \in \Phi(x)\}$. The inverse mapping of Φ, denoted by Φ^{-1} is defined by $\Phi^{-1}(\upsilon) := \{x \in \mathbb{R}^n \mid \upsilon \in \Phi(x)\}$.

2.1 Variational Analysis Background

We start by reviewing some concepts of stability of a set-valued map.

Definition 1 [4] Let $\Phi : \mathbb{R}^n \rightrightarrows \mathbb{R}^q$ be a set-valued map and $(\tilde{x}, \tilde{\upsilon}) \in \text{gph}(\Phi)$. We say that Φ is pseudo-Lipschitz continuous at $(\tilde{x}, \tilde{\upsilon})$ if there exist a neighborhood $\mathbb{V}$ of $\tilde{x}$, a neighborhood $\mathbb{U}$ of $\tilde{\upsilon}$ and $\kappa \geq 0$ such that

$$\Phi(x) \cap \mathbb{U} \subseteq \Phi\left(x'\right) + \kappa \left\|x' - x\right\| \overline{\mathbb{B}}, \quad \forall x', x \in \mathbb{V}.$$

Definition 2 [49] Let $\Phi : \mathbb{R}^n \rightrightarrows \mathbb{R}^q$ be a set-valued map and $\tilde{x} \in \mathbb{R}^n$. We say that Φ is upper-Lipschitz continuous at $\tilde{x}$ if there exist a neighborhood $\mathbb{V}$ of $\tilde{x}$ and $\kappa \geq 0$ such that

$$\Phi(x) \subseteq \Phi(\tilde{x}) + \kappa \|x - \tilde{x}\| \overline{\mathbb{B}}, \quad \forall x \in \mathbb{V}.$$

Definition 3 [64] Let $\Phi : \mathbb{R}^n \rightrightarrows \mathbb{R}^q$ be a set-valued map and $(\tilde{x}, \tilde{\upsilon}) \in \text{gph}(\Phi)$. We say that Φ is calm (or pseudo upper-Lipschitz continuous) at $(\tilde{x}, \tilde{\upsilon})$ if there exist a neighborhood $\mathbb{V}$ of $\tilde{x}$, a neighborhood $\mathbb{U}$ of $\tilde{\upsilon}$ and $\kappa \geq 0$ such that

$$\Phi(x) \cap \mathbb{U} \subseteq \Phi(\tilde{x}) + \kappa \|\tilde{x} - x\| \overline{\mathbb{B}}, \quad \forall x \in \mathbb{V}.$$

It is easy to see from the definition that both pseudo-Lipschitz continuity and the upper-Lipschitz continuity are stronger than the calmness condition. Next we propose a "bounded" version of the psedo-Lipschitz continuity that will be useful in this paper. It is obvious that the "bounded" version is stronger than the original one.

Definition 4 Let $\Phi : \mathbb{R}^n \rightrightarrows \mathbb{R}^q$ be a set-valued map and $(\tilde{x}, \tilde{\upsilon}) \in \text{gph}(\Phi)$. We say that Φ is bounded pseudo-Lipschitz continuous at $(\tilde{x}, \tilde{\upsilon})$ if for any compact set $V \ni \tilde{x}$, there exist a neighborhood $\mathbb{U}$ of $\tilde{\upsilon}$ and $\kappa \geq 0$ such that

$$\Phi(x) \cap \mathbb{U} \subseteq \Phi\left(x'\right) + \kappa \left\|x' - x\right\| \overline{\mathbb{B}}, \quad \forall x', x \in V.$$

Moreover if the above condition holds with the compact set V replaced by the whole space $\mathbb{R}^n$, we say that Φ is globally pseudo-Lipschitz continuous at $\tilde{\upsilon}$.

Note that it is easy to verify that in the definition of the calmness above, the neighborhood $\mathbb{V}$ can be taken as the whole space $\mathbb{R}^n$. Hence Φ is calm at $(\tilde{x}, \tilde{\upsilon})$ if and only if there exist a neighborhood $\mathbb{U}$ of $\tilde{\upsilon}$ and $\kappa \geq 0$ such that

$$\text{dist}\,(\upsilon, \Phi(\tilde{x})) \leq \kappa \text{dist}\left(\tilde{x}, \Phi^{-1}(\upsilon)\right), \quad \forall \upsilon \in \mathbb{U}.$$

Therefore Φ is calm at $(\tilde{x}, \tilde{\upsilon})$ if and only if its inverse map $\Psi := \Phi^{-1}$ is metrically subregular at $(\tilde{\upsilon}, \tilde{x}) \in \text{gph}\,(\Psi)$ in the following sense.

Definition 5 [13] We say that $\Psi : \mathbb{R}^q \rightrightarrows \mathbb{R}^n$ is metrically subregular at $(\tilde{\upsilon}, \tilde{x}) \in \text{gph}\,(\Psi)$ if for some $\epsilon > 0$ there exists $\kappa \geq 0$ such that

$$\text{dist}\left(\upsilon, \Psi^{-1}(\tilde{x})\right) \leq \kappa \text{dist}\,(\tilde{x}, \Psi(\upsilon)), \quad \forall \upsilon \in \mathbb{B}_\epsilon(\tilde{\upsilon}).$$

The following bounded version of the metric subregularity introduced by Zhang and Ng [71] will play an important role.

Definition 6 [71] We say that $\Psi : \mathbb{R}^q \rightrightarrows \mathbb{R}^n$ is bounded metrically subregular at $(\tilde{\upsilon}, \tilde{x}) \in \text{gph}\,(\Psi)$ if for any compact set V such that $\tilde{\upsilon} \in V$, there exists $\kappa \geq 0$ such that

$$\text{dist}\left(\upsilon, \Psi^{-1}(\tilde{x})\right) \leq \kappa \text{dist}\,(\tilde{x}, \Psi(\upsilon)), \quad \forall \upsilon \in V.$$

Note that in all definitions above, we call the constant κ the modulus.

We next propose the following proposition which is inspired by [15, Proposition 6.1.2].

Proposition 1 *For a set-valued map $\Psi : \mathbb{R}^q \rightrightarrows \mathbb{R}^n$, if there exist $\kappa, \eta > 0$ such that*

$$dist\left(\upsilon, \Psi^{-1}(\tilde{x})\right) \leq \kappa dist\,(\tilde{x}, \Psi(\upsilon)), \quad \forall \upsilon \in \mathbb{R}^n \;\; with \;\; dist\,(\tilde{x}, \Psi(\upsilon)) < \eta, \tag{12}$$

then for any $r > 0$ there exists $\kappa_r > 0$ such that

$$dist\left(\upsilon, \Psi^{-1}(\tilde{x})\right) \leq \kappa_r dist\,(\tilde{x}, \Psi(\upsilon)), \quad \forall \upsilon \in r\overline{\mathbb{B}}. \tag{13}$$

Recall that a set-valued map is called a polyhedral multifunction if its graph is the union of finitely many polyhedral convex sets. According to [51, Proposition 1], a polyhedral multifunction always satisfies condition (12). Thanks to Proposition 1, (12) implies (13) and hence a polyhedral multifunction must be bounded metrically subregular at every point in the graph of the set-valued map.

2.2 Interplay Between Regularity Conditions

In order to facilitate our discussion, we first summarize the interplay between some popular regularity conditions. In particular, since F is continuous on $\mathrm{dom}F$, the equivalence among the following four conditions is provable. Note that in this paper, we will add "global" to a condition if the condition holds for all $x \in \mathbb{R}^n$.

Proposition 2 *Given a point $\bar{x} \in \mathcal{X}$, the following conditions are equivalent.*

1) Metric subregularity: There exist $\kappa, \epsilon > 0$ such that

$$dist\,(x, \mathcal{X}) \leq \kappa dist(0, \partial F(x)), \quad \forall x \in \mathbb{B}_\epsilon(\bar{x}).$$

2) Proximal error bound: For any $\gamma > 0$, there exist $\kappa, \epsilon > 0$ such that

$$dist\,(x, \mathcal{X}) \leq \kappa \left\| x - (I + \gamma\partial g)^{-1} (x - \gamma \nabla f(x)) \right\|, \quad \forall x \in \mathbb{B}_\epsilon(\bar{x}) \cap \mathrm{dom}F,$$

where I is the identity matrix of size $n \times n$.

3) Kurdyka–Łojasiewicz (KL) property with exponent $\frac{1}{2}$: There exist $\kappa, \epsilon, r > 0$ such that

$$\kappa\,(F(x) - F(\bar{x}))^{-\frac{1}{2}}\; dist(0, \partial F(x)) \geq 1,\ \forall x \in \mathbb{B}_\epsilon(\bar{x}) \cap \{x \mid F(\bar{x}) < F(x) < F(\bar{x}) + r\}.$$

4) Quadratic growth condition: There exist $\kappa, \epsilon > 0$ such that

$$F(x) \geq F(\bar{x}) + \kappa dist^2\,(x, \mathcal{X}), \quad \forall x \in \mathbb{B}_\epsilon(\bar{x}).$$

Proof 1) $\Leftrightarrow$ 2), see [14, Theorem 3.4, Theorem 3.5]. 3) $\Leftrightarrow$ 4), see [9, Theorem 5, Corollary 6]. 1) $\Rightarrow$ 3), see [68, Theorem 1]. 4) $\Rightarrow$ 1), see [2, Theorem 3.3] or [3, Theorem 2.1]. □

In accordance with the bounded metric subregularity, naturally we may introduce some "bounded" versions of the proximal error bound, the KL property with exponent $\frac{1}{2}$ and the quadratic growth condition respectively. The proof for the equivalence reviewed in Proposition 2 can be trivially extended to that in Proposition 3.

Proposition 3 *Let $V = \overline{\mathbb{B}}_{R_0}(\bar{x}) \cap \{x \mid F(x) \leq F(\bar{x}) + r_0\}$, where $\bar{x}$ is a point in $\mathcal{X}$, $R_0 \in (0, +\infty]$ and $r_0 \in (0, +\infty)$. Then the following are equivalent:*

1) Bounded metric subregularity: There exists $\kappa > 0$ such that

$$dist\,(x, \mathcal{X}) \leq \kappa dist(0, \partial F(x)), \quad \forall x \in V.$$

2) Bounded proximal error bound: For any $\gamma > 0$, there exists $\kappa > 0$ such that

$$dist\,(x, \mathcal{X}) \leq \kappa \left\| x - (I + \gamma\partial g)^{-1} (x - \gamma \nabla f(x)) \right\|, \quad \forall x \in V.$$

3) Bounded KL property with exponent $\frac{1}{2}$: There exists $\kappa > 0$ such that

$$\kappa\,(F(x) - F(\bar{x}))^{-\frac{1}{2}}\; dist(0, \partial F(x)) \geq 1, \quad \forall x \in V \cap \{x \mid F(\bar{x}) < F(x)\}.$$

4) Bounded quadratic growth condition: There exists $\kappa > 0$ such that

$$F(x) \geq F(\bar{x}) + \kappa dist^2\,(x, \mathcal{X}), \quad \forall x \in V.$$

3 First-Order Methods and Linear Convergence Rates

This section is divided into two parts. In the first part, we briefly analyse the linear convergence of the two deterministic methods, i.e., the PGM and PALM under the metric subregularity of ∂F. In the second part, we prove the expected-value type linear convergence of the R-BCPGM under the bounded metric subregularity of ∂F.

3.1 Linear Convergence of the PGM and PALM Under the Metric Subregularity

Recall that the iteration scheme of the PGM applied to problem (1) can be rewritten as

$$x^{k+1} = (I + \gamma \partial g)^{-1} \left(x^k - \gamma \nabla f(x^k) \right).$$

The following result follows directly from [32] and [62].

Proposition 4 (Linear convergence of the PGM under the metric subregularity) *(*[32, Theorem 3.3], [62, Theorems 3.2 and 4.2]*) Assume that the step-size γ in the PGM as in Algorithm 1 satisfies $\gamma < \frac{1}{L}$. Let $\{x^k\}$ be the sequence generated by the PGM, and x^k converges to $\bar{x} \in \mathcal{X}$. Suppose that ∂F is metrically subregular at $(\bar{x}, 0)$. Then the sequence x^k converges to $\bar{x} \in \mathcal{X}$ linearly with respect to the sequence of objective function values, i.e., there exist $k_0 > 0$ and $\sigma \in (0, 1)$, such that for all $k \geq k_0$, we have*

$$F(x^k) - F^* \leq \sigma^k \left(F(x^0) - F^* \right), \quad \forall k = 0, 1, 2, \cdots.$$

Moreover we have for all $k \geq k_0$,

$$\left\| x^k - \bar{x} \right\| \leq \rho_0 \rho^k,$$

for certain $\rho_0 > 0, 0 < \rho < 1$.

The linear convergence of PALM was established in [10, Theorem 1 and Remark 6] if F satisfies KL property with exponent $\frac{1}{2}$ on $\mathcal{X}$ and the set of all limiting points of the iteration sequence is bounded. However, according to the proof of [10, Theorem 1], the KL property with exponent $\frac{1}{2}$ on a specific point $\bar{x} \in \mathcal{X}$ and boundedness of the iteration sequence suffice to guarantee the linear convergence of PALM toward $\bar{x}$. So we slightly modify the result and summarize it in the following proposition. Note that according to Proposition 2, the metric subregularity of ∂F at $(\bar{x}, 0)$ is equivalent to the KL property with exponent $\frac{1}{2}$ of F at $\bar{x}$. We therefore state the linear convergence result under the metric subregularity.

Proposition 5 (Linear convergence of PALM under the metric subregularity) *If $F(x) = f(x_1, x_2, \ldots, x_N) + \sum_{i=1}^N g_i(x_i)$, and let $\{x^k\}$ be the sequence generated by PALM algorithm in Algorithm 2, and x^k converges to $\bar{x} \in \mathcal{X}$. Assume that ∂F is metrically subregular at $(\bar{x}, 0)$. Then the sequence x^k converges to $\bar{x}$ linearly, i.e., there exist $k_0 > 0$ and $\rho_0 > 0, 0 < \rho < 1$, such that for all $k \geq k_0$, we have*

$$\left\| x^k - \bar{x} \right\| \leq \rho_0 \rho^k.$$

Remark 1 The constants σ, ρ, ρ_0 in the linear convergence rate of the PGM toward $\bar{x}$ given in Proposition 4 is closely related to the modulus of the metric subregularity of ∂F at $(\bar{x}, 0)$. In fact, when the modulus of the metric subregularity of ∂F is known, the linear convergence rate of the PGM can be characterized in terms of the modulus, see, e.g., [62, Theorems

3.2 and 4.2] together with [14, Theorem 3.4, Theorem 3.5] for details. Similarly, the linear convergence rate of PALM can also be characterized explicitly by the modulus of the metric subregularity of ∂F, see, e.g., [10, Theorem 1 and Remark 6] together with [68, Theorem 1].

3.2 Linear Convergence of the R-BCPGM Under the Bounded Metric Subregularity

Definition 7 (R-BCPGM-iteration-based error bound) Let $\left\{x^k\right\}$ be an iteration sequence generated by the R-BCPGM method. We say that the R-BCPGM-iteration-based error bound holds for $\left\{x^k\right\}$ if there exists κ such that

$$\operatorname{dist}\left(x^{k+1}, \mathcal{X}\right) \leq \kappa \left\| x^k - (I + \frac{1}{C}\partial g)^{-1}\left(x^k - \frac{1}{C}\nabla f(x^k)\right)\right\|, \quad \text{for all } k, \tag{14}$$

where $C := \sup_k\{c_{i_k}^k\}$.

We are now ready to illustrate our main result in this section. The proof of linear convergence will rely heavily on a technical result developed in [26, Theorem 6]. Instead of using the global KL property with exponent $1/2$ which is equivalent to the global metric subregularity (see Proposition 2) as in [26, Theorem 6], we employ the bounded metric subregularity and show the boundedness of the generated sequence. For succinctness, we give the detail of the proof in Appendix.

Theorem 1 (Linear convergence of the R-BCPGM under the R-BCPGM-iteration-based error bound) *Suppose that $F(x) = f(x_1, x_2, \ldots, x_N) + \sum_{i=1}^N g_i(x_i)$ and $\{x^k\}$ is an iteration sequence generated by the R-BCPGM as in Algorithm 3 and the R-BCPGM-iteration-based error bound holds for $\{x^k\}$. Then the R-BCPGM achieves a linear convergence rate in terms of the expected objective function value, i.e., there exists $\sigma \in (0, 1)$ such that*

$$\mathbb{E}[F(x^k) - F^*] \leq \sigma^k(F(x^0) - F^*), \quad \forall k = 0, 1, 2, \cdots .$$

The R-BCPGM-iteration-based error bound condition, however, depends on the iteration sequence. We now give some sufficient conditions for the R-BCPGM-iteration-based error bound condition that are independent of the iteration sequence. The equivalence between the bounded metric subregularity and the bounded KL property with exponent $\frac{1}{2}$ presented in Proposition 3 immediately yields the following corollary.

Corollary 1 (Linear convergence of the R-BCPGM under the bounded metric subregularity) *Assume that $F(x) = f(x_1, x_2, \ldots, x_N) + \sum_{i=1}^N g_i(x_i)$, and ∂F is bounded metrically subregular at $(\bar{x}, 0)$ for some point $\bar{x} \in \mathcal{X}$. Suppose that the sequence $\{x^k\}$ generated by the R-BCPGM is bounded. Then the R-BCPGM achieves a linear convergence rate in terms of the expected objective value for any iteration sequence generated by the algorithm.*

Proof By assumption, there exist some $r_0 > 0$, $R_0 > 0$ such that $\{x^k\} \subseteq V$, where $V = \overline{\mathbb{B}}_{R_0}(\bar{x}) \cap \{x \mid F(x) \leq F(\bar{x}) + r_0\}$. By definition, since ∂F is bounded metrically subregular at $(\bar{x}, 0)$ we have

$$\operatorname{dist}(x, \mathcal{X}) \leq \kappa \operatorname{dist}(0, \partial F(x)), \quad \forall x \in V.$$

By Proposition 3 it follows that for $\gamma = 1/C$, there exists $\tilde{\kappa} > 0$ such that

$$\operatorname{dist}(x, \mathcal{X}) \leq \tilde{\kappa} \left\| x - (I + \gamma\partial g)^{-1}(x - \gamma\nabla f(x))\right\|, \quad \forall x \in V.$$

It follows that the R-BCPGM-iteration-based error bound holds for $\{x^k\}$ and hence the result follows from Theorem 1. □

In the rest of this section we give a sufficient condition for the boundedness of the iteration sequence. Since the subproblem in each iteration of R-BCPGM with given block is same as the one of PALM, we can have following lemma from [6, Lemma 11.9] or [54, Lemma 3.1] easily.

Lemma 1 *If $F(x) = f(x_1, x_2, \ldots, x_N) + \sum_{i=1}^{N} g_i(x_i)$, let $\{x^k\}$ be a sequence generated by the R-BCPGM. Then $\{F(x^k)\}$ must be bounded.*

According to Lemma 1, if F is level bounded, i.e. $\{x \mid F(x) \leq F(x^0)\}$ is bounded, then the sequence $\{x^k\}$ generated by the R-BCPGM must be bounded as well. We now summerize our discussion above in the following result.

Corollary 2 *Assume that $F(x) = f(x_1, x_2, \ldots, x_N) + \sum_{i=1}^{N} g_i(x_i)$ and F is level bounded and ∂F is bounded metrically subregular at $(\bar{x}, 0)$ for some $\bar{x} \in \mathcal{X}$. Then the R-BCPGM achieves a linear convergence rate in terms of the expected objective function value for any iteration sequence generated by the algorithm.*

4 Verifying Subregularity Conditions for Structured Convex Models

We have shown in Section 3 that the PGM applied to (1) and PALM applied to (1) with g having block separable structure converge linearly under the metric subregularity of ∂F. Suppose further that ∂F is bounded metrically subregular, we have proved the linear convergence for the R-BCPGM for solving (2). In this section, we shall study the metric subregularity and the bounded metric subregularity of ∂F. Moreover, for those applications which satisfy the metric subregularity, we shall also propose an approach to estimate the metric subregularity modulus of ∂F. For this purpose, we concentrate on an important category of optimization problem (1) (including (2) with underlying structure) which satisfies the following assumption.

Assumption 2 (Structured Properties of the Function ***f***) *$f : \mathbb{R}^n \to (-\infty, \infty]$ is a function of the form*

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

where A 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 following properties:

- *(i) h is strongly convex on any convex compact subset of $\operatorname{dom} h$.*
- *(ii) h is continuously differentiable on $\operatorname{dom} h$ which is assumed to be open and ∇h is Lipschitz continuous on any compact subset $C \subseteq \operatorname{dom} h$.*

Some commonly used loss functions in machine learning such as linear regression, logistic regression and Poisson regression automatically satisfy the above assumptions. We next verify the desirable metric subregularity conditions under Assumption 2. In order to present our results more clearly, we summarize the roadmap of analysis in Fig. 1. Note that in Fig. 1,

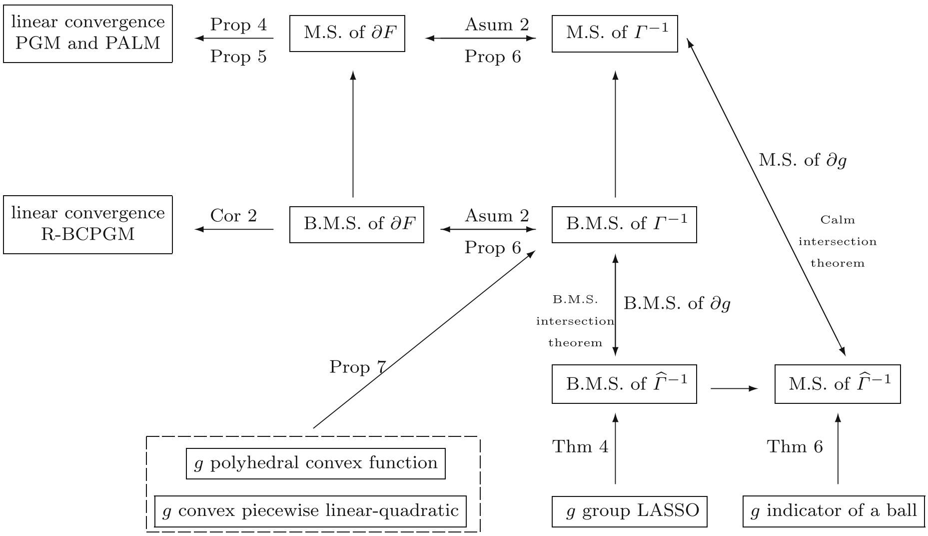

Fig. 1 Roadmap to study linear convergence

M.S. and B.M.S. denote metric subregularity and bounded metric subregularity, respectively; the formula of set-valued map Γ is given in (7) and $\widehat{\Gamma}(p_1) := \Gamma_1(p_1) \cap \Gamma_2(0)$, where both Γ_1 and Γ_2 are given in (10).

Thanks to the strong convexity of h, the following lemma follows from [33, Lemma 2.1] which shows that the affine mapping $x \to Ax$ is invariant over $\mathcal{X}$. It improves [72, Proposition 1] in that no compactness assumption on $\mathcal{X}$ is required.

Lemma 2 *Under Assumption 2, there exist $\bar{y} \in I\!R^m, \bar{\zeta} \in I\!R^n$ defined as in (6) such that $\mathcal{X} = \{x | Ax = \bar{y}, \quad 0 \in \bar{\zeta} + \partial g(x)\}$.*

Proof Since the functions f, g are convex and h is continuous differentiable on the domain of f, by the optimality condition and the chain rule, we have

$$\mathcal{X} = \{x \in \mathbb{R}^n | 0 \in A^T \nabla h(Ax) + q + \partial g(x)\}.$$

By [33, Lemma 2.1], there exists $\bar{y} \in \mathbb{R}^m$ such that $Ax = \bar{y}$ for all $x \in \mathcal{X}$. The result then follows. □

Proposition 6 *Assume that Assumption 2 is satisfied. Then the bounded metric subregularity conditions of Γ^{-1} and $\mathcal{S}^{-1}$ are equivalent. Precisely, given $\bar{x} \in \mathcal{X}$, and a compact set $V \subseteq dom F$ such that $\bar{x} \in V$, the following two statements are equivalent:*

(i) There exists $\kappa_1 > 0$ such that $dist(x, \Gamma(0,0)) \le \kappa_1 dist\left(0, \Gamma^{-1}(x)\right), \forall x \in V$.
(ii) There exists $\kappa_2 > 0$ such that $dist(x, \mathcal{S}(0)) \le \kappa_2 dist\left(0, \mathcal{S}^{-1}(x)\right), \forall x \in V$.

Proof Given $\bar{x} \in \mathcal{X}$ and a compact set V such that $\bar{x} \in V$, suppose that there exists $\kappa_1 > 0$ such that $\text{dist}(x, \Gamma(0,0)) \le \kappa_1 \text{dist}\left(0, \Gamma^{-1}(x)\right)$ for all $x \in V$. For any $x \in V$ and any

$\xi \in \nabla f(x) + \partial g(x)$, by the Lipschitz continuity of ∇h as in Assumption 2, there exists $L_h > 0$ such that

$$\begin{aligned} \mathrm{dist}\,(x, \mathcal{X}) = \mathrm{dist}\,(x, \Gamma(0,0)) &\le \kappa_1 \mathrm{dist}\left(0, \Gamma^{-1}(x)\right) \\ &\le \kappa_1 \left(\|Ax - \bar{y}\| + \|\xi - \nabla f(x) + \bar{\zeta}\|\right) \\ &\le \kappa_1 \left(\|Ax - \bar{y}\| + \|A^T \nabla h(Ax) - A^T \nabla h(\bar{y})\| + \|\xi\|\right) \\ &\le \kappa_1 (1 + \|A\| L_h) \|Ax - \bar{y}\| + \kappa_1 \|\xi\|. \end{aligned} \tag{15}$$

Let $\hat{x}$ be the projection of x on $\mathcal{X}$, since $0 \in \bar{\zeta} + \partial g(\hat{x})$ and ∂g is monotone, we have

$$\langle \xi - \nabla f(x) + \bar{\zeta}, x - \hat{x} \rangle \ge 0.$$

Moreover, since $\bar{\zeta} = A^T \nabla h(\bar{y}) + q$ and $A\hat{x} = \bar{y}$, thanks again to the strong convexity of h, we can find $\sigma > 0$ such that

$$\sigma \|Ax - \bar{y}\|^2 \le \langle \nabla h(Ax) - \nabla h(\bar{y}), Ax - \bar{y} \rangle \le \langle \xi, x - \hat{x} \rangle \le \|\xi\| \|x - \hat{x}\| = \|\xi\| \mathrm{dist}\,(x, \mathcal{X}). \tag{16}$$

Upon combining (15) and (16), we obtain

$$\mathrm{dist}\,(x, \mathcal{X}) \le \frac{\kappa_1 (1 + \|A\| L_h)}{\sqrt{\sigma}} \sqrt{\|\xi\| \mathrm{dist}\,(x, \mathcal{X})} + \kappa_1 \|\xi\|.$$

Consequently,

$$\mathrm{dist}\,(x, \mathcal{X}) \le \tilde{\kappa} \|\xi\|,$$

where

$$\tilde{\kappa} := \kappa_1 + 2c^2 + 2c\sqrt{\kappa_1 + c^2} > 0 \text{ with } c := \frac{\kappa_1 (1 + \|A\| L_h)}{2\sqrt{\sigma}}.$$

Because ξ is arbitrarily chosen in $\nabla f(x) + \partial g(x)$,

$$\mathrm{dist}\,(x, \mathcal{S}(0)) = \mathrm{dist}\,(x, \mathcal{X}) \le \tilde{\kappa} \mathrm{dist}\left(0, \mathcal{S}^{-1}(x)\right).$$

Hence, there exists a $\kappa_2 = \tilde{\kappa} > 0$ such that $\mathrm{dist}\,(x, \mathcal{S}(0)) \le \kappa_2 \mathrm{dist}\left(0, \mathcal{S}^{-1}(x)\right)$ for all $x \in V$.

Conversely, given $\bar{x} \in \mathcal{X}$ and a set compact V such that $\bar{x} \in V$, suppose that there exists a $\kappa_2 > 0$ such that $\mathrm{dist}\,(x, \mathcal{S}(0)) \le \kappa_2 \mathrm{dist}\left(0, \mathcal{S}^{-1}(x)\right)$ for all $x \in V$. For any fixed $x \in V$ and $(p_1, p_2) \in \Gamma^{-1}(x)$, it follows that

$$\begin{aligned} p_1 &= Ax - \bar{y}, \\ p_2 &\in A^T \nabla h(\bar{y}) + q + \partial g(x). \end{aligned}$$

To summarize,

$$p_2 + A^T \nabla h(Ax) - A^T \nabla h(Ax - p_1) \in A^T \nabla h(Ax) + q + \partial g(x).$$

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

$$\begin{aligned} \mathrm{dist}\,(x, \mathcal{X}) = \mathrm{dist}\,(x, \mathcal{S}(0)) &\le \kappa_2 \mathrm{dist}\left(0, \mathcal{S}^{-1}(x)\right) \\ &\le \kappa_2 \|p_2 + A^T \nabla h(Ax) - A^T \nabla h(Ax - p_1)\| \\ &\le \kappa_2 \|A\| L_h \|p_1\| + \kappa_2 \|p_2\|. \end{aligned}$$

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

$$\mathrm{dist}\,(x, \Gamma(0,0)) = \mathrm{dist}\,(x, \mathcal{X}) \le \kappa_2 (\|A\| L_h + 1) \mathrm{dist}\left(0, \Gamma^{-1}(x)\right).$$

Therefore, there exists $\kappa_1 = \kappa_2(\|A\|L_h + 1) > 0$ such that $\mathrm{dist}(x, \Gamma(0,0)) \le \kappa_1 \mathrm{dist}\left(0, \Gamma^{-1}(x)\right)$ for all $x \in V$. □

Observe that if the solution set $\mathcal{X}$ is compact, then Γ is calm at every point $(0, 0, \bar{x})$ where $\bar{x} \in \mathcal{X}$ if and only if there exist $\kappa > 0$, $\rho > 0$ such that

$$\begin{aligned}\mathrm{dist}(x, \mathcal{X}) &\le \kappa \mathrm{dist}\left(x, \Gamma^{-1}(x)\right) \\ &= \kappa\left(\|Ax - \bar{y}\| + \mathrm{dist}(\bar{v}, \partial g(x))\right) \quad \forall x \text{ with } \mathrm{dist}(x, \mathcal{X}) \le \rho.\end{aligned}$$

The above condition is actually the so-called "EBR" condition in [72]. Hence from Proposition 6 one can obtain [72, Proposition 4] but not vice versa. Since we do not assume the compactness of the solution set $\mathcal{X}$, this result improves the result in [72, Proposition 4].

We then concentrate on the metric subregularity of Γ^{-1} and conduct our analysis systematically according to different application-driven scenarios of ∂g. Given $\bar{x} \in \mathcal{X}$, we will show the following results regarding metric subregularity. In fact, when A is of full column rank, straightforwardly F is strongly convex, which implies that $\mathcal{S}^{-1}$ is metrically subregular at $(\bar{x}, 0)$, and thus, Γ^{-1} should be metrically subregular at $(\bar{x}, 0, 0)$. We are interested in the nontrivial cases of scenarios 1 - 3, where A is not of full column rank.

Scenario 1. If ∂g is a polyhedral multifunction, then Γ^{-1} is bounded metrically subregular at $(\bar{x}, 0, 0)$.

Scenario 2. If ∂g is bounded metrically subregular (not necessarily a polyhedral multifunction) at any $(\bar{x}, \bar{v}) \in gph(\partial g)$, then Γ^{-1} is bounded metrically subregular at $(\bar{x}, 0, 0)$ provided that $\Gamma_2(0)$ is a convex polyhedral set.

Scenario 3. If ∂g is metrically subregular (not necessarily bounded metrically subregular) at any $(\bar{x}, \bar{v}) \in gph(\partial g)$, then Γ^{-1} is metrically subregular at $(\bar{x}, 0, 0)$ provided that $\widehat{\Gamma}(p_1) := \Gamma_1(p_1) \cap \Gamma_2(0)$ is calm at $(0, \bar{x})$.

In particular, for each scenario, we will first delineate how to prove the theoretical arguments. We will also classify some popular models in statistics and machine learning as applications in accordance with scenarios 1-3.

4.1 Scenario 1: ∂g is Polyhedral Multifunction

In this scenario, f satisfies Assumption 2 and ∂g is a polyhedral multifunction. In this case Γ is a polyhedral multifunction and hence Γ^{-1} is also a polyhedral multifunction. Consequently, Γ satisfies condition (12) at (p_1, p_2) for any point $(x, p_1, p_2) \in gph(\Gamma^{-1})$ according to the polyhedral multifunction theory of Robinson [51, Proposition 1]. Therefore, Γ^{-1} is bounded metrically subregular at (x, p_1, p_2) by Proposition 1. According to Proposition 6, $\mathcal{S}^{-1}$ is bounded metrically subregular at $(\bar{x}, 0)$ for any given $\bar{x} \in \mathcal{X}$. Theorem 2 summarizes our discussion above.

Theorem 2 *Suppose that f satisfies Assumption 2 and ∂g is a polyhedral multifunction. Then ∂F is bounded metrically subregular at $(\bar{x}, 0)$ for any given $\bar{x} \in \mathcal{X}$.*

According to the equivalence theorem in Proposition 3, a consequence of Theorem 2 is that for convex problems with convex piecewise linear-quadratic regularizes, the objective function must satisfy KL property with exponent $\frac{1}{2}$ without the compactness assumption of the solution set $\mathcal{X}$ as in [29, Proposition 4.1].

It is interesting to see that although ∂F is not necessarily a polyhedral multifunction, through Γ we can verify that ∂F is not only metrically subregular but also bounded metrically subregular at $(\bar{x}, 0)$ for given $\bar{x} \in \mathcal{X}$.

Application of Scenario 1 polyhedral convex or piecewise quadratic function

Proposition 7 *Let $g : \mathbb{R}^n \to (-\infty, \infty]$ be convex, proper lsc and continuous on domg. Then ∂g is a polyhedral multifunction if one of following conditions hold:*

- *(i)* *g is a polyhedral convex function (see e.g.* [52] *for a definition), which includes the indicator function of a polyhedral set and the polyhedral convex regularizer.*
- *(ii)* *g is convex piecewise linear-quadratic function (see e.g.* [53, Definition 10.20] *for a definition).*

In particular, the proof for the polyhedral convex case in Proposition 7(i) can be referred to [50, Proposition 3]. This case covers scenarios where g is the LASSO regularizer (see, e.g., [57]), the $l_\infty-$norm regularizer, the fused LASSO regularizer (see, e.g., [58]), the octagonal selection and clustering algorithm for regression (OSCAR) regularizer (see, e.g., [11]). The definitions of these polyhedral convex regularizers are summarized in Table 2, where λ, λ_1 and λ_2 are given nonnegative parameters.

4.2 Scenario 2: ∂g is Bounded Metrically Subregular

In this scenario, f satisfies Assumption 2 and ∂g is bounded metrically subregular, not necessarily a polyhedral multifunction. In this case, since Γ is not a polyhedral multifunction, it is not automatically bounded metrically subregular. Note that $\Gamma(p_1, p_2)$ is the intersection of two set-valued maps $\Gamma_1(p_1)$ and $\Gamma_2(p_2)$ and the system in $\Gamma_1(0)$ is linear and hence the set-valued map Γ_1^{-1} is bounded metrically subregular at any point in its graph. If the set-valued map Γ_2^{-1} is bounded metrically subregular as well, can one claim that the set-valued map Γ^{-1} is bounded metrically subregular? The answer is negative unless some additional information is given. Here we construct a counter-example to show it is possible that the desired bounded metric subregularity of Γ^{-1} fails to hold while Γ_2^{-1} is bounded metrically subregular.

Example 1 Consider the following ball constrained optimization problem where $x \in \mathbb{R}^2$,

$$\min_x \quad F(x) := \frac{1}{2}(x_2 - 1)^2 + \delta_{\overline{\mathbb{B}}}(x). \tag{17}$$

It can be easily calculated that $\bar{x} = (0, 1)$ is the only point in solution set,

$$\Gamma_2(p_2) = \{x \mid p_2 \in \partial\delta_{\overline{\mathbb{B}}}(x)\},$$

Table 2 Polyhedral convex regularizers

Regularizers	LASSO	$l_\infty-$norm	fused LASSO	OSCAR
g(x)	$\lambda\|x\|_1$	$\lambda\|x\|_\infty$	$\lambda_1\|x\|_1 + \lambda_2 \sum_i \|x_i - x_{i+1}\|$	$\lambda_1\|x\|_1 + \lambda_2 \sum_{i<j} \max\{\|x_i\|, \|x_j\|\}$

and

$$\Gamma(p_1, p_2) = \{x \mid p_1 = (0, x_2) - (0, 1), \quad p_2 \in \partial\delta_{\overline{\mathbb{B}}}(x)\}.$$

It should be noted that $\Gamma_2^{-1}(x) = \partial\delta_{\overline{\mathbb{B}}}(x)$ is bounded metrically subregular at $(\bar{x}, 0)$ (see, e.g., Lemma 6). However, the metric subregularity of Γ^{-1} does not hold. Indeed, we may consider the sequence

$$(x_1^k, x_2^k) = (\cos(\theta^k), \sin(\theta^k)), \qquad \text{with} \qquad \theta^k \in (0, \frac{\pi}{2}), \quad \theta^k \to \pi/2,$$

and

$$p_1^k = (0, \sin(\theta^k) - 1), \quad p_2^k = 0.$$

Then we have $x^k \in \Gamma(p_1^k, p_2^k)$ and $x^k \to (0, 1)$. Since

$$\begin{aligned} \operatorname{dist}(x^k, \Gamma(0,0)) &= \operatorname{dist}(x^k, \{(0,1)\}) \\ &= \sqrt{\cos^2(\theta^k) + (\sin(\theta^k) - 1)^2} \\ &= \sqrt{2 - 2\sin(\theta^k)}, \end{aligned}$$

it follows that

$$\frac{\operatorname{dist}(x^k, \Gamma(0,0))}{\operatorname{dist}(0, \Gamma^{-1}(x^k))\|} \geq \frac{\operatorname{dist}(x^k, \Gamma(0,0))}{\|0 - (p_1^k, p_2^k)\|} = \sqrt{\frac{2}{1 - \sin(\theta^k)}} \to \infty.$$

Hence Γ^{-1} is not metrically subregular at $(\bar{x}, 0)$.

The following proposition represents a "bounded" version of the "calm intersection theorem" initiated in [27, Theorem 3.6]; see also Proposition 9. It can be used to derive concrete sufficient conditions under which $\Gamma(p_1, p_2)$ is bounded metrically subregular.

Proposition 8 (Bounded metric subregular intersection theorem) *Let $T_1 : \mathbb{R}^{q_1} \rightrightarrows \mathbb{R}^n$, $T_2 : \mathbb{R}^{q_2} \rightrightarrows \mathbb{R}^n$ be two set-valued maps. Define set-valued maps*

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

Given $\tilde{x} \in T(0,0)$, suppose that T_1^{-1} is bounded metrically subregular and bounded pseudo-Lipschitz at $(\tilde{x}, 0)$, and T_2^{-1} is bounded metrically subregular $(\tilde{x}, 0)$. Then T^{-1} is bounded metrically subregular at $(\tilde{x}, 0, 0)$ if and only if $\widehat{T}^{-1}$ is bounded metrically subregular at $(\tilde{x}, 0)$.

Proof Since the bounded metric subregularity of T^{-1} at $(\tilde{x}, 0, 0)$ implies that of $\widehat{T}^{-1}$ at $(\tilde{x}, 0)$ trivially, it suffices to show that the bounded metric subregularity of $\widehat{T}^{-1}$ at $(\tilde{x}, 0)$ implies that of T^{-1} at $(\tilde{x}, 0, 0)$.

Suppose that $\widehat{T}^{-1}$ is bounded metrically subregular at $(\tilde{x}, 0)$. Given any compact set V such that $\tilde{x} \in V$. Suppose that $\kappa_1 > 0$ and $\kappa_2 > 0$ are the modulus for the bounded metric subregularity of T_1^{-1} and T_2^{-1} at $(\tilde{x}, 0)$, respectively. Then there exist $\kappa := \max\{\kappa_1, \kappa_2\}$ such that for any $x \in V$ and $x \in T(p_1, p_2) = T_1(p_1) \cap T_2(p_2)$ with $\max\{\|p_1\|, \|p_2\|\} < \sigma$, we can find $x' \in T_1(0)$, $x'' \in T_2(0)$ satisfying

$$\max\{\operatorname{dist}(x, x'), \operatorname{dist}(x, x'')\} \leq \kappa \max\{\operatorname{dist}(0, p_1), \operatorname{dist}(0, p_2)\}.$$

Moreover, suppose that $\tilde{L} > 0$ is the modulus for the bounded pseudo-Lipschitz continuity of T_1^{-1} at $(\tilde{x}, 0)$ for the bounded set $V + \kappa\sigma\overline{\mathbb{B}}$. Since $0 \in T_1^{-1}(x')$, $x', x'' \in V + \kappa\sigma\overline{\mathbb{B}}$, and $p_1' \in T_1^{-1}(x'')$, we have

$$\mathrm{dist}(0, p_1') \leq \tilde{L}\,\mathrm{dist}(x', x'').$$

Note that $\tilde{L} > 0$ is independent of x and only dependent on V, κ, σ. Suppose that for the given V, the modulus of the bounded metric subregularity for $\widehat{T}^{-1}$ at $(\tilde{x}, 0)$ is $\kappa_T > 0$. Then since $x'' \in T_1(p_1') \cap T_2(0) = \widehat{T}(p_1')$, and $\xi \in \widehat{T}(0) = T(0, 0)$, we have

$$\mathrm{dist}(x'', \xi) \leq \kappa_T\,\mathrm{dist}(0, p_1').$$

By the inequalities proven above, we have

$$\begin{aligned}
\mathrm{dist}(x, \xi) &\leq \mathrm{dist}(x'', x) + \mathrm{dist}(x'', \xi) \\
&\leq \kappa \max\{\mathrm{dist}(p_1, 0), \mathrm{dist}(p_2, 0)\} + \kappa_T\,\mathrm{dist}(p_1', 0) \\
&\leq \kappa \max\{\mathrm{dist}(p_1, 0), \mathrm{dist}(p_2, 0)\} + \kappa_T \tilde{L}\,\mathrm{dist}(x', x'') \\
&\leq (1 + 2\kappa_T \tilde{L})\kappa\ \max\{\mathrm{dist}(p_1, 0), \mathrm{dist}(p_2, 0)\}.
\end{aligned}$$

In summary, for any compact set V such that $\tilde{x} \in V$, there exists a positive constant

$$\tilde{\kappa} := 1 + 2\kappa_T \tilde{L},$$

where $\tilde{L}$ is the modulus for the bounded pseudo-Lipschitz continuity of T_1^{-1} for the bounded set $V + \kappa\sigma\overline{\mathbb{B}}$, κ_T is the modulus for the bounded metric subregularity of $\widehat{T}^{-1}$, and a neighborhood $\mathbb{U}$ of $(0, 0)$ such that

$$\mathrm{dist}\,(x, T\,(0, 0)) \leq \tilde{\kappa}\,\mathrm{dist}\left(0, T^{-1}\,(x) \cap \mathbb{U}\right), \quad \forall x \in V,$$

i.e., T^{-1} is bounded metrically subregular at $(\tilde{x}, 0, 0)$. □

The bounded metric subregular intersection theorem in Proposition 8 plays a key role in our investigation in the sense that it provides a verifiable equivalent condition for the bounded metric subregularity of a multifunction with underlying structures. Before we can show the bounded metric subregularity of $\Gamma(p_1, p_2)$ by the bounded metric subregular intersection theorem, we need the following lemma as a preparation.

Lemma 3 *For the set-valued map $\Gamma_1(p_1) = \{x \mid p_1 = Ax - \bar{y}\}$ and any point $(\bar{p}_1, \bar{x}) \in$ gph Γ_1, Γ_1^{-1} is bounded metrically subregular at $(\bar{x}, \bar{p}_1)$, and globally pseudo-Lipschitz continuous at $\bar{p}_1$ with modulus $\|A\|$.*

Proof First, since $\Gamma_1^{-1}(x) = Ax - \bar{y}$ is a polyhedral multifunction, according to Proposition 1, Γ_1^{-1} is bounded metrically subregular at $(\bar{x}, \bar{p}_1)$ with modulus $\frac{1}{\tilde{\sigma}_{\min}(A)}$, where $\tilde{\sigma}_{\min}(A)$ denotes the smallest nonzero singular value of A. (see, e.g. [24] or Lemma 7). Moreover, for any $x', x'' \in \mathbb{R}^n$, we have

$$\|\Gamma_1^{-1}(x') - \Gamma_1^{-1}(x'')\| \leq \|A\|\|x' - x''\|.$$

By Definition 4, Γ_1^{-1} is globally pseudo-Lipschitz continuous at $\bar{p}_1$ with modulus $\|A\|$. □

The underlying property of Γ_1 which represents a perturbed linear system allows us to use the bounded metric subregular intersection theorem to characterize the bounded metric subregularity of ∂F. In terms of the bounded metric subregular intersection theorem, the bounded metric subregularity of ∂F is equivalent to the bounded metric subregularity of a linear system $0 = Ax - \bar{y}$ perturbed on an abstract set $\Gamma_2(0)$. This underlying property was neglected in [72]. However, this discovery is insightful as it reveals an important fact that the (bounded) metric subregularity conditions are automatically satisfied for certain structured convex problems because nothing else but the celebrated Robinson's polyhedral multifunction theory is needed, see, e.g., Theorem 3 and Corollary 3. In particular, upon combining Proposition 8 and Lemma 3, we obtain the main result in this part.

Theorem 3 *Suppose that f satisfies Assumption 2, ∂g is bounded metrically subregular at $(\bar{x}, -\bar{\zeta})$ where $\bar{x} \in \mathcal{X}$ and $\bar{\zeta}$ is defined as in (6). If $\Gamma_2(0) = \{x | 0 \in \bar{\zeta} + \partial g(x)\}$ is a convex polyhedral set, then ∂F is bounded metrically subregular at $(\bar{x}, 0)$.*

Proof According to Proposition 6, $\mathcal{S}^{-1} = \partial F$ is bounded metrically subregular at $(\bar{x}, 0)$ if and only if Γ^{-1} is bounded metrically subregular at $(\bar{x}, 0, 0)$. So it suffices to show that Γ^{-1} is bounded metrically subregular at $(\bar{x}, 0, 0)$ by using the bounded metrically subregular intersection theorem in Proposition 8.

First by Lemma 3, Γ_1^{-1} is bounded metrically subregular at $(\bar{x}, 0)$, and globally pseudo-Lipschitz continuous at 0. Secondly by assumption, $\Gamma_2^{-1}(x) = \bar{\zeta} + \partial g(x)$ is bounded metrically subregular at $(\bar{x}, 0)$. Since $\Gamma_2(0)$ is a convex polyhedral set, $\widehat{\Gamma}(p_1) := \Gamma_1(p_1) \cap \Gamma_2(0)$ is a polyhedral multifunction. Hence, $\widehat{\Gamma}$ satisfies condition (12) at $(0, \bar{x})$ according to the polyhedral multifunction theory of Robinson [51, Proposition 1]. Thanks to Proposition 1, $\widehat{\Gamma}^{-1}$ is bounded metrically subregular at $(\bar{x}, 0)$. Therefore, by virtue of Proposition 8, Γ^{-1} is bounded metrically subregular at $(\bar{x}, 0, 0)$ and the proof of the theorem is completed. □

Application of Scenario 2 group LASSO regularizer

Now as an application of scenario 2 we consider the group LASSO regularizer, i.e., $g(x) := \sum_{J \in \mathcal{J}} \omega_J \|x_J\|_2$, where $\omega_J \geq 0$ and $\mathcal{J}$ is a partition of $\{1, \ldots, n\}$. Throughout this paper we assume that the index set $\mathcal{J}_1 := \{J | w_J > 0$ is a nonempty. The group LASSO was introduced in [66] 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 polyhedral multifunction unless g degenerates to the LASSO regularizer. Note that as ω_J is allowed to be zero for some $J \in \mathcal{J}$, the solution set $\mathcal{X}$ to the group LASSO is not necessarily compact. Reference [72], however, requires such compactness assumption which is restrictive in some practice.

In this part, we will first show that the group LASSO falls into the category of scenario 2. The following lemma is an improvement of [72, Proposition 8] which proved that $\partial \| \cdot \|_2$ is metrically subregular at $(\bar{x}, \bar{v})$. Although the proof may look similar to [72, Proposition 8], we need to provide the detailed proof here since we are proving a stronger condition, i.e., the bounded metric subregularity. Moreover, we shall need the detailed characterization of the metric subregularity modulus for further discussion.

Lemma 4 *Let $(\bar{x}, \bar{v}) \in gph\, \partial \| \cdot \|_2$. Then $\partial \| \cdot \|_2$ is bounded metrically subregular at $(\bar{x}, \bar{v})$ with modulus $\kappa = \frac{M}{1 - \|\bar{v}\|}$ if $\|\bar{v}\| < 1$ and $\kappa = M$ if $\|\bar{v}\| = 1$.*

Proof Given an arbitrary bounded set V such that $\bar{x} \in V$, there exists $M > 0$ such that $\|x\| \leq M$ for all $x \in V$. Recall that

$$\partial \|x\|_2 = \begin{cases} x / \|x\| & \text{if } x \neq 0 \\ \overline{\mathbb{B}} & \text{if } x = 0. \end{cases}$$

Hence $(\bar{x}, \bar{v}) \in gph\, \partial\|\cdot\|_2$ implies that $\|\bar{v}\| \le 1$. Consider first the case when $\|\bar{v}\| < 1$. Then $(\partial\|\cdot\|_2)^{-1}(\bar{v}) = \{0\}$ in this case. Thus,

$$\text{dist}\left(x, (\partial\|\cdot\|_2)^{-1}(\bar{v})\right) = \|x\| \le M, \qquad \forall x \in V,$$

and

$$\text{dist}\,(\bar{v}, \partial\|x\|_2) \ge 1 - \|\bar{v}\| > 0, \qquad \forall x \in V\backslash\{0\}.$$

Therefore

$$\text{dist}\left(x, (\partial\|\cdot\|_2)^{-1}(\bar{v})\right) \le \frac{M}{1-\|\bar{v}\|}(1 - \|\bar{v}\|) \le \frac{M}{1-\|\bar{v}\|}\text{dist}\,(\bar{v}, \partial\|x\|_2)\,, \forall x \in V\backslash\{0\}.$$

Since $\text{dist}\left(0, (\partial\|\cdot\|_2)^{-1}(\bar{v})\right) = 0$, it follows that

$$\text{dist}\left(x, (\partial\|\cdot\|_2)^{-1}(\bar{v})\right) \le \frac{M}{1-\|\bar{v}\|}\text{dist}\,(\bar{v}, \partial\|x\|_2)\,, \forall x \in V. \tag{18}$$

Next we consider the case when $\|\bar{v}\| = 1$. In this case $(\partial\|\cdot\|_2)^{-1}(\bar{v}) \subseteq \{\alpha\bar{v} \mid \alpha > 0\}$ and hence

$$\text{dist}\left(x, (\partial\|\cdot\|_2)^{-1}(\bar{v})\right) \le \|x - \|x\|\cdot\bar{v}\| = \|x\|\|\frac{x}{\|x\|} - \bar{v}\|, \qquad \forall x \in V,$$

and thus

$$\text{dist}\left(x, (\partial\|\cdot\|_2)^{-1}(\bar{v})\right) \le \|x\|\text{dist}\,(\bar{v}, \partial\|x\|_2) \le M\text{dist}\,(\bar{v}, \partial\|x\|_2)\,, \forall x \in V\backslash\{0\}.$$

Again since $\text{dist}\left(0, (\partial\|\cdot\|_2)^{-1}(\bar{v})\right) = 0$, it follows that

$$\text{dist}\left(x, (\partial\|\cdot\|_2)^{-1}(\bar{v})\right) \le M\text{dist}\,(\bar{x}, \partial\|x\|_2)\,, \qquad \forall x \in V. \tag{19}$$

Combining (18) and (19), we conclude that

$$\text{dist}\left(x, (\partial\|\cdot\|_2)^{-1}(\bar{v})\right) \le \kappa\text{dist}\,(\bar{v}, \partial\|x\|_2)\,, \qquad \forall x \in V,$$

where $\kappa := \frac{M}{1-\|\bar{v}\|}$ if $\|\bar{v}\| < 1$ and $\kappa := M$ if $\|\bar{v}\| = 1$, i.e., $\partial\|\cdot\|_2$ is bounded metrically subregular at $(\bar{x}, \bar{v})$. □

Lemma 5 *Let $g(x) := \sum_{J\in\mathcal{J}} \omega_J g_J(x_J)$, where $\omega_j \ge 0$, $g_J(x_J) : I\!R^{|J|} \to I\!R$ is a convex function and $\mathcal{J}$ is a partition of $\{1, \ldots, n\}$. Then ∂g is bounded metrically subregular at any $(\bar{x}, \bar{v}) \in gph\, \partial g$ with modulus $\kappa = \max_{J\in\mathcal{J}_+} \frac{\kappa_{\bar{v}_J/\omega_J}}{\omega_J}$ with $\mathcal{J}_+ := \{J \mid \omega_J > 0\}$, if each ∂g_J is bounded metrically subregular at $(\bar{x}_J, \bar{v}_J/\omega_J)$ for any $\omega_J > 0$ with modulus $\kappa_{\bar{v}_J/\omega_J}$.*

Proof Since each g_J is a convex function and $\omega_J \ge 0$, we have

$$\partial g(x) = \prod_{J\in\mathcal{J}} \omega_J \partial g_J(x_J),$$

where $\Pi_{J\in\mathcal{J}} C_J$ denotes the Cartesian product of sets C_J. Denote by $\mathcal{J}_+ := \{J \mid \omega_J > 0\}$. For any given bounded set $V \ni \bar{x}$, we denote by V_J the projection of V to the space $\mathbb{R}^{|J|}$. Then for each $x \in V$, $x_J \in V_J$ for all $J \in \mathcal{J}$. Therefore we have

$$\begin{aligned}
\operatorname{dist}\left(x, (\partial g)^{-1}(\bar{v})\right) &\le \sum_{J\in\mathcal{J}_+} \operatorname{dist}\left(x_J, (\omega_J \partial g_J)^{-1}(\bar{v}_J)\right) \\
&= \sum_{J\in\mathcal{J}_+} \operatorname{dist}\left(x_J, (\partial g_J)^{-1}(\bar{v}_J/\omega_J)\right) \\
&\le \sum_{J\in\mathcal{J}_+} \kappa_{\bar{v}_J/\omega_J} \operatorname{dist}\left(\bar{v}_J/\omega_J, \partial g_J(x_J)\right) \\
&= \sum_{J\in\mathcal{J}_+} \frac{\kappa_{\bar{v}_J/\omega_J}}{\omega_J} \operatorname{dist}\left(\bar{v}_J, \omega_J \partial g_J(x_J)\right) \\
&\le \kappa \operatorname{dist}\left(\bar{v}, \partial g(x)\right),
\end{aligned}$$

where the second inequality follows from the bounded metric subregularity of each ∂g_J at $(\bar{x}_J, \bar{v}_J/\omega_J)$ with modulus $\kappa_{\bar{v}_J/\omega_J}$ on V_J, and $\kappa := \max_{J\in\mathcal{J}_+} \frac{\kappa_{\bar{v}_J/\omega_J}}{\omega_J}$. □

We now derive the main result of this part in Theorem 4.

Theorem 4 *Suppose that f satisfies Assumption 2 and g represents the group LASSO regularizer. Then ∂F is bounded metrically subregular at $(\bar{x}, 0)$ for any $\bar{x} \in \mathcal{X}$.*

Proof By [72, Proposition 7], $\Gamma_2(0) = (\partial g)^{-1}(-\bar{\zeta})$ is a polyhedral convex set and by Lemmas 4 and 5, $\partial g(x)$ is bounded metrically subregular. The result follows from Theorem 3. □

4.3 Scenario 3: ∂g is Metrically Subregular

In this scenario, f satisfies Assumption 2, ∂g is metrically subregular but not necessarily bounded metrically subregular. We recall the calm intersection theorem in [27, Theorem 3.6], which is in fact a localized version of Proposition 8.

Proposition 9 (Calm intersection theorem) *Let $T_1 : \mathbb{R}^{q_1} \rightrightarrows \mathbb{R}^n$, $T_2 : \mathbb{R}^{q_2} \rightrightarrows \mathbb{R}^n$ be two set-valued maps. Define set-valued maps*

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

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 T is calm at $(0, 0, \tilde{x})$ if and only if $\widehat{T}$ is calm at $(0, \tilde{x})$.

By Lemma 3, Γ_1^{-1} is metrically subregular and pseudo-Lipschitz continuous at any point on its graph. Applying Proposition 9 yields a sufficient condition for the metric subregularity of ∂F as follows.

Theorem 5 *Suppose that f satisfies Assumption 2. Given any $\bar{x} \in \mathcal{X}$, if ∂g is metrically subregular at $(\bar{x}, -\bar{\zeta})$ where $\bar{\zeta}$ is defined as in (6) and $\widehat{\Gamma}(p_1) := \Gamma_1(p_1) \cap \Gamma_2(0)$ is calm at $(0, \bar{x})$, then ∂F is metrically subregular at $(\bar{x}, 0)$.*

Compared to ∂F, the intersection of $\Gamma_1(p_1)$ and $\Gamma_2(0)$, i.e.,

$$\widehat{\Gamma}(p_1) := \Gamma_1(p_1) \cap \Gamma_2(0) = \{x \mid p_1 = Ax - \bar{y}, 0 \in \bar{\zeta} + \partial g(x)\}$$

possesses more informative structures. Theorem 5 reveals that we may focus on the sufficient condition ensuring the calmness of $\widehat{\Gamma}$ instead of that of $\mathcal{S}$. Indeed, if $\Gamma_2(0) = \{x \mid 0 \in \bar{\zeta} + \partial g(x)\}$ is a convex polyhedral set then $\widehat{\Gamma}$ is a polyhedral multifunction. By the polyhedral multifunction theory of Robinson [51, Proposition 1], given any $\bar{x} \in \mathcal{X}$, $\widehat{\Gamma}$ is upper-Lipschitz continuous at $\bar{x}$ hence calm at $(0, \bar{x})$. As a direct consequence of Theorem 5, we obtain the metric subregularity of ∂F at $(\bar{x}, 0)$.

Corollary 3 *Suppose that f satisfies Assumption 2. Given any $\bar{x} \in \mathcal{X}$, suppose that ∂g is metrically subregular at $(\bar{x}, -\bar{\zeta})$ where $\bar{\zeta}$ is defined as in (6), and $\Gamma_2(0)$ is a convex polyhedral set, then ∂F is metrically subregular at $(\bar{x}, 0)$.*

Application of Scenario 3 the indicator function of a ball constraint

We next demonstrate an application of scenario 3. To this end, let g represent the indicator function of a closed ball, i.e., $g(x) = \delta_{\overline{\mathbb{B}}_r(0)}(x)$. According to [52, Page 215], $\partial g(x) = \partial \delta_{\overline{\mathbb{B}}_r(0)}(x) = \mathcal{N}_{\overline{\mathbb{B}}_r(0)}(x)$ where $\mathcal{N}_{\mathcal{C}}(c)$ denotes the normal cone to set $\mathcal{C}$ at c.

Lemma 6 *Let $g(x) := \delta_{\overline{\mathbb{B}}_r(0)}(x)$, where r is a positive constant. Then for any point $(\bar{x}, \bar{v}) \in gph\ \partial g$, ∂g is metrically subregular at $(\bar{x}, \bar{v})$. Specially, ∂g is bounded metrically subregular at $(\bar{x}, \bar{v})$ provided $\bar{v} = 0$.*

Proof Consider first the case where $\bar{v} = 0$. Obviously, $\mathcal{N}^{-1}_{\overline{\mathbb{B}}_r(0)}(\bar{v}) = \overline{\mathbb{B}}_r(0)$ in this case. Given an arbitrary bounded set V such that $\bar{x} \in V$ and any $\kappa > 0$, if $x \in V \cap \overline{\mathbb{B}}_r(0)$, we have

$$\text{dist}\left(x, \mathcal{N}^{-1}_{\overline{\mathbb{B}}_r(0)}(\bar{v})\right) = \text{dist}\left(x, \overline{\mathbb{B}}_r(0)\right) = 0 \leq \kappa\, \text{dist}\left(\bar{v}, \mathcal{N}_{\overline{\mathbb{B}}_r(0)}(x)\right). \tag{20}$$

On the other hand, as $\partial g(x) = \mathcal{N}_{\overline{\mathbb{B}}_r(0)}(x) = \emptyset$ if $x \notin \overline{\mathbb{B}}_r(0)$, (20) holds for any $x \notin \overline{\mathbb{B}}_r(0)$. Thus ∂g is bounded metrically subregular at $(\bar{x}, \bar{v})$ when $\bar{v} = 0$. Consider the other case where $\bar{v} \neq 0$. It follows from [52, Corollary 23.5.1] that $(\partial g(x))^{-1} = \partial g^*$, where $g^*(v) := \sup_x\{\langle v, x\rangle - g(x)\}$ denotes the conjugate function of g. As $g(x) = \delta_{\overline{\mathbb{B}}_r(0)}(x)$, it can be easily calculated that $g^*(v) = r\|v\|$. Since when $\bar{v} \neq 0$, $g^*(v) = r\|v\|$ is second-order continuously differentiable at $\bar{v}$, then $\partial g^* = \nabla g^*$ and ∇g^* is locally Lipschitz continuous around $\bar{v}$. Thus ∂g^* is calm at $(\bar{v}, \bar{x})$ with modulus $\frac{2r}{\|\bar{v}\|}$ and ∂g is metrically subregular at $(\bar{x}, \bar{v})$. Combining the case $\bar{v} = 0$ and $\bar{v} \neq 0$, we have shown that ∂g is metrically subregular at $(\bar{x}, \bar{v})$. □

Proposition 10 *Let $g(x) = \delta_{\overline{\mathbb{B}}_r(0)}(x)$, where r is a positive constant. Given any $\bar{x} \in \mathcal{X}$, if one of the following statements is satisfied:*

1. *$\bar{x} \in \mathcal{X} \cap bd\mathbb{B}_r(0)$ and $\bar{\zeta}$ as defined in (6) is nonzero, or*
2. *$\bar{x} \in \mathcal{X} \cap int\mathbb{B}_r(0)$,*

then $\widehat{\Gamma}(p_1) := \Gamma_1(p_1) \cap \Gamma_2(0) = \{x \mid p_1 = Ax - \bar{y}, 0 \in \bar{\zeta} + \partial g(x)\}$ is calm at $(0, \bar{x})$.

Proof Case 1: $\bar{x} \in \mathcal{X} \cap \text{int}\mathbb{B}_r(0)$. In this case we have $\partial g(\bar{x}) = \mathcal{N}_{\overline{\mathbb{B}}_r(0)}(\bar{x}) = \{0\}$. It follows that $\bar{\zeta} = 0$ and thus in this case

$$\Gamma_2(0) = \{x \mid 0 \in \mathcal{N}_{\overline{\mathbb{B}}_r(0)}(x)\} = \overline{\mathbb{B}}_r(0). \tag{21}$$

Let $\epsilon > 0$ be such that $\mathbb{B}_\epsilon(\bar{x}) \subset \mathbb{B}_r(0)$. For any $x \in \mathbb{B}_\epsilon(\bar{x}) \cap \widehat{\Gamma}(p_1)$, let x_1 be the projection of x onto $\Gamma_1(0) = \{x \mid 0 = Ax - \bar{y}\}$, i.e., $x_1 \in \Gamma_1(0)$ and $\|x - x_1\| = \text{dist}(x, \Gamma_1(0))$. Since $\bar{x} \in \Gamma_1(0)$, we have $\|x - x_1\| \leq \|x - \bar{x}\| < \epsilon$, and thus $x_1 \in \mathbb{B}_r(0) \subset \Gamma_2(0)$ by virtue of (21). Hence we have $\text{dist}(x, \Gamma_1(0) \cap \Gamma_2(0)) \leq \|x - x_1\| = \text{dist}(x, \Gamma_1(0))$ for any $x \in \mathbb{B}_\epsilon(\bar{x}) \cap \widehat{\Gamma}(p_1)$. Since $\Gamma_1(0)$ is the solution to a linear system, thanks to Hoffman's error bound, there exists $\kappa = \tilde{\sigma}_{\min}(A)^{-1} > 0$ such that

$$\text{dist}(x, \widehat{\Gamma}(0)) \leq \text{dist}(x, \Gamma_1(0)) \leq \kappa \|p_1\|, \quad \forall\, x \in \mathbb{B}_\epsilon(\bar{x}) \cap \widehat{\Gamma}(p_1),$$

or equivalently

$$\text{dist}(x, \widehat{\Gamma}(0)) \leq \kappa \text{dist}(0, \widehat{\Gamma}^{-1}(x)), \quad \forall\, x \in \mathbb{B}_\epsilon(\bar{x}).$$

By definition, $\widehat{\Gamma}(p_1)$ is calm at $(0, \bar{x})$.

Case 2: $\bar{x} \in \mathcal{X} \cap \text{bd}\mathbb{B}_r(0)$. Observe that when x lies on the boundary of a closed ball, the normal cone to the closed ball at x must contain all rays in the direction of x, i.e., $\mathcal{N}_{\overline{\mathbb{B}}_r(0)}(x) = \{\alpha x \mid \alpha \geq 0\}$ for any $x \in \overline{\mathbb{B}}_r(0)$. Since $-\bar{\zeta} \neq 0$ is a fixed direction and $-\bar{\zeta} \in \mathcal{N}_{\overline{\mathbb{B}}_r(0)}(\bar{x}) = \{\alpha \bar{x} \mid \alpha \geq 0\}$, it follow that

$$\Gamma_2(0) = \{x \mid 0 \in \bar{\zeta} + \mathcal{N}_{\overline{\mathbb{B}}_r(0)}(x)\} = \{\bar{x}\}.$$

Moreover since $A\bar{x} = \bar{y}$, we have

$$\widehat{\Gamma}(p_1) := \Gamma_1(p_1) \cap \Gamma_2(0) = \{x \mid p_1 = Ax - \bar{y}\} \cap \{\bar{x}\} = \emptyset, \quad \text{whenever } p_1 \neq 0.$$

In this case, the calmness of $\widehat{\Gamma}(p_1)$ at $(0, \bar{x})$ holds trivially by definition.

Combining cases 1 and 2, we obtain the calmness of $\widehat{\Gamma}(p_1)$ at $(0, \bar{x})$. □

Thanks to Theorem 5, Lemma 6 and Proposition 10, we obtain the following result directly.

Theorem 6 *Suppose that f satisfies Assumption 2 and g represents the indicator function of a closed ball with modulus $r > 0$. Given any $\bar{x} \in \mathcal{X}$, if one of the following statements is satisfied:*

1. $\bar{x} \in \mathcal{X} \cap bd\mathbb{B}_r(0)$ *and* $\bar{\zeta} \neq 0$, *or*
2. $\bar{x} \in \mathcal{X} \cap int\mathbb{B}_r(0)$,

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

Remark 2 The assumption that $\bar{\zeta} \neq 0$ whenever $\bar{x} \in \text{bd}\mathbb{B}_r(0)$ in Theorem 6 may not be dismissed for the ball constrained problem. We can use the following ball constrained optimization problem given in Example 1,

$$\min_x \quad F(x) := \frac{1}{2}(x_2 - 1)^2 + \delta_{\overline{\mathbb{B}}}(x),$$

to show it is possible that the desired metric subregularity fails to hold while the assumption is not satisfied. It should be noted that $\bar{x} = (0, 1)$ is the only point in solution set, and the function $f(x) := \frac{1}{2}(x_2 - 1)^2$ has global minima at any $(x_1, 1)$ with a zero gradient $\nabla f(x_1, 1) = (0, 0)$. However, for this example, the metric subregularity of ∂F does not hold. Indeed, as shown in Example 1, Γ^{-1} is not metrically subregular at $(\bar{x}, 0)$. Then, Proposition 6 tells us that ∂F can not be metrically subregular at $(\bar{x}, 0)$.

Remark 3 Recently in [29, Proposition 4.2], the authors showed that for the ball constrained problem, when $\min_x f(x) < \min_x F(x)$, F satisfies KL property with exponent $\frac{1}{2}$ on each $\bar{x} \in \mathcal{X}$. According to Proposition 2, ∂F is then metrically subregular at $(\bar{x}, 0)$ for any $\bar{x} \in \mathcal{X}$. Indeed, the assumption $\min_x f(x) < \min_x F(x)$ is equivalent to saying that $\bar{\zeta} \neq 0$.

In this regard, Theorem 6 slightly improves [29, Proposition 4.2] in that we have shown that for those $\bar{x} \in \mathcal{X} \cap \text{int}\mathbb{B}_r(0)$, F satisfies the KL property with exponent $\frac{1}{2}$ at $\bar{x}$ automatically without any restriction on $\bar{\zeta}$. Moreover, in the next section, we will show how to estimate the calmness modulus for ball constrained problem, see, e.g., Theorem 7. This error bound estimation significantly improves [29, Proposition 4.2].

5 Calculus of Modulus of Metric Subregularity

So far we have verified the calmness conditions for $\mathcal{S}$ under Scenarios 1-3. In this section, we focus on calculating the calmness modulus of $\mathcal{S}$ which relies heavily on the calm intersection theorem. Indeed, the calm intersection theorem bridges the metric subregularity of $\widehat{\Gamma}^{-1}$ with the metric subregularity of ∂F. As a consequence, it enhances our understanding on the metric subregularity of ∂F for some applications such as the LASSO and the group LASSO. Moreover, it leads us to an interesting observation that the modulus of the metric subregularity of ∂F is now computable. Therefore, thanks to Remark 1, the linear convergence rate of the PGM or PALM can be explicitly calculated. In fact, by summarizing the proofs in Propositions 6 and 8, and Lemma 3, we are now in the position to estimate the modulus of the metric subregularity of ∂F through those of ∂g and $\widehat{\Gamma}^{-1}$ together with some problem data.

Theorem 7 *Given $\bar{x} \in \mathcal{X}$. Suppose that f satisfies Assumption 2, ∂g is metrically subregular at $(\bar{x}, -\bar{\zeta})$ with modulus κ_g and $\widehat{\Gamma}(p_1)$ is calm at $(0, \bar{x})$ with modulus κ, i.e., there exist $\kappa, \kappa_g > 0$ and $\epsilon > 0$ such that for all $x \in \mathbb{B}_\epsilon(\bar{x}) \subseteq dom\ f$,*

$$dist\left(x, \widehat{\Gamma}(0)\right) \leq \kappa dist\left(0, \widehat{\Gamma}^{-1}(x)\right),$$

and

$$dist\left(x, (\partial g)^{-1}(-\bar{\zeta})\right) \leq \kappa_g dist\left(-\bar{\zeta}, \partial g(x)\right).$$

Then ∂F is metrically subregular at $(\bar{x}, 0)$, i.e.,

$$dist\left(x, (\partial F)^{-1}(0)\right) \leq \tilde{\kappa} dist\left(0, \partial F(x)\right), \forall x \in \mathbb{B}_\epsilon(\bar{x}),$$

where the modulus $\tilde{\kappa} := \kappa_1 + 2c^2 + 2c\sqrt{\kappa_1 + c^2} > 0$ with $c := \frac{\kappa_1(1+\|A\|L_h)}{2\sqrt{\sigma}}$, $\kappa_1 := (1+\kappa\|A\|)\max\{\frac{1}{\tilde{\sigma}_{\min}(A)}, \kappa_g\}$, σ and L_h are the strong convexity modulus of h and Lipschitz continuity constant of ∇h on $\mathbb{B}_\epsilon(\bar{x})$, respectively.

Proof By Lemma 3, Γ_1^{-1} is bounded metrically subregular with modulus $\frac{1}{\tilde{\sigma}_{\min}(A)}$ and globally pseudo-Lipschitz continuous at 0 with modulus $\|A\|$. Hence it follows by the proof of Proposition 8 that $\Gamma^{-1}(p_1, p_2)$ is bounded metrically subregular at $(\bar{x}, 0, 0)$ with modulus

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

Now applying Proposition 6, with parameter $c = \frac{\kappa_1(1+\|A\|L_h)}{2\sqrt{\sigma}}$, we have that $\partial F(x)$ is bounded metrically subregular at $(\bar{x}, 0)$ with modulus $\tilde{\kappa} = \kappa_1 + 2c^2 + 2c\sqrt{\kappa_1 + c^2} > 0$. □

Thanks to Theorem 7, as long as we know the moduli of the metric subregularity of ∂g and $\widehat{\Gamma}^{-1}$, the modulus of the metric subregularity of ∂F can be estimated. The main difficulty is associated with the estimation of the calmness modulus of $\widehat{\Gamma}$. Based on different problem structures, we may divide our discussion for the calmness modulus calculation of $\widehat{\Gamma}$ for applications into following two classes.

Class 1. As we observed in Sections 4.1 and 4.2, $\widehat{\Gamma}$ actually represents a perturbed linear system on a convex polyhedral set for a wide range of applications, including the LASSO, the fused LASSO, the OSCAR and the group LASSO. Although computing the calmness modulus is always a challenging task, thanks to the calm intersection theorem, $\widehat{\Gamma}$ can be recharacterized as a partially perturbed polyhedral set for the mentioned applications. Hence, the calmness modulus of $\widehat{\Gamma}$ is achievable by the Hoffman's error bound theory (see Lemma 7) or its variant (see Lemma 8).

Class 2. The calmness modulus calculation for ball constrained problem, however, is a little different. In fact, in the proof of Proposition 10, the calmness modulus of $\widehat{\Gamma}$ has been explicitly characterized. Together with Lemma 6 where the metric subregularity modulus of ∂g has been calculated, the calculus rule for the metric subregularity modulus of ∂F presented in Theorem 7 is therefore applicable.

Remark 4 We can see from the discussions above that for both two classes, the calmness modulus of $\widehat{\Gamma}$ is achieved by using the constant in Hoffman's error bound or its variant. Hence, the main difficulty arises from the estimation of the Hoffman's error bound constant. In the two illustrative examples associated with the LASSO and group LASSO given below, though we can give the formulae of the constants of Hoffman's error bound and its variant through (24) and (27), it is still a challenging task to calculate a sharp estimation for the Hoffman's error bound constant or its variant through such formulae. Recently, [45, 46] have proposed tractable numerical algorithms for computing Hoffman constants. These algorithms can be used in the calmness modulus calculation of $\widehat{\Gamma}$.

We next show how to calculate the calmness modulus on specific application problems. We take the LASSO and group LASSO as illustrative examples while the extension to other problems is purely technical and hence omitted.

Calculus of calmness modulus for the LASSO Suppose that f satisfies Assumption 2 and $g(x) = \lambda\|x\|_1$ with $\lambda > 0$ in problem (1). Recall that $-\bar{\zeta} \in \partial g(\bar{x})$ is defined as in (6). By Lemma 5, $\partial g(x)$ is bounded metrically subregular at $(\bar{x}, -\bar{\zeta})$. That is, for any $M > 0$,

$$\text{dist}\left(x, (\partial g)^{-1}(-\bar{\zeta})\right) \leq \kappa_{l_1} \text{dist}\left(-\bar{\zeta}, \partial g\,(x)\right), \quad \forall \|x\| \leq M,$$

where $\kappa_{l_1} = \frac{\kappa_{-\bar{\zeta}/\lambda}}{\lambda}$ and $\kappa_{-\bar{\zeta}/\lambda}$ is the metric subregularity modulus of $\partial\|\cdot\|_1$ at $(\bar{x}, -\bar{\zeta}/\lambda)$. It follows by Lemma 4 that $\kappa_{-\bar{\zeta}/\lambda} = \frac{M}{\lambda(1-\bar{c})}$ with

$$\bar{c} = \max_{\{i:|\bar{\zeta}_i/\lambda|<1\}} |\bar{\zeta}_i/\lambda|; \quad \bar{c} = 0 \text{ if } \{i : |\bar{\zeta}_i/\lambda|<1\} = \emptyset. \tag{22}$$

We are left to estimate the calmness modulus of $\widehat{\Gamma}(p)$. Again under the setting that $g(x) = \lambda\|x\|_1$ for some $\lambda > 0$, given $\bar{\zeta}$, we shall define index sets

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

Moreover, we shall need the following notations.

- e_i denotes the vector whose ith entry is 1 and other entries are zero,
- $E_0 \in \mathbb{R}^{|I_0|\times n}$ denotes a matrix whose rows are $\{e_i\}_{i\in I_0}$,
- $E_1 \in \mathbb{R}^{|I_+|\times n}$ denotes a matrix whose rows are $\{e_i\}_{i\in I_+}$,
- $E_2 \in \mathbb{R}^{|I_-|\times n}$ denotes a matrix whose rows are $\{-e_i\}_{i\in I_-}$.

By constructing two matrices as

$$\tilde{A} := \begin{pmatrix} A \\ E_0 \end{pmatrix}, \quad \tilde{E} := \begin{pmatrix} E_1 \\ E_2 \end{pmatrix},$$

$\widehat{\Gamma}$ can be rewritten as a perturbed system of linear equality and inequality constraints:

$$\widehat{\Gamma}(p) = \{x \in \mathbb{R}^n \mid p = Ax - \bar{y}, E_0 x = 0, \tilde{E}x \leq 0\}. \tag{23}$$

We are in the position to apply Hoffman's error bound to calculate the calmness modulus of $\widehat{\Gamma}$. We first recall the Hoffman's error bound theory.

Lemma 7 (Hoffman's error bound) [19, 24, 28] *Let P be the polyhedral convex set $P := \{x \mid \tilde{A}x = \tilde{b}, \tilde{C}x \leq \tilde{d}\}$, where $\tilde{A}, \tilde{C}$ are given matrices and $\tilde{b}, \tilde{d}$ are given vectors of appropriate sizes. Then for any x, it holds*

$$dist\,(x, P) \leq \theta(\tilde{A}, \tilde{C}) \left\| \left(\tilde{A}x - \tilde{b}, \max\{0, \tilde{C}x - \tilde{d}\}\right) \right\|,$$

where

$$\theta(\tilde{A}, \tilde{C}) := \sup_{u,v} \left\{ \|(u, v)\| \;\middle|\; \begin{array}{l} \|\tilde{A}^T u + \tilde{C}^T v\| = 1, v \geq 0, \\ \text{The corresponding rows of } \tilde{A}, \tilde{C} \text{ to } u, v\text{'s} \\ \text{non-zero elements are linearly independent} \end{array} \right\}. \tag{24}$$

Applying the Hoffman's error bound to the perturbed linear system in (23), we obtain the modulus for the calmness of the set-valued map $\widehat{\Gamma}(p)$.

Proposition 11 *Suppose that $g(x) = \lambda\|x\|_1$. For a given $\bar{\zeta}$, $\widehat{\Gamma}(p)$ is globally calm with modulus $\theta(\tilde{A}, \tilde{E})$ defined as in (24), i.e.,*

$$dist\left(x, \widehat{\Gamma}(0)\right) \leq \theta(\tilde{A}, \tilde{E}) dist\left(0, (\widehat{\Gamma})^{-1}(x)\right), \quad \forall x.$$

In Theorem 7, using $\kappa_g = \frac{M}{\lambda(1-\bar{c})}$ and $\kappa = \theta(\tilde{A}, \tilde{E})$ we are eventually able to calculate the calmness modulus for the LASSO.

Theorem 8 *Consider the LASSO problem. That is, f satisfies Assumption 2 and $g(x) = \lambda\|x\|_1$ with $\lambda > 0$ in problem (1). For any given positive number M such that h is strongly convex on $M\bar{\mathbb{B}}$ with modulus σ and ∇h is Lipschitz continuous on $M\bar{\mathbb{B}}$ with constant L_h, there exists $\tilde{\kappa}_{Lasso} > 0$ such that*

$$dist\left(x, (\partial F)^{-1}(0)\right) \leq \tilde{\kappa}_{Lasso} dist\,(0, \partial F\,(x))\,, \quad \forall \|x\| \leq M.$$

In particular, $\tilde{\kappa}_{Lasso} = \kappa_{Lasso} + 2{c_{Lasso}}^2 + 2c_{Lasso}\sqrt{\kappa_{Lasso} + {c_{Lasso}}^2} > 0$ with

$$c_{Lasso} = \frac{\kappa_{Lasso}(1 + \|A\|L_h)}{2\sqrt{\sigma}}, \quad \kappa_{Lasso} = (1 + \theta(\tilde{A}, \tilde{E})\|A\|) \max\{\frac{1}{\tilde{\sigma}_{\min}(A)}, \frac{M}{\lambda(1-\bar{c})}\}$$

where $\bar{c}$ is defined as in (22).

Calculus of calmness modulus for the group LASSO Suppose that f satisfies Assumption 2 and $g(x) := \sum_{J\in\mathcal{J}} \omega_J \|x_J\|_2$, where $\omega_J \geq 0$ and $\mathcal{J}$ is a partition of $\{1,\dots,n\}$ is the group LASSO regularizer. Recall that $-\bar{\zeta} \in \partial g(\bar{x})$ is defined as in (6). According to Lemmas 4 and 5, $\partial g(x)$ is bounded metrically subregular at $(\bar{x}, -\bar{\zeta})$ and for any given positive number M,

$$\mathrm{dist}\left(x, (\partial g)^{-1}(-\bar{\zeta})\right) \leq \tilde{\kappa}_g \,\mathrm{dist}\left(-\bar{\zeta}, \partial g(x)\right), \quad \forall \|x\| \leq M,$$

where

$$\tilde{\kappa}_g := \max_{J\in\mathcal{J}_+} \left\{\frac{\kappa_{-\bar{\zeta}_J/\omega_J}}{\omega_J}\right\} = \max_{J\in\mathcal{J}_+} \{\kappa_{g_J}\}, \tag{25}$$

with $\kappa_{g_J} = \frac{M}{\omega_J - \|\bar{\zeta}_J\|}$, if $\|\bar{\zeta}_J\| < \omega_J$; $\kappa_{g_J} = M$, if $\|\bar{\zeta}_J\| = \omega_J$, and $\mathcal{J}_+ := \{J \mid \omega_J > 0\}$.

We shall next estimate the calmness modulus of $\widehat{\Gamma}$. Define the index set

$$\mathcal{J}_1 := \{J \in \mathcal{J}_+ \mid \|\bar{\zeta}_J\| = \omega_J, \omega_J > 0\}.$$

Moreover, for simplicity we shall need the following notations.

- For any $J \in \mathcal{J}_1$, $\tilde{g}_J$ denotes the vector whose jth entry is $\bar{\zeta}_j$ for $j \in J$ and other entries are zero,
- $K \in \mathbb{R}^{p\times n}$ denotes a matrix whose rows are $\{e_i\}_{i\in J_+}$ with $J_+ = \cup_{J\in\mathcal{J}_+} J$, and $p = |J_+|$,
- $D \in \mathbb{R}^{n\times|\mathcal{J}_1|}$ denotes a matrix whose columns are $\{-\tilde{g}_J\}_{J\in\mathcal{J}_1}$.

From the formula for the subdifferential of the group LASSO in [72, Proposition 7], it is easy to see that $\widehat{\Gamma}$ can be characterized in terms of the above notation as

$$\widehat{\Gamma}(p) = \{x \in \mathbb{R}^n \mid p = Ax - \bar{y}, Kx = KD\alpha, \alpha \geq 0\}. \tag{26}$$

Note that for the LASSO, $\widehat{\Gamma}(p)$ can be characterized as a perturbed linear equality and inequality system, see (23). The group LASSO, however, is to some extent different. Actually, one may not easily characterize $\widehat{\Gamma}(p)$ in terms of linear equality and inequality explicitly. Instead, as shown in (26), $\widehat{\Gamma}(p)$ can be expressed as a linear system perturbed over a convex cone. Unfortunately, the Hoffman's error bound theory is not directly applicable for $\widehat{\Gamma}(p)$ in (26). In order to estimate the calmness modulus of $\widehat{\Gamma}$ for the group LASSO, we shall first establish an error bound result analogous to Hoffman's error bound, which is inspired by [28, 70].

Lemma 8 (Partial error bound over a convex cone) *Let P be a polyhedral set $P := \{x \in \mathbb{R}^n \mid \tilde{A}x = \tilde{b},\ \tilde{K}x \in \mathcal{D}\}$, where $\tilde{A}$ is a matrix of size $m \times n$, $\tilde{K}$ is a matrix of size $p \times n$, $\tilde{b} \in \mathbb{R}^m$, $\mathcal{D} := \{z \mid z = \sum_{i=1}^{l} \alpha_i d_i, \alpha_i \geq 0\}$, and $\{d_i\}_{i=1}^{l} \subseteq \mathbb{R}^p$. Then*

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

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

$$\bar{\theta}(\mathcal{M}) := \sup_{\lambda,\mu,\nu} \left\{ \|\lambda\| \;\middle|\; \begin{array}{l} \|\mathcal{M}(\lambda,\mu,\nu)\| = 1, \nu \geq 0, \\ \textit{The corresponding rows of } \mathcal{M} \textit{ to } \lambda, \mu, \nu\textit{'s} \\ \textit{non-zero elements are linearly independent.} \end{array} \right\}. \tag{27}$$

Proof For any $x \notin P$, let $\bar{x}$ denote the projector of x on P, i.e., $\bar{x} = \arg\min_{y \in P} \|x - y\|$. Hence there is $\bar{\alpha} \geq 0$ such that $(\bar{x}, \bar{\alpha})$ is an optimal solution of the following problem

$$\begin{aligned} \min_{y,\alpha} \quad & \tfrac{1}{2}\|y - x\|^2 \\ s.t. \quad & \tilde{A}y = \tilde{b},\ \tilde{D}\alpha = \tilde{K}y,\ \alpha \geq 0. \end{aligned}$$

The Karush-Kuhn-Tucker (KKT) optimality condition for above problem yields the existence of multipliers λ, μ, ν such that

$$\begin{aligned} \bar{x} - x + \tilde{A}^T\lambda - \tilde{K}^T\mu &= 0, \\ \tilde{D}^T\mu - \nu &= 0, \\ \nu \geq 0,\ \nu_i\bar{\alpha}_i &= 0. \end{aligned} \tag{28}$$

Define index set $I' := \{i \in \{1, \ldots, l\} \mid \bar{\alpha}_i = 0\}$, matrix $E_{I'} \in \mathbb{R}^{l \times l}$ as $E_{I'ii} = 1$ if $i \in I'$ and $E_{I'ij} = 0$ otherwise. Let $\mathcal{M}_{I'} := \begin{bmatrix} \tilde{A}^T & -\tilde{K}^T & 0 \\ 0 & \tilde{D}^T & -E_{I'} \end{bmatrix}$, and

$$\mathcal{F}_{I'} := \{(\lambda, \mu, \nu) \mid \mathcal{M}_{I'}(\lambda, \mu, \nu) = (x - \bar{x}, 0, 0),\ \nu \geq 0\}.$$

According to [56, Theorem 2.12.4] or [28, Lemma 2.1], there is $(\hat{\lambda}, \hat{\mu}, \hat{\nu}) \in \mathcal{F}_{I'}$ such that the corresponding columns of $\mathcal{M}_{I'}$ to $(\hat{\lambda}, \hat{\mu}, \hat{\nu})$'s non-zero elements are linearly independent. It is easy to see that $(\hat{\lambda}, \hat{\mu}, \hat{\nu})$ also satisfies (28). It follows that

$$\begin{aligned} \|\bar{x} - x\|^2 &= -(\bar{x} - x)^T\tilde{A}^T\hat{\lambda} + (\bar{x} - x)^T\tilde{K}^T\hat{\mu} \\ &= (\tilde{A}x - \tilde{b})^T\hat{\lambda} + (\bar{x} - x)^T\tilde{K}^T\hat{\mu}, \end{aligned}$$

where the second equality comes from $\tilde{A}\bar{x} = \tilde{b}$. Moreover, $\bar{x}^T\tilde{K}^T\hat{\mu} = \bar{\alpha}^T\tilde{D}^T\hat{\mu} = \bar{\alpha}^T\hat{\nu} = 0$, and because $x \in \mathcal{D}$, there exists $\alpha_x \geq 0$ such that $\tilde{K}^Tx = \tilde{D}\alpha_x$, thus $x^T\tilde{K}^T\hat{\mu} = \alpha_x^T\tilde{D}^T\hat{\mu} = \alpha_x^T\hat{\nu} \geq 0$. Therefore, defining

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

we may conclude that

$$\begin{aligned} \|\bar{x} - x\|^2 &\leq (\tilde{A}x - \tilde{b})^T\hat{\lambda} \\ &\leq \|\tilde{A}x - \tilde{b}\|\|\hat{\lambda}\| \\ &\leq \bar{\theta}(\mathcal{M}_{I'})\|\tilde{A}x - \tilde{b}\|\|\bar{x} - x\|, \end{aligned}$$

and thus

$$\|\bar{x} - x\| \leq \bar{\theta}(\mathcal{M}_{I'})\|\tilde{A}x - \tilde{b}\|.$$

And, it can be easily seen that for any subset I of $\{1, \ldots, l\}$,

$$\bar{\theta}(\mathcal{M}_I) \leq \bar{\theta}(\mathcal{M}).$$

Eventually we arrive at the conclusion, i.e.,

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

□

Lemma 8 can be considered as a variant of Hoffman's error bound theory. Applying Lemma 8 to $\widehat{\Gamma}(p)$ in (26), we obtain the following result.

Proposition 12 *Suppose that* $g(x) := \sum_{J\in\mathcal{J}} \omega_J \|x_J\|_2$*, where* $\omega_J \geq 0$ *and* $\mathcal{J}$ *is a partition of* $\{1,\ldots,n\}$*.* $\widehat{\Gamma}(p)$ *is globally calm with modulus* $\bar{\theta}(\mathcal{M})$*, i.e.,*

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

where

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

and $\bar{\theta}(\mathcal{M})$ *is defined as in* (27).

In Theorem 7, using $\kappa_g = \tilde{\kappa}_g$ and $\kappa = \bar{\theta}(\mathcal{M})$ we obtained the calmness modulus for the group LASSO.

Theorem 9 *Consider the group LASSO. That is, f satisfies Assumption 2 and* $g(x) := \sum_{J\in\mathcal{J}} \omega_J \|x_J\|_2$*, where* $\omega_J \geq 0$ *and* $\mathcal{J}$ *is a partition of* $\{1,\ldots,n\}$*, in problem* (1)*. For any given positive number M such that h is strongly convex on* $M\bar{\mathbb{B}}$ *with modulus* σ *and* ∇h *is Lipschitz continuous on* $M\bar{\mathbb{B}}$ *with constant* L_h*, there exists* $\tilde{\kappa}_{gLasso} > 0$ *such that*

$$dist\left(x, (\partial F)^{-1}(0)\right) \leq \tilde{\kappa}_{gLasso} dist\left(0, \partial F(x)\right), \quad \forall \|x\| \leq M.$$

In particular, $\tilde{\kappa}_{gLasso} = \kappa_{gLasso} + 2{c_{gLasso}}^2 + 2c_{gLasso}\sqrt{\kappa_{gLasso} + {c_{gLasso}}^2} > 0$ *with* $c_{gLasso} = \frac{\kappa_{gLasso}(1+\|A\|L_h)}{2\sqrt{\sigma}}$, $\kappa_{gLasso} = (1 + \bar{\theta}(\mathcal{M})\|A\|) \max\{\frac{1}{\tilde{\sigma}_{\min}(A)}, \tilde{\kappa}_g\}$*, where* $\tilde{\kappa}_g$ *is defined as in* (25).

6 Conclusion

In summary, the variational analysis perspective we propose enhances our understanding of linear convergence of some common first order methods. This new perspective allows us to use extensive advanced tools or techniques in variational analysis literatures to investigate algorithmic convergence behaviors.

- The use of calm intersection theorem offers a tight and equivalent characterization of the calmness of Γ, that is, the calmness of $\widehat{\Gamma} := \Gamma_1(p_1) \cap \Gamma_2(0)$. This equivalence has sheded some light on model structures which ensure the required error bound for linear convergence. Metric subregularity/calmness/error bound are all point-based properties. Therefore usually they cannot satisfy everywhere automatically unless the system under consideration is polyhedral. We justify the essential polyhedral structures for popular applications where the Luo-Tseng error bound conditions are known to be satisfied everywhere automatically. When g represents the nuclear norm regularizer, by its nature $\Gamma_2(0)$ and hence $\widehat{\Gamma}$ are lack of polyhedricity. In this regard, we cannot expect that the calmness of Γ is satisfied automatically in this case. Some recently-developed verifiable sufficient conditions for the calmness of $\widehat{\Gamma}$ could be further considered, see, e.g., [18, 21, 31]. By doing so, the sufficient conditions given in [72] which seem to be very restrictive and hence impractical could be significantly improved.
- This new perspective also inspires deeper understanding on the convergence behaviors of some other first order methods not discussed in this paper, for example, the alternating direction method of multipliers (ADMM) and its variants. In particular, in a recent paper [67], using the calm intersection theorem and other techniques in

variational analysis, the point-based metric subregularity/calmness is proved to hold automatically for a wide range of applications arising in statistical learning. Hence the empirically observed linear convergence of a number of algorithms is tightly proved; and the understanding of linear convergence of ADMM and its variants is significantly enhanced.

Appendix

Proof of Theorem 1 Since L_{i_k} is the Lipschitz constant of $\nabla_{i_k} f$, $c_i^k \geq L_i$ and $x_j^{k+1} = x_j^k, \forall j \neq i_k$, we have

$$f(x^{k+1}) \leq f(x^k) + \left\langle \nabla_{i_k} f(x^k), x_{i_k}^{k+1} - x_{i_k}^k \right\rangle + \frac{c_{i_k}^k}{2} \|x_{i_k}^{k+1} - x_{i_k}^k\|^2,$$

which implies that

$$F(x^{k+1}) \leq F(x^k) + \left\langle \nabla_{i_k} f(x^k), x_{i_k}^{k+1} - x_{i_k}^k \right\rangle + \frac{c_{i_k}^k}{2} \|x_{i_k}^{k+1} - x_{i_k}^k\|^2 + g_{i_k}(x_{i_k}^{k+1}) - g_{i_k}(x_{i_k}^k).$$

Combining with the iteration scheme (3), we have

$$F(x^{k+1}) - F(x^k) \leq \min_{t_{i_k}} \left\{ \left\langle \nabla_{i_k} f(x^k), t_{i_k} \right\rangle + \frac{c_{i_k}^k}{2} t_{i_k}^2 + g_{i_k}(x_{i_k}^k + t_{i_k}) - g_{i_k}(x_{i_k}^k) \right\},$$

where $t_{i_k} := x_{i_k} - x_{i_k}^k$. Recall that for a given iteration point x^k, the next iteration point x^{k+1} is obtained by using the scheme (3) where the index i_k is randomly chosen from $\{1, \ldots, N\}$ with the uniform probability distribution. Conditioned on x^k and taking expectation with respect to the random index i_k, we obtain

$$\begin{aligned} &\mathbb{E}[F(x^{k+1}) - F(x^k) \mid x^k] \\ &\leq \mathbb{E} \left\{ \min_{t_{i_k}} \left\langle \nabla_{i_k} f(x^k), t_{i_k} \right\rangle + \frac{c_{i_k}^k}{2} t_{i_k}^2 + g_{i_k}(x_{i_k}^k + t_{i_k}) - g_{i_k}(x_{i_k}^k) \;\middle|\; x^k \right\}. \end{aligned} \tag{29}$$

Now, we are going to estimate the right hand side in the above inequality

$$\begin{aligned} &\mathbb{E} \left\{ \min_{t_{i_k}} \left\langle \nabla_{i_k} f(x^k), t_{i_k} \right\rangle + \frac{c_{i_k}^k}{2} t_{i_k}^2 + g_{i_k}(x_{i_k}^k + t_{i_k}) - g_{i_k}(x_{i_k}^k) \;\middle|\; x^k \right\} \\ &= \frac{1}{N} \sum_{i=1}^N \left\{ \min_{t_i} \left\langle \nabla_i f(x^k), t_i \right\rangle + \frac{c_i^k}{2} t_i^2 + g_i(x_i^k + t_i) - g_i(x_i^k) \right\} \qquad (30) \\ &\leq \frac{1}{N} \min_t \sum_{i=1}^N \left\{ \left\langle \nabla_i f(x^k), t_i \right\rangle + \frac{C}{2} t_i^2 + g_i(x_i^k + t_i) - g_i(x_i^k) \right\} \\ &= \frac{1}{N} \min_y \left\{ \left\langle \nabla f(x^k), y - x^k \right\rangle + \frac{C}{2} \|y - x^k\|^2 + g(y) - g(x^k) \right\} \\ &= \frac{1}{N} (F_C(x^k) - F(x^k)), \end{aligned}$$

where $t := (t_1, \ldots, t_N)$ and $F_C(x) := \min_y \left\{ f(x) + \langle \nabla f(x), y - x \rangle + \frac{C}{2} \|y - x\|^2 + g(y) \right\}$.

Furthermore, we set

$$\begin{aligned}\hat{x}^k &:= (I + \frac{1}{C}\partial g)^{-1}\left(x^k - \frac{1}{C}\nabla f(x^k)\right)\\ &= \arg\min_y \left\{\left\langle \nabla f(x^k), y - x^k\right\rangle + \frac{C}{2}\|y - x^k\|^2 + g(y)\right\}.\end{aligned}$$

It follows immediately that

$$g\left(x^k\right) \geq g\left(\hat{x}_k\right) - \left\langle \nabla f(x^k) + C(\hat{x}_k - x^k), x^k - \hat{x}_k\right\rangle,$$

which yields that,

$$f\left(x^k\right) + g\left(x^k\right) \geq f\left(x^k\right) + \left\langle \nabla f(x^k), \hat{x}_k - x^k\right\rangle + \frac{C}{2}\|\hat{x}_k - x^k\|^2 + g\left(\hat{x}_k\right) + \frac{C}{2}\|\hat{x}_k - x^k\|^2,$$

and hence

$$F(x^k) \geq F_C(x^k) + \frac{C}{2}\|\hat{x}_k - x^k\|^2. \tag{31}$$

Let $\tilde{x} := \mathrm{Proj}_{\mathcal{X}}(x)$ for any x and thus $f(\tilde{x}) + g(\tilde{x}) = F^*$. Then we have

$$\begin{aligned}F_C(x) - F^* &= \min_y \left\{f(x) + \langle \nabla f(x), y - x\rangle + \frac{C}{2}\|y - x\|^2 + g(y)\right\} - f(\tilde{x}) - g(\tilde{x})\\ &\leq f(x) - f(\tilde{x}) + \langle \nabla f(x), \tilde{x} - x\rangle + \frac{C}{2}\|\tilde{x} - x\|^2\\ &\leq \frac{L + C}{2}\|\tilde{x} - x\|^2 = \frac{L + C}{2}\mathrm{dist}\,(x, \mathcal{X})^2,\end{aligned}$$

where L is the Lipschitz constant of ∇f. Plugging $x = x^k$ into the above inequalities, we have

$$\begin{aligned}F_C(x^k) - F^* &\leq \frac{L + C}{2}\mathrm{dist}\left(x^k, \mathcal{X}\right)^2\\ &\leq \frac{\kappa(L + C)}{2}\|\hat{x}_k - x^k\|^2\\ &\leq \kappa(1 + L/C)(F(x^k) - F_C(x^k)),\end{aligned}$$

where the second inequality follows from (14) and the third inequality is a direct consequence of (31). Then we have

$$\begin{aligned}F(x^k) - F^* &= F(x^k) - F_C(x^k) + F_C(x^k) - F^*\\ &\leq (1 + \kappa(1 + L/C))(F(x^k) - F_C(x^k)).\end{aligned} \tag{32}$$

By (29), (30) and (32), we have

$$\mathbb{E}[F(x^{k+1}) - F(x^k) \mid x^k] \leq \frac{1}{N}(F_C(x^k) - F(x^k)) \leq \frac{1}{N} \cdot \frac{1}{1 + \kappa(1 + L/C)}(F^* - F(x^k)),$$

therefore

$$\mathbb{E}[F(x^{k+1}) - F^* \mid x^k] \leq \left(1 - \frac{1}{N(1 + \kappa(1 + L/C))}\right)(F(x^k) - F^*).$$

For any $l \geq 1$, combining the above inequality over $k = 0, 1, \ldots, l - 1$, taking expectation with all the history, we obtain

$$\mathbb{E}[F(x^l) - F^*] \leq \sigma^l(F(x^0) - F^*),$$

where $\sigma = \left(1 - \frac{1}{N(1+\kappa(1+L/C))}\right) \in (0, 1)$, and hence the R-BCPGM achieves a linear convergence rate in terms of the expected objective value. □

Acknowledgments We are grateful to two anonymous referees for their suggestions and comments which have helped us improve the paper substantially.

References

1. Agarwal, A., Negahban, S.N., Wainwright, M.J.: Fast global convergence of gradient methods for high-dimensional statistical recovery, vol. 40 (2012)
2. Aragón Artacho, F.J., Geoffroy, M.H.: Characterization of metric regularity of subdifferentials. J. Convex Anal. **15**, 365–380 (2008)
3. Aragón Artacho, F.J., Geoffroy, M.H.: Metric subregularity of the convex subdifferential in Banach spaces. J. Nonlinear Convex Anal. **15**, 35–47 (2014)
4. Aubin, J.: Lipschitz behavior of solutions to convex minimization problems. Math. Oper. Res. **9**, 87–111 (1984)
5. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. Found. Trends R Mach. Learn. **4**(1), 1–106 (2012)
6. Beck, A.: First-order methods in optimization, vol. 25. SIAM, New Delhi (2017)
7. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. **2**, 183–202 (2009)
8. Bishop, C.M.: Pattern recognition and machine learning. Springer-Verlag, New York (2006)
9. Bolte, J., Nguyen, T.P., Peypouquet, J., Suter, B.W.: From error bounds to the complexity of first-order descent methods for convex functions. Math. Program. **165**, 471–507 (2017)
10. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**, 459–494 (2014)
11. Bondell, H.D., Reich, B.J.: Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. Biometrics **64**, 115–123 (2008)
12. Candès, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted ℓ_1 minimization. J. Fourier Anal. Appl. **14**(5), 877–905 (2008)
13. Dontchev, A.L., Rockafellar, R.T.: Regularity and conditioning of solution mappings in variational analysis. Set-Valued Anal. **12**, 79–109 (2004)
14. Drusvyatskiy, D., Lewis, A.S.: Error bounds, quadratic growth, and linear convergence of proximal methods. Math. Oper. Res. **43**, 919–948 (2018)
15. Facchinei, F., Pang, J.S.: Finite-dimensional variational inequalities and complementarity problems. Springer Science & Business Media, Berlin (2007)
16. Fercoq, O., Richtrik, P.: Optimization in high dimensions via accelerated, parallel, and proximal coordinate descent. SIAM Rev. **58**, 739–771 (2016)
17. Friedman, J., Hastie, T., Tibshirani, R.: A note on the group lasso and a sparse group lasso. arXiv:1606.00269 (2010)
18. Gfrerer, H.: On directional metric regularity, subregularity and optimality conditions for nonsmooth mathematical programs. Set-Valued Variat. Anal. **21**, 151–176 (2013)
19. Güler, O., Hoffman, A.J., Rothblum, U.G.: Approximations to solutions to systems of linear inequalities. SIAM J. Matrix Anal. Appl. **16**, 688–696 (1995)
20. Guo, L., Ye, J.J., Zhang, J.: Mathematical programs with geometric constraints in Banach spaces: enhanced optimality, exact penalty, and sensitivity. SIAM J. Optim. **23**, 2295–2319 (2013)
21. Gfrerer, H., Ye, J.J.: New constraint qualifications for mathematical programs with equilibrium constraints via variational analysis. SIAM J. Optim. **27**, 842–865 (2017)
22. Henrion, R., Jourani, A., Outrata, J.: On the calmness of a class of multifunctions. SIAM J. Optim. **13**, 603–618 (2002)
23. Henrion, R., Outrata, J.: Calmness of constraint systems with applications. Math. Program. **104**, 437–464 (2005)
24. Hoffman, A.J.: On approximate solutions of systems of linear inequalities. J. Research Nat. Bur. Standards **49**, 263–265 (1952)
25. Hong, M., Wang, X., Razaviyayn, M., Luo, Z.Q.: Iteration complexity analysis of block coordinate descent methods. Math. Program. **163**, 85–114 (2017)

26. Karimi, H., Nutini, J., Schmidt, M.: Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In: Joint European conference on machine learning and knowledge discovery in databases, pp. 795–811. Springer (2016)
27. Klatte, D., Kummer, B.: Constrained minima and lipschitzian penalties in metric spaces. SIAM J. Optim. **13**, 619–633 (2002)
28. Klatte, D., Thiere, G.: Error bounds for solutions of linear equations and inequalities. Zeitschrift fü,r Oper. Res. **41**, 191–214 (1995)
29. Li, G., Pong, T.K.: Calculus of the exponent of kurdykaŁojasiewicz inequality and its applications to linear convergence of first-order methods. Found. Comput. Math. **18**, 1199–1232 (2018)
30. Li, X., Zhao, T., Arora, R., Liu, H., Hong, M.: An improved convergence analysis of cyclic block coordinate descent-type methods for strongly convex minimization. Artif. Intell. Stat., 491–499 (2016)
31. Liu, Y.L., Bi, S.J., Pan, S.H.: Several classes of stationary points for rank regularized minimization problems. SIAM J. Optim. **30**(2), 1756–1775 (2020)
32. Luke, D.R., Nguyen, H.T., Tam, M.K.: Quantitative convergence analysis of iterated expansive, set-valued mappings. Math. Oper. Res. **43**, 1143–1176 (2018)
33. Luo, Z.Q., Tseng, P.: On the linear convergence of descent methods for convex essentially smooth minimization. SIAM J. Control. Optim. **30**, 408–425 (1992)
34. Luo, Z.Q., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Ann. Oper. Res. **46**, 157–178 (1993)
35. Martinet, B.: Brève communication régularisation d'inéquations variationnelles par approximations successives. Revue française d'informatique et de Recherche Opérationnelle, Sé,rie Rouge **4**, 154–158 (1970)
36. Mordukhovich, B.: Variational analysis and generalized differentiation i: basic theory, II: applications. Springer Science & Business Media, Berlin (2006)
37. Necoara, I., Clipici, D.: Efficient parallel coordinate descent algorithm for convex optimization problems with separable constraints: Application to distributed MPC, vol. 23 (2013)
38. Necoara, I., Clipici, D.: Parallel random coordinate descent method for composite minimization: Convergence analysis and error bounds. SIAM J. Optim. **26**, 197–226 (2016)
39. Necoara, I., Nesterov, Y., Glineur, F.: Linear convergence of first order methods for non-strongly convex optimization. Math. Program. **175**, 69–107 (2019)
40. Necoara, I., Nesterov, Y., Glineur, F.: Random block coordinate descent methods for linearly constrained optimization over networks. J. Optim. Theory Appl. **173**, 227–254 (2017)
41. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J. Optim. **22**, 341–362 (2012)
42. Nesterov, Y.: Introductory lectures on convex optimization. Kluwer Academic, Dordrecht (2004)
43. O'donoghue, B., Candès, E.: Adaptive restart for accelerated gradient schemes. Found. Comput. Math. **15**, 715–732 (2015)
44. Passty, G.: Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. J. Math. Anal. Appl. **72**, 383–390 (1979)
45. Peña, J., Vera, J.C., Zuluaga, L.F.: An algorithm to compute the Hoffman constant of a system of linear constraints. arXiv:1804.08418 (2018)
46. Peña, J., Vera, J.C., Zuluaga, L.F.: New characterizations of Hoffman constants for systems of linear constraints. Math. Prog. (2020)
47. Polyak, B.T.: Introduction to optimization, optimization software incorporation. Publications Division, New York (1987)
48. Richtárik, P., Takáč, M.: Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Math. Program. **144**, 1–38 (2014)
49. Robinson, S.M.: Stability theory for systems of inequalities. Part i: Linear systems. SIAM J. Numer. Anal. **12**, 754–769 (1975)
50. Robinson, S.M.: An implicit-function theorem for generalized variational inequalities Technical report (WISCONSIN UNIV MADISON MATHEMATICS RESEARCH CEN- TER, 1976)
51. Robinson, S.M.: Some continuity properties of polyhedral multifunctions. Math. Program. Study **14**, 206–214 (1981)
52. Rockafellar, R.T.: Convex analysis. Princeton University Press, Princeton (1970)
53. Rockafellar, R.T., Wets, R.: Variational analysis. Springer Science & Business Media, Berlin (2009)
54. Shefi, R., Teboulle, M.: On the rate of convergence of the proximal alternating linearized minimization algorithm for convex problems. EURO J. Comput. Optim. **4**, 27–46 (2016)
55. Schmidt, M., Roux, N., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. Adv. Neural Inf. Process. Sys. **24**, 1458–1466 (2011)

56. Stoer, J., Witzgall, C.: Convexity and optimization in finite dimensions I. Springer Science & Business Media, Berlin (2012)
57. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Royal Stat. Soc. Series B (Methodological) **73**, 267–288 (1996)
58. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. J. Royal Stat. Soc. Ser. B (Statistical Methodology) **67**, 91–108 (2005)
59. Tseng, P.: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Math. Program. **125**, 263–295 (2010)
60. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Math. Program. **117**, 387–423 (2009)
61. Wang, P.W., Lin, C.J.: Iteration complexity of feasible descent methods for convex optimization. J. Mach. Learn. Res. **15**, 1523–1548 (2014)
62. Wang, X., Ye, J.J., Yuan, X., Zeng, S., Zhang, J.: Perturbation techniques for convergence analysis of proximal gradient method and other first-order algorithms via variational analysis. Set-Valued Variat. Anal. (2021)
63. Xiao, L., Zhang, T.: A proximal-gradient homotopy method for the sparse least-squares problem. SIAM J. Optim. **23**, 1062–1091 (2013)
64. Ye, J.J., Ye, X.Y.: Necessary optimality conditions for optimization problems with variational inequality constraints. Math. Oper. Res. **22**, 977–997 (1997)
65. Ye, J.J., Zhou, J.C.: Verifiable sufficient conditions for the error bound property of second-order cone complementarity problems. Math. Program. **171**, 361–395 (2018)
66. Yuan, M., Lin, Y.: Model selection and estimation in regression with group variables. J. Royal Stat. Soc. Series B (Statistical Methodology) **68**, 49–67 (2006)
67. Yuan, X., Zeng, S., Zhang, J.: Discerning the linear convergence of ADMM for structured convex optimization through the lens of variational analysis. J. Mach. Learn. Res. **21**, 1–75 (2020)
68. Zhang, H.: New analysis of linear convergence of gradient-type methods via unifying error bound conditions. Math. Program. **180**(1), 371–416 (2020)
69. Zhang, H., Jiang, J., Luo, Z.Q.: On the linear convergence of a proximal gradient method for a class of nonsmooth convex minimization problems. J. Oper. Res. Soc. China **1**, 163–186 (2013)
70. Zhang, S.: Global error bounds for convex conic problems. SIAM J. Optim. **10**, 836–851 (2000)
71. Zheng, X.Y., Ng, K.F.: Metric subregularity of piecewise linear multifunctions and applications to piecewise linear multiobjective optimization. SIAM J. Optim. **24**, 154–174 (2014)
72. Zhou, Z., So, A.M.-C.: A unified approach to error bounds for structured convex optimization problems. Math. Program. **165**, 689–728 (2017)
73. Zhou, Z., Zhang, Q., So, A.M.-C.: $L_{1,p}$-norm regularization: error bounds and convergence rate analysis of first-order methods. In: International conference on machine learning, pp. 1501–1510 (2015)

Affiliations

Jane J. Ye[1] · Xiaoming Yuan[2] · Shangzhi Zeng[2] · Jin Zhang[3]

Jane J. Ye
janeye@uvic.ca

Shangzhi Zeng
zengsz@connect.hku.hk

Jin Zhang
zhangj9@sustech.edu.cn

[1] Department of Mathematics and Statistics, University of Victoria, Victoria, BC V8P 5C2, Canada

[2] Department of Mathematics, The University of Hong Kong, Hong Kong, China

[3] Department of Mathematics, SUSTech International Center for Mathematics, Southern University of Science and Technology, National Center for Applied Mathematics Shenzhen, Shenzhen, China