Journal of Machine Learning Research 25 (2024) 1-48 Submitted 12/23; Revised 9/24; Published 11/24

An Inexact Projected Regularized Newton Method for Fused Zero-norms Regularization Problems

Yuqia Wu YUQIA.WU@CONNECT.POLYU.HK
Department of Applied Mathematics
The Hong Kong Polytechnic University
Kowloon, Hong Kong

Shaohua Pan SHHPAN@SCUT.EDU.CN
School of Mathematics
South China University of Technology
Guangzhou, China

Xiaoqi Yang* MAYANGXQ@POLYU.EDU.HK
Department of Applied Mathematics
The Hong Kong Polytechnic University
Kowloon, Hong Kong

Editor: Silvia Villa

Abstract

This paper concerns structured ℓ_0-norms regularization problems, with a twice continuously differentiable loss function and a box constraint. This class of problems have a wide range of applications in statistics, machine learning and image processing. To the best of our knowledge, there is no efficient algorithm in the literature for solving them. In this paper, we first provide a polynomial-time algorithm to find a point in the proximal mapping of the fused ℓ_0-norms with a box constraint based on dynamic programming principle. We then propose a hybrid algorithm of proximal gradient method and inexact projected regularized Newton method to solve structured ℓ_0-norms regularization problems. The iterate sequence generated by the algorithm is shown to be convergent by virtue of a non-degeneracy condition, a curvature condition and a Kurdyka-Łojasiewicz property. A superlinear convergence rate of the iterates is established under a locally Hölderian error bound condition on a second-order stationary point set, without requiring the local optimality of the limit point. Finally, numerical experiments are conducted to highlight the features of our considered model, and the superiority of our proposed algorithm.

Keywords: fused ℓ_0-norms regularization problems; inexact projected regularized Newton algorithm; global convergence; superlinear convergence; KL property.

1. Introduction

Given a matrix $B \in \mathbb{R}^{p\times n}$, parameters $\lambda_1 > 0$ and $\lambda_2 > 0$, and vectors $l \in \mathbb{R}^n_-$ and $u \in \mathbb{R}^n_+$, we are interested in the structured ℓ_0-norms regularization problem with a box constraint:

$$\min_{x\in\mathbb{R}^n} \; F(x) := f(x) + \lambda_1\|Bx\|_0 + \lambda_2\|x\|_0 \;\; \text{s.t.} \;\; l \le x \le u, \tag{1}$$

∗. Corresponding author.

where $f : \mathbb{R}^n \to \overline{\mathbb{R}} := (-\infty, \infty]$ is twice continuously differentiable on an open set $\mathcal{O}$ containing the box set $\Omega := \{x \in \mathbb{R}^n \mid l \leq x \leq u\}$, and $\|\cdot\|_0$ denotes the ℓ_0-norm (or cardinality) function. This model encourages sparsity of both variable x and its linear transformation Bx. Throughout this paper, we write $g(\cdot) := \lambda_1\|B\cdot\|_0 + \lambda_2\|\cdot\|_0 + \delta_\Omega(\cdot)$, where $\delta_\Omega(\cdot)$ denotes the indicator function of Ω.

1.1 Motivation

Given a data matrix $A \in \mathbb{R}^{m\times n}$ and its response $b \in \mathbb{R}^m$, the common regression model is to minimize $f(x) := h(Ax - b)$, where $h: \mathbb{R}^m \to \overline{\mathbb{R}}$ is continuously differentiable on $A(\mathcal{O}) - b$ with its minimum attained at the origin. When $h(\cdot) = \frac{1}{2}\|\cdot\|^2$, f is the least-squares loss function of the linear regression. It is known that one of the popular models for seeking a sparse vector while minimizing f is the following ℓ_0-norm regularization problem

$$\min_{x\in\mathbb{R}^n} \; f(x) + \lambda_2\|x\|_0, \tag{2}$$

where the ℓ_0-norm term is used to identify a set of influential components by shrinking some small coefficients to 0. However, the ℓ_0-norm regularizer only takes the sparsity of x into consideration, but ignores its spatial nature, which sometimes needs to be considered in real-world applications. For example, in the context of image processing, the variables often represent the pixels of images, which are correlated with their neighboring ones. To recover the blurred images, Rudin et al. (1992) took into account the differences between adjacent variables and used the total variation regularization, which penalizes the changes of the neighboring pixels and hence encourages smoothness in the solution. In addition, Land and Friedman (1997) studied the phoneme classification on TIMIT database, for which there is a high chance that every sampled point is close or identical to its neighboring ones because each phoneme is composed of a series of consecutively sampled points. Land and Friedman (1997) considered imposing a fused penalty on the coefficients vector x, and proposed the following models with zero-order variable fusion and first-order variable fusion respectively to train the classifier:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \lambda_1\|\widehat{B}x\|_0, \tag{3}$$

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \lambda_1\|\widehat{B}x\|_1, \tag{4}$$

where $A \in \mathbb{R}^{m\times n}$ represents the phoneme data, $b \in \mathbb{R}^m$ is the label vector, $\widehat{B} \in \mathbb{R}^{(n-1)\times n}$ with $\widehat{B}_{ii} = 1$ and $\widehat{B}_{i,i+1} = -1$ for all $i \in \{1, \ldots, n-1\}$ and $\widehat{B}_{ij} = 0$ otherwise. In the sequel, we call (1) with $f(\cdot) = \frac{1}{2}\|A\cdot - b\|^2$ and $B = \widehat{B}$ a fused ℓ_0-norms regularization problem with a box constraint.

Additionally taking the sparsity of x into consideration, Tibshirani et al. (2005) proposed the fused Lasso, given by

$$\min_{x\in\mathbb{R}^n} \; \frac{1}{2}\|Ax - b\|^2 + \lambda_1\|\widehat{B}x\|_1 + \lambda_2\|x\|_1, \tag{5}$$

and presented its nice statistical properties. Friedman et al. (2007) demonstrated that the proximal mapping of the function $\lambda_1\|\widehat{B}\cdot\|_1 + \lambda_2\|\cdot\|_1$ can be obtained through a process, which is known as "prox-decomposition" later. Based on the accessibility of this proximal mapping, various efficient algorithms were proposed to address model (5), see Liu et al. (2009, 2010); Li et al. (2018); Molinari et al. (2019). In particular, Li et al. (2018) proposed a semismooth Newton augmented Lagrangian method (SSNAL) to solve the dual of (5). The numerical results reported in their paper indicate that SSNAL is highly efficient.

It was claimed in Land and Friedman (1997) that both (3) and (4) perform well in signal regression, but the zero-order fusion model (3) produces simpler estimated coefficient vectors. This observation suggests that model (1) with $f = \frac{1}{2}\|A\cdot - b\|^2$ and $B = \widehat{B}$ may be able to effectively find a simpler solution while perform as well as the fused Lasso does. Compared with regularization problems with ℓ_0-norm, those using $\|Bx\|_0$ regularization remain less explored in terms of algorithm development. According to Land and Friedman (1997), the global optimal solution of (3) is unavailable. However, one of its stationary points can be obtained. In fact, Jewell et al. (2020) has revealed by virtue of dynamic programming principle that a point in the proximal mapping of $\lambda_1\|\widehat{B}\cdot\|_0$ can be exactly determined within polynomial time, which allows one to use the well-known proximal gradient (PG) method to find a stationary point of problem (3). However, the highly nonconvex and nonsmooth nature of model (1) poses significant challenges in computing the proximal mapping of g when $B = \widehat{B}$ and in developing effective optimization algorithms to solve it. As far as we know, no specific algorithms have yet been designed to solve these challenging problems.

Another motivation for this work comes from our previous research (Wu et al. (2023)). In that work, we considered the model (2) with the ℓ_0-norm replaced by the ℓ_q quasi-norm $\|x\|_q^q$, where $q \in (0,1)$ and $\|x\|_q := \left(\sum_{i=1}^n |x_i|^q\right)^{1/q}$. For this class of nonconvex and nonsmooth problems, we proposed a hybrid of PG and subspace regularized Newton methods (HpgSRN), which restricts the subproblems of Newton steps on a subspace within which their objective functions are smooth, and thus a regularized Newton method can be applied. It is worth noting that the subspace is induced by the support of the current iterate x^k. PG step is executed in every iteration, but it does not necessarily run a Newton step unless a switch condition is satisfied. The full convergence of the iterate sequence was established under a curvature condition and the Kurdyka-Łojasiewicz (KL) property (Attouch et al. (2010)) of the objective function, and a superlinear convergence rate was achieved under an additional local error bound condition on a second-order stationary point set. Due to the desirable convergence result and numerical performance of HpgSRN, we aim to adopt a similar subspace regularized Newton algorithm to solve (1), in which the subspace is induced by the combined support of Bx^k and x^k.

1.2 Related work

In recent years, many optimization algorithms have been well developed to solve the ℓ_0-norm regularization problems of the form (2), which includes iterative hard thresholding (Herrity et al. (2006); Blumensath and Davies (2008, 2010); Lu (2014)), the penalty decomposition (Lu and Zhang (2013)), the smoothing proximal gradient method (Bian and Chen (2020)), the accelerated iterative hard thresholding (Wu and Bian (2020)) and NL0R (Zhou et al. (2021)). Among all these algorithms, NL0R is the only Newton-type method, which employs Newton method to solve a series of stationary point equations confined to the subspaces identified by the support of the solution obtained by the proximal mapping of $\lambda_2\|\cdot\|_0$.

The PG method is able to effectively cope with model (1) if the proximal mapping of g can be exactly computed. The PG method belongs to first-order methods, which have a low computation cost and require weak global convergence conditions, but they achieve at most a linear convergence rate. On the other hand, the Newton method has a faster convergence rate, but it can only be applied to minimize sufficiently smooth objective functions. In recent years, there have been active investigations into the Newton-type methods for nonsmooth composite optimization problems of the form

$$\min_{x\in\mathbb{R}^n}\ \Psi(x) := \psi(x) + \varphi(x), \tag{6}$$

where $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper lower semicontinuous, and ψ is twice continuously differentiable on an open subset of $\mathbb{R}^n$ containing the domain of φ. The proximal Newton-type method is able to address (6) with a convex or weakly convex φ, and a convex ψ (see Bertsekas (1982); Lee et al. (2014); Yue et al. (2019); Mordukhovich et al. (2023)) or nonconvex ψ (Liu et al. (2024)). Another popular second-order method for solving (6) is to minimize the forward-backward envelop (FBE) of Ψ, see Stella et al. (2017); Themelis et al. (2018, 2019); Ahookhosh et al. (2021). In particular, for those Ψ with the proximal mapping of φ being available, Themelis et al. (2018) proposed an algorithm called ZeroFPR, based on the quasi-Newton method, for minimizing the FBE of Ψ. They achieved the global convergence of the iterate sequence by means of the KL property of the FBE and its local superlinear rate under the Dennis-Moré condition and the strong local minimality property of the limit point. An algorithm similar to ZeroFPR but minimizing the Bregman FBE of Ψ was proposed in Ahookhosh et al. (2021), which achieves a superlinear convergence rate without requiring the strong local minimality of the limit point. For the case that ψ is smooth and φ admits a computable proximal mapping, Bareilles et al. (2023) proposed an algorithm, alternating between a PG step and a Riemannian Newton method, which was proved to have a quadratic convergence rate under a positive definiteness assumption on the Riemannian Hessian at the limit point.

1.3 Main contributions

This work aims to design a hybrid of PG and inexact projected regularized Newton methods (PGiPN) to solve the structured ℓ_0-norms regularization problem (1). Let $x^k \in \Omega$ be the current iterate. Our method first runs a PG step with line search at x^k to produce $\overline{x}^k$ via

$$\overline{x}^k \in \text{prox}_{\overline{\mu}_k^{-1} g}(x^k - \overline{\mu}_k^{-1}\nabla f(x^k)), \tag{7}$$

where $\text{prox}_{\overline{\mu}_k^{-1} g}(\cdot)$ is the proximal mapping of g, $\overline{\mu}_k > 0$ is a constant such that the objective function F of (1) gains a sufficient decrease from x^k to $\overline{x}^k$, and then judges whether the iterate x^k enters Newton step or not in terms of some switch condition, which takes the following forms of structured stable supports:

$$\text{supp}(x^k) = \text{supp}(\overline{x}^k) \ \text{ and } \ \text{supp}(Bx^k) = \text{supp}(B\overline{x}^k). \tag{8}$$

If this switch condition does not hold, we set $x^{k+1} = \overline{x}^k$ and return to the PG step. Otherwise, by the nature of ℓ_0-norm, the restriction of the function $x \mapsto \lambda_1\|Bx\|_0 + \lambda_2\|x\|_0$

on the supports $\text{supp}(Bx^k)$ and $\text{supp}(x^k)$, i.e., $\lambda_1\|(Bx)_{\text{supp}(Bx^k)}\|_0 + \lambda_2\|x_{\text{supp}(x^k)}\|_0$, is a constant near x^k and does not provide any useful information. In this case, unlike dealing with the ℓ_q-norm regularization problem in Wu et al. (2023), we introduce the following multifunction $\Pi : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$:

$$\begin{aligned}\Pi(z) &:= \{x \in \Omega \mid \text{supp}(x) \subset \text{supp}(z),\ \text{supp}(Bx) \subset \text{supp}(Bz)\}\\ &= \left\{x \in \Omega \mid x_{[\text{supp}(z)]^c} = 0,\ (Bx)_{[\text{supp}(Bz)]^c} = 0\right\},\end{aligned} \tag{9}$$

and consider the associated subproblem

$$\min_{x\in\mathbb{R}^n} f(x) + \delta_{\Pi_k}(x) \quad \text{with} \quad \Pi_k = \Pi(x^k). \tag{10}$$

It is noted that the set $\Pi(x^k)$ containing all the points whose supports are a subset of the support of x^k as well as the supports of their linear transformation is a subset of the support of the linear transformation of x^k. It is worth pointing out that the multifunction Π is not closed but closed-valued.

We will show that every stationary point of (10) is one for problem (1). Thus, instead of a subspace regularized Newton step in Wu et al. (2023), following the projected Newton method in Bertsekas (1982) and the proximal Newton method in Lee et al. (2014); Yue et al. (2019); Mordukhovich et al. (2023) and Liu et al. (2024), our projected regularized Newton step minimizes the following second-order approximation of (10) on Π_k:

$$\arg\min_{x\in\mathbb{R}^n} \Theta_k(x) := f(x^k) + \langle \nabla f(x^k), x - x^k\rangle + \frac{1}{2}\langle x - x^k, G_k(x - x^k)\rangle + \delta_{\Pi_k}(x). \tag{11}$$

Among others, G_k in (11) is an approximation to the Hessian $\nabla^2 f(x^k)$ satisfying

$$G_k \succeq b_1\|\overline{\mu}_k(x^k - \overline{x}^k)\|^\sigma I, \tag{12}$$

where $b_1 > 0$, $\sigma \in (0, \frac{1}{2})$ and $\overline{\mu}_k$ is the same as in (7). To cater for the practical computation, our Newton step seeks an inexact solution y^k of (11) satisfying

$$\begin{cases} \Theta_k(y) - \Theta_k(x^k) \le 0, & (13)\\ \text{dist}(0, \partial\Theta_k(y)) \le \dfrac{\min\{\overline{\mu}_k^{-1}, 1\}}{2} \min\left\{\|\overline{\mu}_k(x^k - \overline{x}^k)\|, \|\overline{\mu}_k(x^k - \overline{x}^k)\|^{1+\varsigma}\right\} & (14)\end{cases}$$

with $\varsigma \in (\sigma, 1]$. Set the direction $d^k := y^k - x^k$. A step size $\alpha_k \in (0, 1]$ is found in the direction d^k via backtracking, and let $x^{k+1} := x^k + \alpha_k d^k$. To ensure the global convergence, the next iteration still returns to the PG step. The details of the algorithm are given in Section 3.

The main contributions of the paper are as follows:

• Based on dynamic programming principle, we develop a polynomial-time algorithm in time $O(n^{3+\epsilon})$ with any $\epsilon > 0$ for seeking a point $\overline{x}_k$ in the proximal mapping (7) of g with $B = \widehat{B}$. This generalizes the corresponding result in Jewell et al. (2020) for finding $\overline{x}^k$ in (7) from $g(\cdot) = \lambda_1\|B\cdot\|_0$ to $g(\cdot) = \lambda_1\|B\cdot\|_0 + \lambda_2\|\cdot\|_0 + \delta_\Omega(\cdot)$ with $B = \widehat{B}$, and also provides the core of PG algorithms for solving (1). We also establish a uniform lower bound for

$\mathrm{prox}_{\mu^{-1}g}(\cdot)$ with μ from a closed interval on a compact set. This plays a crucial role in the convergence analysis of the proposed algorithm, as well as generalizes the corresponding results in Lu (2014) for ℓ_0-norm and in Wu et al. (2023) for ℓ_q-norm with $0 < q < 1$, respectively.

• We design a hybrid algorithm (PGiPN) of PG and inexact projected regularized Newton methods to solve the structured ℓ_0-norms regularization problem (1), which includes the fused ℓ_0-norms regularization problem with a box constraint as a special case. We obtain the global convergence of the algorithm by showing that the structured stable supports (8) hold when the iteration number is sufficiently large. Moreover, we establish a superlinear convergence rate under a Hölderian error bound on a second-order stationary point set, without requiring the local optimality of the limit point.

• The numerical experiments show that our PGiPN is more effective than some existing algorithms in the literature in terms of solution quality and running time.

The rest of the paper is organized as follows. In Section 2 we recall some preliminary knowledge and characterize the stationary point condition of model (1). In Section 3, we prove the prox-regularity of g, characterize a uniform lower bound of the proximal mapping of g, and provide an algorithm for finding a point in the proximal mapping of g with $B = \widehat{B}$. In Section 4, we introduce our algorithm and show that it is well defined. Section 5 is devoted to the convergence analysis of the proposed algorithm. The implementation details of our algorithm and the numerical experiments are included in Section 6.

1.4 Notation

Throughout this paper, $\mathbb{B}(x,\epsilon) := \{z \mid \|z - x\| \le \epsilon\}$ denotes the ball centered at x with radius $\epsilon > 0$, and $\mathbf{B} := \mathbb{B}(0,1)$. Let I and $\mathbf{1}$ be an identity matrix and a vector of all ones, respectively, whose dimension is known from the context. For any two integers $0 \le j < k$, define $[j:k] := \{j, j+1, \ldots, k\}$ and $[k] := [1:k]$. For a closed and convex set $\Xi \subset \mathbb{R}^n$, $\mathrm{ri}(\Xi)$ denotes the relative interior of Ξ, $\mathrm{proj}_{\Xi}(\cdot)$ represents the projection operator onto Ξ, and for a given $x \in \Xi$, $\mathcal{N}_{\Xi}(x)$ and $\mathcal{T}_{\Xi}(x)$ denote the normal cone and tangent cone of Ξ at x, respectively. For a closed set $\Xi' \subset \mathbb{R}^n$, $\mathrm{dist}(z, \Xi') := \min_{x \in \Xi'} \|x - z\|$. For an index set $T \subset [n]$, $|T|$ means the number of the elements of T and write $T^c := [n] \backslash T$. For $t \in \mathbb{R}$, $\mathrm{sign}(t)$ denotes the sign of t, i.e., $\mathrm{sign}(0) = 0$ and $\mathrm{sign}(t) = t/|t|$ for $t \neq 0$, and $t_+ := \max\{t, 0\}$. For a given $x \in \mathbb{R}^n$, $\mathrm{supp}(x) := \{i \in [n] \mid x_i \neq 0\}$, $\mathrm{sign}(x)$ denotes the vector with $[\mathrm{sign}(x)]_i = \mathrm{sign}(x_i)$, $|x|_{\min} := \min_{i \in \mathrm{supp}(x)} |x_i|$. For a vector $x \in \mathbb{R}^n$ and an index set $T \subset [n]$, $x_T \in \mathbb{R}^{|T|}$ is the vector obtained by removing those x_j's with $j \notin T$, and $x_{j:k}$ means $x_{[j:k]}$. Given a real symmetric matrix H, $\lambda_{\min}(H)$ denotes the smallest eigenvalue of H, and $\|H\|_2$ is the spectral norm of H. For a matrix $A \in \mathbb{R}^{m \times n}$ and $S \subset [m]$ (resp. $T \subset [n]$), $A_{S\cdot}$ (resp. $A_{\cdot T}$) denotes the matrix obtained by removing those rows (resp. columns) of A whose indices are not in S (resp. T). For a proper lower semicontinuous function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$, its domain is denoted by $\mathrm{dom}\, h := \{x \in \mathbb{R}^n \mid h(x) < \infty\}$, and its proximal mapping of h associated with a parameter $\mu > 0$ is defined as

$$\mathrm{prox}_{\mu h}(z) := \mathop{\arg\min}_{x \in \mathbb{R}^n} \left\{ \frac{1}{2\mu} \|x - z\|^2 + h(x) \right\} \quad \forall z \in \mathbb{R}^n. \tag{15}$$

For a nonnegative real number sequence $\{a_n\}$, $O(a_n)$ represents a sequence such that $O(a_n) \le c_1 a_n$ for some $c_1 > 0$. The symbol $\mathcal{F}: \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ means that $\mathcal{F}$ is a set-valued mapping (or multifunction), i.e., its image at every point is a set.

2. Preliminaries

Note that the structured ℓ_0-norms function is lower semicontinuous and problem (1) involves a compact box constraint, so its set of global optimal solutions is nonempty and compact. Moreover, the continuity of $\nabla^2 f$ on an open set containing Ω and the compactness of Ω implies that ∇f is Lipschitz continuous on Ω, i.e., there exists $L_{\nabla f} > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \le L_1 \|x - y\| \ \text{ for all } x, y \in \Omega. \tag{16}$$

The above basic facts are often used in the subsequent sections.

2.1 Stationary conditions

For an extended real-valued $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $\overline{x} \in \operatorname{dom} h$, we denote the regular subdifferential of h at $\overline{x}$ by $\widehat{\partial} h(\overline{x})$, and the general subdifferential of h at $\overline{x}$ by $\partial h(\overline{x})$ (Rockafellar and Wets, 2009, Definition 8.3). Now we introduce two classes of stationary points for the general composite problem (6), which includes (1) as a special case.

Definition 1 *A vector $x \in \mathbb{R}^n$ is called a stationary point of problem* (6) *if $0 \in \partial \Psi(x)$. A vector $x \in \mathbb{R}^n$ is called an L-stationary point of problem* (6) *if there exists a constant $\mu > 0$ such that $x \in \operatorname{prox}_{\mu^{-1}\varphi}(x - \mu^{-1}\nabla \psi(x))$.*

Recall that $\Psi = \psi + \varphi$, where ψ is twice continuously differentiable and φ is proper and lower semicontinuous. If in addition φ is assumed to be convex, then

$$0 \in \partial \Psi(x) \Leftrightarrow 0 \in \mu(x - (x - \mu^{-1}\nabla \psi(x))) + \partial \varphi(x) \Leftrightarrow x = \operatorname{prox}_{\mu^{-1}\varphi}(x - \mu^{-1}\nabla \psi(x)).$$

This means that x is a stationary point of problem (6) if and only if x is an L-stationary point. To extend this equivalence to the class of prox-regular functions, we need to recall the definition of prox-regularity, which acts as a surrogate of local convexity.

Definition 2 *(Rockafellar and Wets, 2009, Definition 13.27) A function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is prox-regular at a point $\overline{x} \in \operatorname{dom} h$ for $\overline{v} \in \partial h(\overline{x})$ if h is locally lower semicontinuous at $\overline{x}$, and there exist $r \ge 0$ and $\varepsilon > 0$ such that $h(x') \ge h(x) + v^\top (x' - x) - \frac{r}{2}\|x' - x\|^2$ for all $\|x' - \overline{x}\| \le \varepsilon$, whenever $v \in \partial h(x)$, $\|v - \overline{v}\| < \varepsilon$, $\|x - \overline{x}\| < \varepsilon$ and $h(x) < h(\overline{x}) + \varepsilon$. If h is prox-regular at $\overline{x}$ for all $\overline{v} \in \partial h(\overline{x})$, we say that h is prox-regular at $\overline{x}$.*

The following proposition reveals that under the prox-regularity of φ, the set of stationary points of problem (6) coincides with that of its L-stationary points. Since the proof is similar to that in (Wu et al., 2023, Remark 2.5), the details are omitted here.

Proposition 3 *If $\overline{x}$ is an L-stationary point of problem* (6)*, then $0 \in \partial \Psi(\overline{x})$. If φ is prox-regular at $\overline{x}$ for $-\nabla \psi(\overline{x})$ and prox-bounded*[1]*, the converse is also true.*

1. For the definition of prox-boundedness, see (Rockafellar and Wets, 2009, Definitions 1.23).

Next we provide the stationary point conditions of problem (1) by characterizing the subdifferential of function F. The closed-valuedness of the multifunction Π in (9) is used.

Lemma 4 *Consider any $z \in \Omega$. The following statements are true.*

(i) *$z \in \Pi(z)$, and $\widehat{\partial} g(z) = \partial g(z) = \mathcal{N}_{\Pi(z)}(z)$.*

(ii) *$\partial F(z) = \nabla f(z) + \partial g(z) = \nabla f(z) + \mathcal{N}_{\Pi(z)}(z)$.*

(iii) *for any $x \in \Omega$, $0 \in \nabla f(x) + \mathcal{N}_{\Pi(z)}(x)$ implies that $0 \in \partial F(x)$.*

Proof **(i)** Clearly, $z \in \Pi(z)$. We first argue that $\widehat{\partial} g(z) \subset \mathcal{N}_{\Pi(z)}(z)$. Let $h(x) := \lambda_1 \|Bx\|_0 + \lambda_2 \|x\|_0$. Pick any $v \in \widehat{\partial} g(z)$. By invoking (Rockafellar and Wets, 2009, Definition 8.3),

$$
\begin{aligned}
0 &\le \liminf_{z \neq y \to z} \frac{h(y) + \delta_\Omega(y) - h(z) - \delta_\Omega(z) - \langle v, y - z \rangle}{\|y - z\|} \\
&\le \liminf_{z \neq y \to z, y \in \Pi(z)} \frac{h(y) + \delta_\Omega(y) - h(z) - \delta_\Omega(z) - \langle v, y - z \rangle}{\|y - z\|} \\
&= \liminf_{z \neq y \to z, y \in \Pi(z)} \frac{-\langle v, y - z \rangle}{\|y - z\|} = - \limsup_{z \neq y \to z, y \in \Pi(z)} \frac{\langle v, y - z \rangle}{\|y - z\|},
\end{aligned}
$$

which by (Rockafellar and Wets, 2009, Definition 6.3) implies that $v \in \widehat{\mathcal{N}}_{\Pi(z)}(z)$. From the arbitrariness of $v \in \widehat{\partial} g(z)$ and the convexity of $\Pi(z)$, we conclude that $\widehat{\partial} g(z) \subset \mathcal{N}_{\Pi(z)}(z)$. Next we prove that $\partial g(z) \subset \mathcal{N}_{\Pi(z)}(z)$. Pick any $v \in \partial g(z)$, there exists $z^k \xrightarrow{g} z$ and $v^k \in \widehat{\partial} g(z^k)$ with $v^k \to v$ as $k \to \infty$. As $\mathrm{supp}(Bz^k) \supset \mathrm{supp}(Bz)$ and $\mathrm{supp}(z^k) \supset \mathrm{supp}(z)$ as $k \to \infty$, we deduce from $z^k \xrightarrow{g} z$ that $\Pi(z^k) = \Pi(z)$. Therefore, $v^k \in \mathcal{N}_{\Pi(z^k)}(z^k)$ and $v \in \mathcal{N}_{\Pi(z)}(z)$, which yields the desired inclusion.

Let $h_1(x) := \lambda_1 \|Bx\|_0 + \delta_\Omega(x)$ and $h_2(x) := \lambda_2 \|x\|_0$ for $x \in \mathbb{R}^n$. From (Pan et al., 2023, Lemma 2.2 (iii)), $\partial h_1(z) = \widehat{\partial} h_1(z) = \mathrm{Range}(B^\top_{[\mathrm{supp}(z)]^c \cdot}) + \mathcal{N}_\Omega(z)$ and $\partial h_2(z) = \widehat{\partial} h_2(z) = \{v \in \mathbb{R}^n \,|\, \mathrm{supp}(v) \subset [\mathrm{supp}(z)]^c\}$. As $g = h_1 + h_2$, by the definition of regular subdifferential,

$$
\partial h_1(z) + \partial h_2(z) = \widehat{\partial} h_1(z) + \widehat{\partial} h_2(z) \subset \widehat{\partial} g(z) \subset \partial g(z) \subset \mathcal{N}_{\Pi(z)}(z).
$$

Let $\Pi^1(z) := \{x \in \mathbb{R}^n \,|\, \mathrm{supp}(Bx) \subset \mathrm{supp}(Bz)\}$ and $\Pi^2(z) := \{x \in \mathbb{R}^n \,|\, \mathrm{supp}(x) \subset \mathrm{supp}(z)\}$. Observe that $\Pi^1(z)$ and $\Pi^2(z)$ are the subspaces with $\mathcal{N}_{\Pi^1(z)}(z) = \mathrm{Range}(B^\top_{[\mathrm{supp}(Bz)]^c \cdot})$ and $\mathcal{N}_{\Pi^2(z)}(z) = \{v \in \mathbb{R}^n \,|\, \mathrm{supp}(v) \subset [\mathrm{supp}(z)]^c\}$. Along with the above arguments, we have

$$
\mathcal{N}_{\Pi^1(z)}(z) + \mathcal{N}_{\Pi^2(z)}(z) + \mathcal{N}_\Omega(z) = \partial h_1(z) + \partial h_2(z) \subset \widehat{\partial} g(z) \subset \partial g(z) \subset \mathcal{N}_{\Pi(z)}(z).
$$

Since $\Pi(z) = \Omega \cap \Pi^1(z) \cap \Pi^2(z)$ and $z \in \Pi(z)$, by (Rockafellar, 1970, Theorem 23.8), $\mathcal{N}_{\Pi(z)}(z) = \mathcal{N}_\Omega(z) + \mathcal{N}_{\Pi^1(z)}(z) + \mathcal{N}_{\Pi^2(z)}(z)$. Thus, the desired conclusion holds.

(ii)-(iii) The first equality of part (ii) follows by (Rockafellar and Wets, 2009, Exercise 8.8), and the second one is implied by part (i). Next we consider part (iii). Suppose that $0 \in \nabla f(x) + \mathcal{N}_{\Pi(z)}(x)$. Obviously, $x \in \Pi(z)$. From the definition of $\Pi(\cdot)$, we have $\Pi(x) \subset \Pi(z)$, which along with their convexity and $x \in \Pi(x)$ implies that $\mathcal{N}_{\Pi(z)}(x) \subset \mathcal{N}_{\Pi(x)}(x)$. Combining with part (ii) leads to the desired result. ■

Remark 5 *Lemma 4 (ii) provides a way to seek a stationary point of* F*. Indeed, for any given* $z \in \Omega$*, if* x *is a stationary point of problem* $\min_{y\in\mathbb{R}^n}\{f(y) \mid y \in \Pi(z)\}$*, i.e.,* $0 \in \nabla f(x) + \mathcal{N}_{\Pi(z)}(x)$*, then by Lemma 4 (ii) it necessarily satisfies* $0 \in \partial F(x)$*. This implication will be utilized in the design of our algorithm, that is, when obtaining a good estimate of the stationary point, say* x^k*, we run a Newton step to minimize* f *over the polyhedral set* $\Pi(x^k)$ *so as to enhance the speed of the algorithm.*

2.2 Kurdyka-Łojasiewicz property

Next we introduce the Kurdyka-Łojasiewicz (KL) property of an extended real-valued function, which plays an important role in the convergence analysis of first-order algorithms for nonconvex and nonsmooth optimization problems (see, e.g., Attouch et al. (2010, 2013)). In this work, we will use it to establish the global convergence property of our algorithm.

Definition 6 *For any given* $\eta > 0$*, we denote by* Υ_η *the set consisting of all continuous concave* $\varphi : [0, \eta) \to \mathbb{R}_+$ *that are continuously differentiable on* $(0, \eta)$ *with* $\varphi(0) = 0$ *and* $\varphi'(s) > 0$ *for all* $s \in (0, \eta)$*. A proper function* $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is said to have the KL property at* $\overline{x} \in \operatorname{dom}\partial h$ *if there exist* $\eta \in (0, \infty]$*, a neighborhood* $\mathcal{U}$ *of* $\overline{x}$ *and a function* $\varphi \in \Upsilon_\eta$ *such that for all* $x \in \mathcal{U} \cap \big[h(\overline{x}) < h < h(\overline{x}) + \eta\big]$*,* $\varphi'(h(x) - h(\overline{x}))\mathrm{dist}(0, \partial h(x)) \geq 1$*. If* h *has the KL property at each point of* $\operatorname{dom}\partial h$*, then* h *is called a KL function.*

The KL property is ubiquitous, and the functions definable in an o-minimal structure over the real field admit this property; see (Attouch et al., 2010, Theorem 4.1). The functions definable in an o-minimal structure cover a wide range of functions, such as semi-algebraic functions and globally subanalytic functions; see (Van den Dries and Miller, 1996, Example 2.5). Moreover, from (Attouch et al., 2010, Section 4), we know that definable sets and functions are closed under some common calculus rules in optimization; for example, finite unions or finite intersections of definable sets are definable, compositions of definable mappings are definable, and subdifferentials of definable functions are definable.

3. Prox-regularity and proximal mapping of g

3.1 Prox-regularity of g

In this subsection, we aim at proving the prox-regularity of g, which together with Proposition 3 and the prox-boundedness of g indicates that the set of stationary points of problem (1) coincides with that of its L-stationary points.

We remark here that the prox-regularity of g cannot be obtained from the existing chain calculus of prox-regularity. It was revealed in (Poliquin and Rockafellar, 2010, Theorem 3.2) that, for proper f_i, $i = 1, 2$ with f_i being prox-regular at $\overline{x}$ for $v_i \in \partial f_i(\overline{x})$, by letting $v := v_1 + v_2$, and $f_0 := f_1 + f_2$, a sufficient condition for f_0 to be prox-regular at $\overline{x}$ for v is

$$w_1 + w_2 = 0 \text{ with } w_i \in \partial^\infty f_i(\overline{x}) \Longrightarrow w_i = 0,\ i = 1, 2, \tag{17}$$

where ∂^∞ denotes the horizon subdifferential (Rockafellar and Wets, 2009, Definition 8.3). We give a counter example to illustrate that the above constraint qualification does not hold for $f_i : \mathbb{R}^4 \to \mathbb{R}$ with $f_1 = \|\widehat{B}\cdot\|_0$ and $f_2 = \|\cdot\|_0$. Let $\overline{x} = (0, 0, 0, 1)^\top$. Then,

$$\partial^\infty f_1(\overline{x}) = \partial f_1(\overline{x}) = \mathrm{Range}((\widehat{B}_{[2]\cdot})^\top),\ \partial^\infty f_2(\overline{x}) = \partial f_2(\overline{x}) = \mathrm{Range}((I_{[3]\cdot})^\top).$$

By the expressions of $\partial^\infty f_1(\overline{x})$ and $\partial^\infty f_2(\overline{x})$, it is immediate to check that the constraint qualification in (17) does not hold. Next, we give our proof toward the prox-regularity of g.

Lemma 7 *The function g is prox-bounded, and is prox-regular on its domain Ω, so the set of stationary points of model (1) coincides with its set of L-stationary points.*

Proof The prox-boundedness of g is immediate by (Rockafellar and Wets, 2009, Definition 1.23). It suffices to prove that g is prox-regular on Ω. Fix any $\overline{x} \in \Omega$ and pick any $\overline{v} \in \partial g(\overline{x})$. Let $\lambda := \min\{\lambda_1, \lambda_2\}$ and $C := [B; I]$. Pick any $\varepsilon \in (0, \min\{\lambda, \frac{\lambda}{3(\|\overline{v}\|+\lambda)}\})$ such that for all $x \in \mathbb{B}(\overline{x}, \varepsilon)$, $\mathrm{supp}(Cx) \supset \mathrm{supp}(C\overline{x})$. We will prove that

$$g(x') \geq g(x) + v^\top(x'-x), \text{ for all } \|x'-\overline{x}\| \leq \varepsilon,\ v \in \partial g(x),\ \|v-\overline{v}\| < \varepsilon \text{ and } x \in \Xi \quad (18)$$

with $\Xi := \{x \mid \|x-\overline{x}\| < \varepsilon,\ g(x) < g(\overline{x}) + \varepsilon\}$, so the function g is prox-regular at $\overline{x}$ for $\overline{v}$.

We first claim that for each $x \in \Xi$, $\mathrm{supp}(Cx) = \mathrm{supp}(C\overline{x})$ and $x \in \Omega$. In fact, by the definition of ε, $\mathrm{supp}(Cx) \supset \mathrm{supp}(C\overline{x})$. If $\mathrm{supp}(Cx) \neq \mathrm{supp}(C\overline{x})$, we have $g(x) \geq g(\overline{x}) + \lambda > g(\overline{x}) + \varepsilon$, which yields that $x \notin \Xi$. Therefore, $\mathrm{supp}(Cx) = \mathrm{supp}(C\overline{x})$. The fact that $x \in \Xi$ implies $x \in \Omega$ is clear. Hence the claimed facts are true.

Fix any $x \in \Xi$. Consider any $x' \in \mathbb{B}(\overline{x}, \varepsilon)$. If $x' \notin \Omega$, since $g(x') = \infty$, it is immediate to see that (18) holds, so it suffices to consider $x' \in \mathbb{B}(\overline{x}, \varepsilon) \cap \Omega$. Note that $\mathrm{supp}(Cx') \supset \mathrm{supp}(C\overline{x}) = \mathrm{supp}(Cx)$. If $\mathrm{supp}(Cx') \neq \mathrm{supp}(Cx)$, then $g(x') \geq g(x) + \lambda$. For any $v \in \partial g(x)$ with $v \in \mathbb{B}(\overline{v}, \varepsilon)$, $\|v\| \leq \|\overline{v}\| + \varepsilon \leq \|\overline{v}\| + \lambda$, which along with $\|x'-x\| \leq \|x'-\overline{x}\| + \|x-\overline{x}\| \leq 2\varepsilon$ implies that $\|v\|\|x'-x\| \leq (\|\overline{v}\| + \lambda)\frac{2\lambda}{3(\|\overline{v}\|+\lambda)} \leq \frac{2\lambda}{3}$, and hence

$$g(x') - g(x) - v^\top(x'-x) \geq \lambda - \|v\|\|x'-x\| > 0.$$

Equation (18) holds. Next we consider the case $\mathrm{supp}(Cx') = \mathrm{supp}(Cx)$. Define

$$\Pi^1(x) := \big\{z \in \mathbb{R}^n \mid (Bz)_{[\mathrm{supp}(Bx)]^c} = 0\big\},\ \Pi^2(x) := \big\{z \in \mathbb{R}^n \mid z_{[\mathrm{supp}(x)]^c} = 0\big\}.$$

Clearly, $\Pi(x) = \Pi^1(x) \cap \Pi^2(x) \cap \Omega$ and $\Pi^1(x), \Pi^2(x)$ and Ω are all polyhedral sets. By (Rockafellar, 1970, Theorem 23.8), for any $v \in \mathcal{N}_{\Pi(x)}(x) = \partial g(x)$, there exist $v_1 \in \mathcal{N}_{\Pi^1(x)}(x)$, $v_2 \in \mathcal{N}_{\Pi^2(x)}(x)$ and $v_3 \in \mathcal{N}_\Omega(x)$ such that $v = v_1 + v_2 + v_3$. Then,

$$\begin{aligned} g(x') - g(x) - v^\top(x'-x) &= \lambda_1\|Bx'\|_0 - \lambda_1\|Bx\|_0 - v_1^\top(x'-x) \\ &\quad + \lambda_2\|x'\|_0 - \lambda_2\|x\|_0 - v_2^\top(x'-x) - v_3^\top(x'-x) \geq 0, \end{aligned}$$

where the inequality follows from $\lambda_1\|Bx'\|_0 - \lambda_1\|Bx\|_0 = 0, v_1^\top(x'-x) = 0,\ \lambda_2\|x'\|_0 - \lambda_2\|x\|_0 = 0, v_2^\top(x'-x) = 0$ and $v_3^\top(x'-x) \leq 0$. Equation (18) is true. Thus, by the arbitrariness of $\overline{x} \in \Omega$ and $\overline{v} \in \partial g(\overline{x})$, we conclude that g is prox-regular on set Ω. ■

3.2 Lower bound of the proximal mapping of g

Given $\lambda > 0$ and $x \in \mathbb{R}^n$, any $z \in \mathrm{prox}_{\lambda\|\cdot\|_0}(x)$ satisfies $|z_i| \geq \sqrt{2\lambda}$ for $i \in \mathrm{supp}(z)$ (Lu, 2014, Lemma 3.3). This indicates that $|z|_{\min}$ with $z \in \mathrm{prox}_{\lambda\|\cdot\|_0}(x)$ has a uniform lower bound. Such a uniform lower bound is shown to hold for the proximal mapping of the ℓ_q-norm with

$0<q<1$ and played a crucial role in the convergence analysis of the algorithms involving subspace Newton method (see Wu et al. (2023)). Next, we show that such a uniform lower bound exists for the proximal mapping of g.

Lemma 8 *For any given compact set $\Xi\subset\mathbb{R}^n$ and constants $0<\underline{\mu}<\overline{\mu}$, define*

$$\mathcal{Z}:=\bigcup\nolimits_{z\in\Xi,\mu\in[\underline{\mu},\overline{\mu}]}\ \mathrm{prox}_{\mu^{-1}g}(z).$$

Then, there exists $\nu>0$ (depending on $\Xi,\underline{\mu}$ and $\overline{\mu}$) such that $\inf_{u\in\mathcal{Z}\backslash\{0\}}|[B;I]u|_{\min}\ge\nu$.

Proof Let $C:=[B;I]$. By invoking (Bauschke et al., 1999, Corollary 3) and the compactness of Ω, there exists $\kappa>0$ such that for all index set $J\subset[n+p]$,

$$\mathrm{dist}(x,\mathrm{Null}(C_{J\cdot})\cap\Omega)\le\kappa\mathrm{dist}(x,\mathrm{Null}(C_{J\cdot}))\quad\text{for any }x\in\Omega. \tag{19}$$

Since the index sets $J\subset[n+p]$ are finite, there exists $\sigma>0$ such that for any index set $J\subset[n+p]$ with $C_{J\cdot}$ having full row rank,

$$\lambda_{\min}(C_{J\cdot}C_{J\cdot}^\top)\ge\sigma. \tag{20}$$

For any $z\in\Xi$ and $\mu\in[\underline{\mu},\overline{\mu}]$, define $h_{z,\mu}(x):=\frac{\mu}{2}\|x-z\|^2$ for $x\in\mathbb{R}^n$. By the compactness of Ω, $[\underline{\mu},\overline{\mu}]$ and Ξ, there exists $\delta_0\in(0,1)$ such that for all $z\in\Xi$, $\mu\in[\underline{\mu},\overline{\mu}]$ and $x,y\in\Omega$ with $\|x-y\|<\delta_0$, $\overline{\mu}(\|x\|+\|y\|+2\|z\|)\|x-y\|<\lambda:=\min\{\lambda_1,\lambda_2\}$, and consequently,

$$|h_{z,\mu}(x)-h_{z,\mu}(y)|=\frac{\mu}{2}|\langle x-y,x+y-2z\rangle|\le\frac{\overline{\mu}}{2}(\|x\|+\|y\|+2\|z\|)\|x-y\|<\frac{\lambda}{2}. \tag{21}$$

Now suppose on the contrary that the conclusion does not hold. Then there is a sequence $\{\overline{z}^k\}_{k\in\mathbb{N}}\subset\mathcal{Z}\backslash\{0\}$ such that $|C\overline{z}^k|_{\min}\le\frac{1}{k}$ for all $k\in\mathbb{N}$. Note that C has a full column rank. We also have $|C\overline{z}^k|_{\min}>0$ for each $k\in\mathbb{N}$. By the definition of $\mathcal{Z}$, for each $k\in\mathbb{N}$, there exist $z^k\in\Xi$ and $\mu_k\in[\underline{\mu},\overline{\mu}]$ such that $\overline{z}^k\in\mathrm{prox}_{\mu_k^{-1}g}(z^k)$. Since $|C\overline{z}^k|_{\min}\in\big(0,\frac{1}{k}\big)$ for all $k\in\mathbb{N}$, there exists an infinite index set $\mathcal{K}\subset\mathbb{N}$ and an index $i\in[n+p]$ such that

$$0<|(C\overline{z}^k)_i|=|C\overline{z}^k|_{\min}<\frac{\delta_0\sigma}{\kappa\|C\|_2}\quad\text{for each }k\in\mathcal{K}, \tag{22}$$

where κ and σ are the ones appearing in (19) and (20), respectively. Fix any $k\in\mathcal{K}$. Write $Q_k:=[n+p]\backslash\mathrm{supp}(C\overline{z}^k)$ and choose $J_k\subset Q_k$ such that the rows of $C_{J_k\cdot}$ form a basis of those of $C_{Q_k\cdot}$. Let $\widehat{J}_k:=J_k\cup\{i\}$. Obviously, $\|C_{\widehat{J}_k\cdot}\overline{z}^k\|=|(C\overline{z}^k)_i|$. We claim that $C_{\widehat{J}_k\cdot}$ also has a full row rank. Indeed, if $J_k=\emptyset$, then $C_{\widehat{J}_k\cdot}$ has a full row rank because $C_{\widehat{J}_k\cdot}\ne0$ by (22); if $J_k\ne\emptyset$, then $C_{J_k\cdot}\overline{z}^k=0$, which implies that $C_{\widehat{J}_k\cdot}$ also has a full row rank (if not, $C_{i\cdot}$ is a linear combination of the rows of $C_{J_k\cdot}$, which along with $C_{J_k\cdot}\overline{z}^k=0$ implies that $C_{i\cdot}\overline{z}^k=0$, contradicting to $|(C\overline{z}^k)_i|=|C\overline{z}^k|_{\min}>0$). The claimed fact holds. Let $\widetilde{z}^k:=\mathrm{proj}_{\mathrm{Null}(C_{\widehat{J}_k\cdot})}(\overline{z}^k)$. Then, $C_{\widehat{J}_k\cdot}\widetilde{z}^k=0$, and by the optimality condition of the projection problem, there exists $\xi^k\in\mathbb{R}^{|\widehat{J}_k|}$ such that $\overline{z}^k-\widetilde{z}^k=C_{\widehat{J}_k\cdot}^\top\xi^k$. Since $C_{\widehat{J}_k\cdot}$ has a full row rank and $\|C_{\widehat{J}_k\cdot}\overline{z}^k\|=|(C\overline{z}^k)_i|$, we have

$$|(C\overline{z}^k)_i|=\|C_{\widehat{J}_k\cdot}\overline{z}^k-C_{\widehat{J}_k\cdot}\widetilde{z}^k\|=\|C_{\widehat{J}_k\cdot}C_{\widehat{J}_k\cdot}^\top\xi^k\|\ge\sigma\|\xi^k\|, \tag{23}$$

where the inequality is due to (20). Combining (23) with (22) yields $\|\xi^k\| < \kappa^{-1}\|C\|_2^{-1}\delta_0$. Therefore,

$$\|\overline{z}^k - \widetilde{z}^k\| = \|C_{\widehat{J}_k\cdot}^\top \xi^k\| \le \|C_{\widehat{J}_k\cdot}\|_2\|\xi^k\| \le \|C\|_2\|\xi^k\| < \kappa^{-1}\delta_0. \tag{24}$$

Let $\widehat{z}^k := \mathrm{proj}_{\mathrm{Null}(C_{\widehat{J}_k\cdot})\cap\Omega}(\overline{z}^k)$. From (19) and (24), it follows that

$$\|\overline{z}^k - \widehat{z}^k\| = \mathrm{dist}(\overline{z}^k, \mathrm{Null}(C_{\widehat{J}_k\cdot}) \cap \Omega) \le \kappa\mathrm{dist}(\overline{z}^k, \mathrm{Null}(C_{\widehat{J}_k\cdot})) = \kappa\|\overline{z}^k - \widetilde{z}^k\| < \delta_0. \tag{25}$$

Note that $\widehat{z}^k, \overline{z}^k \in \Omega$. From (25) and (21), it follows that

$$|h_{z^k,\mu_k}(\widehat{z}^k) - h_{z^k,\mu_k}(\overline{z}^k)| < \frac{\lambda}{2}. \tag{26}$$

Next we claim that $\mathrm{supp}(C\widehat{z}^k) \cup \{i\} \subset \mathrm{supp}(C\overline{z}^k)$. Indeed, since the rows of $C_{\widehat{J}_k\cdot}$ form a basis of those of $C_{[Q_k\cup\{i\}]\cdot}$ and $C_{\widehat{J}_k\cdot}\widehat{z}^k = 0$, $C_{[Q_k\cup\{i\}]\cdot}\widehat{z}^k = 0$. Then, $\mathrm{supp}(C_{[Q_k\cup\{i\}]\cdot}\widehat{z}^k) \cup \{i\} = \mathrm{supp}(C_{[Q_k\cup\{i\}]\cdot}\overline{z}^k)$. Since all the entries of $C_{[Q_k\cup\{i\}]^c\cdot}\overline{z}^k$ are nonzero, it holds that $\mathrm{supp}(C_{[Q_k\cup\{i\}]^c\cdot}\widehat{z}^k) \subset \mathrm{supp}(C_{[Q_k\cup\{i\}]^c\cdot}\overline{z}^k)$, which implies that $\mathrm{supp}(C\widehat{z}^k) \cup \{i\} \subset \mathrm{supp}(C\overline{z}^k)$. Thus, the claimed inclusion follows, which implies that $g(\overline{z}^k) - g(\widehat{z}^k) \ge \lambda$. This together with (26) yields $h_{z^k,\mu_k}(\overline{z}^k) + g(\overline{z}^k) - (h_{z^k,\mu_k}(\widehat{z}^k) + g(\widehat{z}^k)) \ge \lambda - \frac{\lambda}{2} = \frac{\lambda}{2}$, contradicting to $\overline{z}^k \in \mathrm{prox}_{\mu_k^{-1}g}(z^k)$. The proof is completed. ■

The result of Lemma 8 will be utilized in Proposition 14 to justify the fact that the sequences $\{|B\overline{x}^k|_{\min}\}_{k\in\mathbb{N}}$ and $\{|\overline{x}^k|_{\min}\}_{k\in\mathbb{N}}$ are uniformly lower bounded, where $\overline{x}^k$ is obtained in (7) (or (38) below). This is a crucial aspect in proving the stability of $\mathrm{supp}(x^k)$ and $\mathrm{supp}(Bx^k)$ when k is sufficiently large.

3.3 Proximal mapping of a fused ℓ_0-norms function with a box constraint

The characterization for the proximal mapping of the fused ℓ_0-norm $\lambda_1\|\widehat{B}\cdot\|_0$ can be traced back to Liebscher and Winkler (1999), where the problem is addressed by using the technique of optimal partitioning of changepoints. For later developments of this technique, please refer to Jackson et al. (2005); Friedrich et al. (2008); Killick et al. (2012); Weinmann et al. (2015); Jewell and Witten (2018). Recently, by using the functional pruning technique introduced in Rigaill (2015) and Maidstone et al. (2017), Jewell et al. (2020) presented a polynomial-time algorithm for computing the proximal mapping of $\lambda_1\|\widehat{B}\cdot\|_0$. Numerical experiments show that this method is more efficient than the one proposed in Jewell and Witten (2018); see also the arguments in (Jewell et al., 2020, Section 2.2). In this subsection, we extend the functional pruning technique to compute the proximal mapping of the fused ℓ_0-norms $\lambda_1\|\widehat{B}\cdot\|_0 + \lambda_2\|\cdot\|_0 + \delta_\Omega(\cdot)$, i.e., for any given $z \in \mathbb{R}^n$ ($n \ge 2$), to seek a global optimal solution of the problem

$$\min_{x\in\mathbb{R}^n} h(x;z) := \frac{1}{2}\|x - z\|^2 + \lambda_1\|\widehat{B}x\|_0 + \lambda_2\|x\|_0 + \delta_\Omega(x). \tag{27}$$

To simplify the deduction, for each $i \in [n]$, define $\omega_i(\alpha) := \lambda_2|\alpha|_0 + \delta_{[l_i,u_i]}(\alpha)$ for $\alpha \in \mathbb{R}$. Clearly, $\lambda_2\|x\|_0 + \delta_\Omega(x) = \sum_{i=1}^n \omega_i(x_i)$ for $x \in \mathbb{R}^n$. Let $H(0) := -\lambda_1$, and for each $s \in [n]$, define

$$H(s) := \min_{y\in\mathbb{R}^s} h_s(y; z_{1:s}) := \frac{1}{2}\|y - z_{1:s}\|^2 + \lambda_1\|\widehat{B}_{[s-1][s]}y\|_0 + \sum_{j=1}^{s} \omega_j(y_j) \tag{28}$$

with $\widehat{B}_{[0][1]} := 0$. It is immediate to see that $H(n)$ is the optimal value to (27). For each $s \in [n]$, define function $P_s\colon [0\!:\!s-1] \times \mathbb{R} \to \overline{\mathbb{R}}$ by

$$P_s(i, \alpha) := H(i) + \frac{1}{2}\|\alpha\mathbf{1} - z_{i+1:s}\|^2 + \sum_{j=i+1}^{s} \omega_j(\alpha) + \lambda_1. \tag{29}$$

For each $s \in [n]$, there is a close relation between P_s and h_s. Indeed, for any given $y \in \mathbb{R}^s$ with $y_s = \alpha$, let i be the smallest integer in $[0\!:\!s-1]$ such that $y_{i+1} = \cdots = y_s = \alpha$. When $i = 0$, $P_s(i, \alpha) = \frac{1}{2}\|y_{1:s} - z_{1:s}\|^2 + \sum_{j=1}^{s} \omega_j(y_j) = h_s(y; z_{1:s})$. When $i \neq 0$, if $y_{1:i}$ is optimal to $\min_{y'\in\mathbb{R}^i} h_i(y'; z_{1:i})$, then by noting that $y = (y_{1:i}; \alpha\mathbf{1})$ and

$$\begin{aligned} h_s(y; z_{1:s}) &= \frac{1}{2}\|y_{1:i} - z_{1:i}\|^2 + \lambda_1\|\widehat{B}_{[i-1][i]}y_{1:i}\|_0 + \sum_{j=1}^{i} \omega_j(y_j) \\ &\quad + \frac{1}{2}\|y_{i+1:s} - z_{i+1:s}\|^2 + \sum_{j=i+1}^{s} \omega_j(y_j) + \lambda_1 \\ &= h_i(y_{1:i}; z_{1:i}) + \frac{1}{2}\|\alpha\mathbf{1} - z_{i+1:s}\|^2 + \sum_{j=i+1}^{s} \omega_j(\alpha) + \lambda_1, \end{aligned} \tag{30}$$

we get $H(i) = h_i(y_{1:i}; z_{1:i})$. Along with the above equality and (29), $h_s(y; z_{1:s}) = P_s(i, \alpha)$.

In the following lemma, we prove that the optimal value of $\min_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i, \alpha)$ is equal to $H(s)$, and apply this result to characterize a global minimizer of $h_s(\cdot; z_{1:s})$.

Lemma 9 *Fix any $s \in [n]$. The following statements are true.*

(i) $H(s) = \min_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i, \alpha)$.

(ii) *If $(i_s^*, \alpha_s^*) \in \mathop{\arg\min}_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i, \alpha)$, then $y^* = (y^*_{1:i_s^*}; \alpha_s^*\mathbf{1})$ with $y^*_{1:i_s^*} \in \mathop{\arg\min}_{v\in\mathbb{R}^{i_s^*}} h_{i_s^*}(v; z_{1:i_s^*})$ is a global optimal solution of the minimization problem $\min_{y\in\mathbb{R}^s} h_s(y; z_{1:s})$.*

Proof **(i)** Let y^* be an optimal solution of problem (28). If $y_i^* = y_j^*$ for any $i, j \in [s]$, let $i_s^* = 0$; otherwise, let $i_s^* \in [s-1]$ be the largest integer such that $y^*_{i_s^*} \neq y^*_{i_s^*+1}$. Set $\alpha_s^* = y^*_{i_s^*+1}$. If $i_s^* \neq 0$, from the definition of $H(\cdot)$, $h_{i_s^*}(y^*_{1:i_s^*}; z_{1:i_s^*}) \geq H(i_s^*)$, which implies that

$$\begin{aligned} \min_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i, \alpha) &\leq H(i_s^*) + \frac{1}{2}\|\alpha_s^*\mathbf{1} - z_{i_s^*+1:s}\|^2 + \sum_{j=i_s^*+1}^{s} \omega_j(\alpha_s^*) + \lambda_1 \\ &\leq h_{i_s^*}(y^*_{1:i_s^*}; z_{1:i_s^*}) + \frac{1}{2}\|y^*_{i_s^*+1:s} - z_{i_s^*+1:s}\|^2 + \sum_{j=i_s^*+1}^{s} \omega_j(y_j^*) + \lambda_1 \\ &= h_s(y^*; z_{1:s}) = H(s), \end{aligned}$$

where the first equality is due to $y^*_{i^*_s+1} \neq y^*_{i^*_s}$ and the expression of $h_s(y^*; z_{1:s})$ by (30). If $i^*_s = 0$,

$$\min_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i,\alpha) \le H(0) + \frac{1}{2}\|y^* - z_{1:s}\|^2 + \sum_{j=1}^{s} \omega_j(y^*_j) + \lambda_1 = H(s).$$

Therefore, $\min_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i,\alpha) \le H(s)$ holds. On the other hand, let (i^*_s, α^*_s) be an optimal solution to $\min_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i,\alpha)$. If $i^*_s \neq 0$, let $y^* \in \mathbb{R}^s$ be such that $y^*_{1:i^*_s} \in \arg\min_{v\in\mathbb{R}^{i^*_s}} h_{i^*_s}(v; z_{1:i^*_s})$ and $y^*_{i^*_s+1:s} = \alpha^*_s \mathbf{1}$. Then, it is clear that

$$\begin{aligned} H(s) \le h_s(y^*; z_{1:s}) &\le h_{i^*_s}(y^*_{1:i^*_s}; z_{1:i^*_s}) + \frac{1}{2}\|y^*_{i^*_s+1:s} - z_{i^*_s+1:s}\|^2 + \sum_{j=i^*_s+1}^{s} \omega_j(y^*_j) + \lambda_1 \\ &= H(i^*_s) + \frac{1}{2}\|\alpha^*_s\mathbf{1} - z_{i^*_s+1:s}\|^2 + \sum_{j=i^*_s+1}^{s} \omega_j(\alpha^*_s) + \lambda_1 = \min_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i,\alpha). \end{aligned}$$

If $i^*_s = 0$, let $y^* = \alpha^*_s\mathbf{1}$. We have

$$H(s) \le h_s(y^*; z_{1:s}) = H(0) + \frac{1}{2}\|y^* - z_{1:s}\|^2 + \sum_{j=1}^{s} \omega_j(\alpha^*_s) + \lambda_1 = \min_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i,\alpha).$$

Therefore, $H(s) \le \min_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i,\alpha)$. The above two inequalities imply the result.
(ii) If $i^*_s \neq 0$, by part (i) and the definitions of α^*_s and i^*_s, it holds that

$$\begin{aligned} H(s) = \min_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i,\alpha) &= H(i^*_s) + \frac{1}{2}\|\alpha^*_s\mathbf{1} - z_{i^*_s+1:s}\|^2 + \sum_{j=i^*_s+1}^{s} \omega_j(\alpha^*_s) + \lambda_1 \\ &= h_{i^*_s}(y^*_{1:i^*_s}; z_{1:i^*_s}) + \frac{1}{2}\|y^*_{i^*_s+1:s} - z_{i^*_s+1:s}\|^2 + \sum_{j=i^*_s+1}^{s} \omega_j(y^*_j) + \lambda_1 \ge h_s(y^*; z_{1:s}), \end{aligned}$$

where the last inequality follows by (30). If $i^*_s = 0$,

$$H(s) = \min_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i,\alpha) = P_s(0,\alpha) = h_s(y^*; z_{1:s}).$$

Therefore, $H(s) \ge h_s(y^*; z_{1:s})$. Along with the definition of $H(s)$, $H(s) = h_s(y^*; z_{1:s})$. ■

Lemma 9 (i) implies that the nonconvex and nonsmooth problem (27) can be recast as a mixed-integer programming with objective function given in (29). Lemma 9 (ii) suggests a recursive method to obtain an optimal solution to (27). In fact, by setting $s = n$, there exists an optimal solution to (27), says x^*, such that $x^*_{i^*_n+1:n} = \alpha^*_n\mathbf{1}$, and $x^*_{1:i^*_n} \in \arg\min_{v\in\mathbb{R}^{i^*_n}} h_{i^*_n}(v; z_{1:i^*_n})$. Next, by setting $s = i^*_n$, we are able to obtain the expression of $x^*_{i^*_s+1:i^*_n}$. Repeating this loop backward until $s = 0$, we can obtain the full expression of an optimal solution to (27). The outline of computing $\mathrm{prox}_{\lambda_1\|\widehat{B}\cdot\|_0+\omega(\cdot)}(z)$ is shown as follows.

$$\begin{cases} \text{Set } s=n. \\ \textbf{While } s>0 \text{ do} \\ \qquad \text{Find } (i_s^*,\alpha_s^*)\in \displaystyle\mathop{\arg\min}_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i,\alpha). \\ \qquad \text{Let } x^*_{i_s^*+1:s}=\alpha_s^*\mathbf{1} \text{ and } s\leftarrow i_s^*. \\ \textbf{End} \end{cases} \tag{31}$$

To obtain an optimal solution to (27), the remaining issue is how to execute the first line in **while** loop of (31), or in other words, for any given $s\in[n]$, how to find $(i_s^*,\alpha_s^*)\in\mathbb{N}\times\mathbb{R}$ appearing in Lemma 9 (ii). The following proposition provides some preparations.

Proposition 10 *For each $s\in[n]$, let $P_s^*(\alpha):=\min_{i\in[0:s-1]}P_s(i,\alpha)$.*

(i) *For any $\alpha\in\mathbb{R}$, it holds that*

$$P_s^*(\alpha)=\begin{cases} \frac{1}{2}(\alpha-z_1)^2+\omega_1(\alpha) & \text{if } s=1,\\ \min\Big\{P_{s-1}^*(\alpha),\min_{\alpha'\in\mathbb{R}}P_{s-1}^*(\alpha')+\lambda_1\Big\}+\frac{1}{2}(\alpha-z_s)^2+\omega_s(\alpha) & \text{if } s\in[2:n].\end{cases}$$

(ii) *Let $\mathcal{R}_1^0:=\mathbb{R}$. For each $s\in[2:n]$ and $i\in[0:s-2]$, let $\mathcal{R}_s^i:=\mathcal{R}_{s-1}^i\cap(\mathcal{R}_s^{s-1})^c$ with*

$$\mathcal{R}_s^{s-1}:=\Big\{\alpha\in\mathbb{R}\ |\ P_{s-1}^*(\alpha)\ge\min_{\alpha'\in\mathbb{R}}P_{s-1}^*(\alpha')+\lambda_1\Big\}. \tag{32}$$

Then, the following assertions hold true.

(a) For each $s\in[2:n]$, $\bigcup_{i\in[0:s-1]}\mathcal{R}_s^i=\mathbb{R}$ and $\mathcal{R}_s^i\cap\mathcal{R}_s^j=\emptyset$ for any $i\neq j\in[0:s-1]$.

(b) For each $s\in[n]$ and $i\in[0:s-1]$, $P_s^*(\alpha)=P_s(i,\alpha)$ when $\alpha\in\mathcal{R}_s^i$.

Proof (i) Fix any $\alpha\in\mathbb{R}$. Note that $P_1^*(\alpha)=P_1(0,\alpha)=H(0)+\frac{1}{2}(\alpha-z_1)^2+\omega_1(\alpha)+\lambda_1=\frac{1}{2}(\alpha-z_1)^2+\omega_1(\alpha)$. Now pick any $s\in[2:n]$. By the definition of P_s^*, we have

$$P_s^*(\alpha)=\min_{i\in[0:s-1]}P_s(i,\alpha)=\min\Big\{\min_{i\in[0:s-2]}P_s(i,\alpha),\ P_s(s-1,\alpha)\Big\}. \tag{33}$$

From the definition of P_s in (29), for each $i\in[0:s-2]$, it holds that

$$\begin{aligned} P_s(i,\alpha)&=H(i)+\frac{1}{2}\|\alpha\mathbf{1}-z_{i+1:s}\|^2+\sum_{j=i+1}^{s}\omega_j(\alpha)+\lambda_1\\ &=H(i)+\frac{1}{2}\|\alpha\mathbf{1}-z_{i+1:s-1}\|^2+\sum_{j=i+1}^{s-1}\omega_j(\alpha)+\lambda_1+\frac{1}{2}(\alpha-z_s)^2+\omega_s(\alpha)\\ &=P_{s-1}(i,\alpha)+\frac{1}{2}(\alpha-z_s)^2+\omega_s(\alpha),\end{aligned}$$

while $P_s(s-1,\alpha) = H(s-1) + \frac{1}{2}(\alpha - z_s)^2 + \omega_s(\alpha) + \lambda_1$. Together with the above equality and (33), we immediately obtain that

$$\begin{aligned} P_s^*(\alpha) &= \min\Big\{ \min_{i\in[0:s-2]} P_{s-1}(i,\alpha),\, H(s-1)+\lambda_1 \Big\} + \frac{1}{2}(\alpha - z_s)^2 + \omega_s(\alpha) \\ &= \min\Big\{ P_{s-1}^*(\alpha), \min_{\alpha'\in\mathbb{R}} P_{s-1}^*(\alpha') + \lambda_1 \Big\} + \frac{1}{2}(\alpha - z_s)^2 + \omega_s(\alpha), \end{aligned} \tag{34}$$

where the second equality is by Lemma 9 (i) and the definition of P_{s-1}^*. Thus, we get the desired result.

(ii) We first prove (a) by induction. When $s=2$, since $\mathcal{R}_1^0 = \mathbb{R}$ and $\mathcal{R}_2^0 = \mathcal{R}_1^0 \cap (\mathcal{R}_2^1)^c$, we have $\mathcal{R}_2^0 \cup \mathcal{R}_2^1 = \mathbb{R}$ and $\mathcal{R}_2^0 \cap \mathcal{R}_2^1 = \emptyset$. Assume that the result holds with $s=j$ for some $j \in [2:n-1]$. We prove that the result holds for $s = j+1$. Since $\mathcal{R}_{j+1}^i = \mathcal{R}_j^i \cap (\mathcal{R}_{j+1}^j)^c$ for all $i \in [0:j-1]$ and $\bigcup_{i\in[0:j-1]} \mathcal{R}_j^i = \mathbb{R}$, it holds that

$$\bigcup_{i\in[0:j]} \mathcal{R}_{j+1}^i = \big[\bigcup_{i\in[0:j-1]} (\mathcal{R}_j^i \cap (\mathcal{R}_{j+1}^j)^c)\big] \cup \mathcal{R}_{j+1}^j = \big(\mathbb{R} \cap (\mathcal{R}_{j+1}^j)^c\big) \cup \mathcal{R}_{j+1}^j = \mathbb{R}.$$

The first part holds. For the second part, by definition, $\mathcal{R}_{j+1}^i \cap \mathcal{R}_{j+1}^j = \emptyset$ for all $i \in [0:j-1]$, so it suffices to prove that $\mathcal{R}_{j+1}^i \cap \mathcal{R}_{j+1}^k = \emptyset$ for any $i \neq k \in [0:j-1]$. By definition,

$$\mathcal{R}_{j+1}^i \cap \mathcal{R}_{j+1}^k = \big[\mathcal{R}_j^i \cap (\mathcal{R}_{j+1}^j)^c\big] \cap \big[\mathcal{R}_j^k \cap (\mathcal{R}_{j+1}^j)^c\big] = \emptyset,$$

where the second equality is due to $\mathcal{R}_j^i \cap \mathcal{R}_j^k = \emptyset$. Thus, the second part follows.

Next we prove (b). When $s=1$, since for any $\alpha \in \mathbb{R} = \mathcal{R}_1^0$, $P_1^*(\alpha) = P_1(0,\alpha)$, the result holds. For $s \in [2:n]$ and $i = s-1$, by the definition of $\mathcal{R}_s^{s-1}$ and part (i), for all $\alpha \in \mathcal{R}_s^{s-1}$,

$$P_s^*(\alpha) = \min_{\alpha'\in\mathbb{R}} P_{s-1}^*(\alpha') + \lambda_1 + \frac{1}{2}(\alpha - z_s)^2 + \omega_s(\alpha) = P_s(s-1,\alpha),$$

where the second equality is obtained by using $H(s) = \min_{\alpha'\in\mathbb{R}} P_{s-1}^*(\alpha')$ and (29). Next we consider $s \in [2:n]$ and $i \in [0:s-2]$. We argue by induction that $P_s^*(\alpha) = P_s(i,\alpha)$ when $\alpha \in \mathcal{R}_s^i$. Indeed, when $s=2$, since $\mathcal{R}_2^0 = \mathcal{R}_1^0 \cap (\mathcal{R}_2^1)^c = (\mathcal{R}_2^1)^c$, for any $\alpha \in \mathcal{R}_2^0$, from (32) we have $P_1^*(\alpha) < \min_{\alpha'\in\mathbb{R}} P_1^*(\alpha') + \lambda_1$, which by part (i) implies that

$$P_2^*(\alpha) = P_1^*(\alpha) + \frac{1}{2}(\alpha - z_2)^2 + \omega_2(\alpha) = P_1(0,\alpha) + \frac{1}{2}(\alpha - z_2)^2 + \omega_2(\alpha) = P_2(0,\alpha).$$

Assume that the result holds when $s = j$ for some $j \in [2:n-1]$. We consider the case for $s = j+1$. For any $i \in [0:j-1]$, from $\mathcal{R}_{j+1}^i = \mathcal{R}_j^i \cap (\mathcal{R}_{j+1}^j)^c$ and (34), for any $\alpha \in \mathcal{R}_{j+1}^i$,

$$\begin{aligned} P_{j+1}^*(\alpha) &= P_j^*(\alpha) + \frac{1}{2}(\alpha - z_{j+1})^2 + \omega_{j+1}(\alpha) = P_j(i,\alpha) + \frac{1}{2}(\alpha - z_{j+1})^2 + \omega_{j+1}(\alpha) \\ &= H(i) + \frac{1}{2}\|\alpha\mathbf{1} - z_{i+1:j}\|^2 + \sum_{k=i+1}^{j} w_k(\alpha) + \lambda_1 + \frac{1}{2}(\alpha - z_{j+1})^2 + \omega_{j+1}(\alpha) \\ &= H(i) + \frac{1}{2}\|\alpha\mathbf{1} - z_{i+1:j+1}\|^2 + \sum_{k=i+1}^{j+1} w_k(\alpha) + \lambda_1 = P_{j+1}(i,\alpha), \end{aligned}$$

where the second equality is using $P_j^*(\alpha) = P_j(i, \alpha)$ implied by induction. Hence, the conclusion holds for $s = j+1$ and any $i \in [0\!:\!s-2]$. The proof is completed. ■

Now we take a closer look at Proposition 10. Part (i) provides a recursive method to compute $P_s^*(\alpha)$ for all $s \in [n]$. For each $s \in [n]$, by the expression of ω_s, $P_s(i, \cdot)$ is a piecewise lower semicontinuous linear-quadratic function whose domain is a closed interval, relative to which $P_s(i, \cdot)$ has an expression of the form $H(i) + \frac{1}{2}\|\alpha\mathbf{1} - z_{i+1:s}\|^2 + (s-i)|\alpha|_0 + \lambda_1$. While $P_s^*(\cdot) = \min\{P_s(0, \cdot), P_s(1, \cdot), \ldots, P_s(s-1, \cdot)\}$, and for each $i \in [0\!:\!s-1]$, the optimal solution to $\min_{\alpha\in\mathbb{R}} P_s(i, \alpha)$ is easily obtained (in fact, all the possible candidates of the global solutions are 0, $\frac{\sum_{j=i+1}^{s} z_j}{s-i}, \max_{j\in[i+1:s]}\{l_j\}, \min_{j\in[i+1:s]}\{u_j\}$), so is $\arg\min_{\alpha'\in\mathbb{R}} P_s^*(\alpha')$. Part (ii) suggests a way to search for i_s^* such that $P_s^*(\alpha_s^*) = P_s(i_s^*, \alpha_s^*)$ for each $s \in [n]$. Obviously, $P_s(i_s^*, \alpha_s^*) = \min_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i, \alpha)$. This inspires us to propose Algorithm 1 for solving $\mathrm{prox}_{\lambda_1\|\widehat{B}\cdot\|_0+\omega(\cdot)}(z)$, whose iteration steps are described as follows.

Algorithm 1 (Computing $\mathrm{prox}_{\lambda_1\|\widehat{B}\cdot\|_0+\omega(\cdot)}(z)$)

1. **Initialize:** Compute $P_1^*(\alpha) = \frac{1}{2}(z_1 - \alpha)^2 + \omega_1(\alpha)$ and set $\mathcal{R}_1^0 = \mathbb{R}$.
2. **For** $s = 2, \ldots, n$ **do**
3. $\quad P_s^*(\alpha) := \min\{P_{s-1}^*(\alpha), \min_{\alpha'\in\mathbb{R}} P_{s-1}^*(\alpha') + \lambda_1\} + \frac{1}{2}(\alpha - z_s)^2 + \omega_s(\alpha)$.
4. $\quad$ Compute $\mathcal{R}_s^{s-1}$ by (32).
5. $\quad$ **For** $i = 0, \ldots, s-2$ **do**
6. $\qquad \mathcal{R}_s^i = \mathcal{R}_{s-1}^i \cap (\mathcal{R}_s^{s-1})^c$.
7. $\quad$ **End**
8. **End**
9. Set $s = n$.
10. **While** $s > 0$ **do**
11. $\quad$ Find $\alpha_s^* \in \arg\min_{\alpha\in\mathbb{R}} P_s^*(\alpha)$, and $i_s^* = \{i \mid \alpha_s^* \in \mathcal{R}_s^i\}$.
12. $\quad x_{i_s^*+1:s}^* = \alpha_s^*\mathbf{1}$ and $s \leftarrow i_s^*$.
13. **End**

For every $s \in [n]$, as P_s^* is a piecewise lower semicontinuous linear-quadratic function, in the implementation of Algorithm 1, we store the parameters to identify this function via a matrix, whose each row records the parameters of P_s^* and its domain. Similarly, each $\mathcal{R}_s^i$ is stored via a vector which records its endpoints. The main computation cost of Algorithm 1 comes from lines 3 and 6, in which the number of pieces of the linear-quadratic functions involved in P_s^* plays a key role. The following lemma gives a worst-case estimation for it.

Lemma 11 *Fix any $s \in [2:n]$. The function P_s^* in line 3 of Algorithm 1 has at most $O(s^{1+\epsilon})$ linear-quadratic pieces, where ϵ is any small positive constant.*

Proof For each $i \in [0:s-2]$, let $h_i(\alpha) := H(i) + \frac{1}{2}\|\alpha\mathbf{1} - z_{i+1:s}\|^2 + \lambda_1 + (s-i)\lambda_2|\alpha|_0 + \sum_{j=i+1}^{s} \delta_{[l_j,u_j]}(\alpha)$ for $\alpha \in \mathbb{R}$. Obviously, every h_i is a piecewise lower semicontinuous linear-quadratic function whose domain is a closed interval, and every piece is continuous on the closed interval except $\alpha = 0$. Therefore, for each $i \in [0:s-2]$, $h_i = \min\{h_{i,1}, h_{i,2}, h_{i,3}\}$ with $h_{i,1}(\alpha) := h_i(\alpha) - (s-i)\lambda_2|\alpha|_0 + (s-i)\lambda_2 + \delta_{(-\infty,0]}(\alpha)$, $h_{i,2}(\alpha) := h_i(\alpha) + \delta_{\{0\}}(\alpha)$ and $h_{i,3}(\alpha) := h_i(\alpha) - (s-i)\lambda_2|\alpha|_0 + (s-i)\lambda_2 + \delta_{[0,\infty)}(\alpha)$. Obviously, $h_{i1}, h_{i,2}$ and $h_{i,3}$ are piecewise linear-quadratic functions with domain being a closed interval. In addition, write $h_{s-1}(\alpha) := \min_{\alpha'\in\mathbb{R}} P^*_{s-1}(\alpha') + \frac{1}{2}(\alpha - z_s)^2 + \lambda|\alpha|_0 + \lambda_1 + \delta_{[l_s,u_s]}(\alpha)$ for $\alpha \in \mathbb{R}$. Obviously, h_{s-1} is a piecewise lower semicontinuous linear-quadratic function whose domain is a closed interval. Similarly, $h_{s-1} = \min\{h_{s-1,1}, h_{s-1,2}, h_{s-1,3}\}$ where each $h_{s-1,j}$ for $j = 1, 2, 3$ is a piecewise linear function whose domain is a closed interval. Combining the above discussion with line 3 of Algorithm 1 and the definition of P^*_{s-1}, for any $\alpha \in \mathbb{R}$,

$$\begin{aligned} P^*_s(\alpha) &= \min_{i\in[0:s-2]} \Big\{ P_{s-1}(i,\alpha) + \frac{1}{2}(\alpha - z_s)^2 + \omega_s(\alpha), \min_{\alpha'\in\mathbb{R}} P^*_{s-1}(\alpha') + \frac{1}{2}(\alpha - z_s)^2 + \omega_s(\alpha) + \lambda_1 \Big\} \\ &= \Big\{ h_0(\alpha), h_1(\alpha), \ldots, h_{s-2}(\alpha), h_{s-1}(\alpha) \Big\} = \min_{i\in[0:s-1], j\in[3]} \big\{ h_{i,j}(\alpha) \big\}. \end{aligned}$$

Notice that any $h_{i,j}$ and $h_{i',j'}$ with $i \neq i' \in [0 : s-1]$ or $j \neq j' \in [3]$ crosses at most 2 times. From (Sharir, 1995, Theorem 2.5) the maximal number of linear-quadratic pieces involved in P^*_s is bounded by the maximal length of a $(3s, 4)$ Davenport-Schinzel sequence, which by (Davenport and Schinzel, 1965, Theorem 3) is $3c_1 s \exp(c_2\sqrt{\log 3s})$. Here, c_1, c_2 are positive constants independent of s. Thus, we conclude that the maximal number of linear-quadratic pieces involved in P^*_s is $O(s^{1+\epsilon})$ for any $\epsilon > 0$. The proof is finished. ■

By invoking Lemma 11, we are able to provide a worst-case estimation for the complexity of Algorithm 1. Indeed, the main cost of Algorithm 1 consists in lines 3 and 5-7. The computation cost involved in line 3 depends on the number of pieces of P^*_{s-1}, which by Lemma 11 requires $O(s^{1+\epsilon})$ operations with any small $\epsilon > 0$. From part (b) of Proposition (ii), for each $i \in [0 : s-1]$, $\mathcal{R}^i_s$ consists of at most $O(s^{1+\epsilon})$ intervals, which means that line 6 requires at most $O(s^{1+\epsilon})$ operations and then the computation complexity of lines 5-7 is $O(s^{2+\epsilon})$ with any small $\epsilon > 0$. Thus, the worst-case complexity of Algorithm 1 is $\sum_{s=2}^{n} O(s^{2+\epsilon}) = O(n^{3+\epsilon})$ with any small $\epsilon > 0$.

4. A hybrid of PG and inexact projected regularized Newton methods

In the hybrid frameworks owing to Themelis et al. (2018) and Bareilles et al. (2023), the PG and Newton steps are alternating. Consider that the PG step is more cost-effective than the Newton step when the iterates are far from a stationary point. We introduce a switch condition (8) into our algorithm, a hybrid of PG and inexact projected regularized Newton methods (PGiPN) for problem (1), to control when the Newton steps are executed.

Now we describe the details of our algorithm. Let $x^k \in \Omega$ be the current iterate. It is noted that the PG step is always executed, and when condition (8) is met, we need to solve (11), which involves constructing G_k to satisfy (12). Such G^k can be easily achieved for the following two cases. One is the case that f can be expressed as $f(x) = h(Ax - b)$ for some $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and separable twice continuously differentiable h. Now $\nabla^2 f(x) = A^\top \nabla^2 h(Ax - b)A$ with $\nabla^2 h(Ax - b)$ being a diagonal matrix. Since $\nabla^2 f(x^k)$ is not necessarily positive definite, following the same way as in Liu et al. (2024), we construct

$$G^1_k := \nabla^2 f(x^k) + b_2[-\lambda_{\min}(\nabla^2 h(Ax^k - b))]_+ A^\top A + b_1\|\overline{\mu}_k(x^k - \overline{x}^k)\|^\sigma I \quad \text{with } b_2 \geq 1. \tag{35}$$

However, for highly nonconvex h, $[-\lambda_{\min}(\nabla^2 h(Ax^k-b))]_+$ is large, for which G_k^1 is a poor approximation to $\nabla^2 f(x^k)$. To avoid this drawback, we consider the following

$$G_k^2 := A^\top[\nabla^2 h(Ax^k-b)]_+ A + b_1\|\overline{\mu}_k(x^k-\overline{x}^k)\|^\sigma I. \tag{36}$$

When $\nabla^2 h(Ax^k-b)\succeq 0$, $G_k^1=G_k^2$. If $\nabla^2 h(Ax^k-b)\not\succeq 0$, it is immediate to see that $\|G_k^1-\nabla^2 f(x^k)\|_2\ge\|G_k^2-\nabla^2 f(x^k)\|_2$, which means that G_k^2 is a better approximation to $\nabla^2 f(x^k)$ than G_k^1. The other is the case that f has no special structure, and in this case we form $G_k:=G_k^3$ as in Ueda and Yamashita (2010) and Wu et al. (2023), where

$$G_k^3 := \nabla^2 f(x^k)+\big(b_2[-\lambda_{\min}(\nabla^2 f(x^k))]_+ + b_1\|\overline{\mu}_k(x^k-\overline{x}^k)\|^\sigma\big)I. \tag{37}$$

It is not hard to check that for $i=1,2,3$, G_k^i meets the requirement in (12). We remark here that the subsequent convergence analysis holds for the above three G_k^i, and we write them by G_k for simplicity. The iterates of PGiPN are described as follows.

Algorithm 2 (a hybrid of PG and inexact projected regularized Newton methods)

Initialization: Choose $\epsilon\ge 0$ and parameters $\mu_{\max}>\mu_{\min}>0$, $\tau>1$, $\alpha>0, b_1>0, b_2\ge 1$, $\varrho\in(0,\frac{1}{2})$, $\sigma\in(0,\frac{1}{2})$, $\varsigma\in(\sigma,1]$ and $\beta\in(0,1)$. Choose an initial $x^0\in\Omega$ and let $k:=0$.

PG Step:

(1a) Select $\mu_k\in[\mu_{\min},\mu_{\max}]$. Let m_k be the smallest nonnegative integer m such that

$$F(\overline{x}^k)\le F(x^k)-\frac{\alpha}{2}\|x^k-\overline{x}^k\|^2 \ \text{ with } \ \overline{x}^k\in \operatorname{prox}_{(\mu_k\tau^m)^{-1}g}(x^k-(\mu_k\tau^m)^{-1}\nabla f(x^k)). \tag{38}$$

(1b) Let $\overline{\mu}_k=\mu_k\tau^{m_k}$. If $\overline{\mu}_k\|x^k-\overline{x}^k\|\le\epsilon$, stop and output x^k; otherwise, go to step (1c).

(1c) If condition (8) holds, go to Newton step; otherwise, let $x^{k+1}=\overline{x}^k$. Set $k\leftarrow k+1$ and return to step (1a).

Newton step:

(2a) Seek an inexact solution y^k of (11) with G_k from (36) or (37) such that (13)-(14) hold.

(2b) Set $d^k:=y^k-x^k$. Let t_k be the smallest nonnegative integer t such that

$$f(x^k+\beta^t d^k)\le f(x^k)+\varrho\beta^t\langle\nabla f(x^k),d^k\rangle. \tag{39}$$

(2c) Let $\alpha_k=\beta^{t_k}$ with $x^{k+1}=x^k+\alpha_k d^k$. Set $k\leftarrow k+1$ and return to PG step.

Remark 12 (i) *Our PGiPN benefits from the PG step in two aspects. First, the incorporation of the PG step can guarantee that the sequence generated by PGiPN remains in a right position for convergence. Second, the PG step helps to identify adaptively the subspace used in the Newton step, and as will be shown in Proposition 16, when k is sufficiently large, switch condition* (8) *always holds and the supports of* $\{Bx^k\}_{k\in\mathbb{N}}$ *and* $\{x^k\}_{k\in\mathbb{N}}$ *keep unchanged, so that Algorithm 2 will reduce to an inexact projected regularized Newton method for solving* (10) *with* $\Pi_k\equiv\Pi_*$*, where* $\Pi_*\subset\mathbb{R}^n$ *is a polytope defined in* (49)*. In this sense, the PG step plays a crucial role in transforming the original challenging problem* (1) *into a problem that can be efficiently solved by the inexact projected regularized Newton method.*

(ii) *When x^k enters the Newton step, from the inexact criterion* (13) *and the expression of* Θ_k, $0 \geq \Theta_k(x^k+d^k) - \Theta_k(x^k) = \langle \nabla f(x^k), d^k\rangle + \frac{1}{2}\langle d^k, G_k d^k\rangle$, *and then*

$$\langle \nabla f(x^k), d^k\rangle \leq -\frac{1}{2}\langle d^k, G_k d^k\rangle \leq -\frac{b_1}{2}\|\overline{\mu}_k(x^k-\overline{x}^k)\|^\sigma \|d^k\|^2 < 0, \tag{40}$$

where the second inequality is due to (12)*. In addition, the inexact criterion* (13) *implies that* $y^k \in \Pi_k$*, which along with* $x^k \in \Pi_k$ *and the convexity of* Π_k *yields that* $x^k + \alpha d^k \in \Pi_k$ *for any* $\alpha \in (0,1]$*. By the definition of* Π_k, $\mathrm{supp}(B(x^k+\alpha d^k)) \subset \mathrm{supp}(Bx^k)$ *and* $\mathrm{supp}(x^k+\alpha d^k) \subset \mathrm{supp}(x^k)$*, so* $g(x^k+\alpha d^k) \leq g(x^k)$ *for any* $\alpha \in (0,1]$*. This together with* (40) *shows that the iterate along the direction* d^k *will reduce the value of* F *at* x^k*.*
(iii) *When* $\epsilon = 0$*, by Definition 1 the output* x^k *of Algorithm 2 is an L-stationary point of* (1)*, which is also a stationary point of problem* (10) *from Proposition 3 and Lemma 4 (i). Let* $r_k\colon \mathbb{R}^n \to \mathbb{R}^n$ *be the KKT residual mapping of* (10) *defined by*

$$r_k(x) := \overline{\mu}_k[x - \mathrm{proj}_{\Pi_k}(x - \overline{\mu}_k^{-1}\nabla f(x))]. \tag{41}$$

It is not difficult to verify that when x^k *satisfies condition* (8)*, the following relation holds*

$$r_k(x^k) = \overline{\mu}_k(x^k - \overline{x}^k). \tag{42}$$

Indeed, we only need to argue that $\overline{x}^k = \mathrm{proj}_{\Pi_k}(x^k - \overline{\mu}_k^{-1}\nabla f(x^k))$*. Suppose that this does not hold. Then, there exists* $\overline{z}^k \in \Pi_k$ *such that* $\widetilde{h}_k(\overline{z}^k) < \widetilde{h}_k(\overline{x}^k)$*, where* $\widetilde{h}_k(x) := \frac{\overline{\mu}_k}{2}\|x - (x^k - \overline{\mu}_k^{-1}\nabla f(x^k))\|^2$*. Since* $\overline{z}^k \in \Pi_k$*, we have* $\mathrm{supp}(B\overline{z}^k) \subset \mathrm{supp}(B\overline{x}^k)$ *and* $\mathrm{supp}(\overline{z}^k) \subset \mathrm{supp}(\overline{x}^k)$*, which implies that* $g(\overline{z}^k) \leq g(\overline{x}^k)$ *and then* $\widetilde{h}_k(\overline{z}^k) + g(\overline{z}^k) < \widetilde{h}_k(\overline{x}^k) + g(\overline{x}^k)$*, which yields a a contradiction to* $\overline{x}^k \in \mathrm{prox}_{\overline{\mu}_k^{-1}g}(x^k - \overline{\mu}_k^{-1}\nabla f(x^k))$*.*
(iv) *By using* (16) *and the descent lemma (Bertsekas, 1997, Proposition A.24), the line search in step (1a) must stop after a finite number of backtrackings. In fact, the line search in step (1a) is satisfied whenever the nonnegative integer* m *is such that* $\mu_k\tau^m \geq L_1 + \alpha$*, and consequently, for each* $k \in \mathbb{N}$, $\overline{\mu}_k = \mu_k\tau^{m_k} \leq \widetilde{\mu} := \tau(L_1+\alpha)$*.*
(v) *Note that problem* (11) *is a strongly convex quadratic program over a polyhedral set, for which many successful algorithms have been developed such as interior point algorithms. In our numerical experiments, we call the commercial software GUROBI (Gurobi Optimization, LLC (2024)) to solve it, which uses an interior point method as the solver.*

By Remark 12 (iv), to show that Algorithm 2 is well defined, we only need to argue that the Newton steps in Algorithm 2 are well defined, which is implied by the following lemma.

Lemma 13 *For each* $k \in \mathbb{N}$*, define the KKT residual mapping* $R_k\colon \mathbb{R}^n \to \mathbb{R}^n$ *of* (11) *by*

$$R_k(y) := \overline{\mu}_k[y - \mathrm{proj}_{\Pi_k}(y - \overline{\mu}_k^{-1}(G_k(y-x^k) + \nabla f(x^k)))].$$

Then, for those x^k*'s satisfying* (8)*, the following statements are true.*

(i) *For any* y *close enough to the optimal solution of* (11)*,* $y - \overline{\mu}_k^{-1}R_k(y)$ *satisfies* (13)*-*(14)*.*

(ii) *The line search step in* (39) *is well defined, and* $\alpha_k \geq \min\left\{1, \frac{(1-\varrho)b_1\beta}{L_1}\|\overline{\mu}_k(x^k - \overline{x}^k)\|^\sigma\right\}$*.*

(iii) *The inexact criterion* (14) *implies that* $\|R_k(y^k)\| \leq \frac{1}{2}\min\left\{\|r_k(x^k)\|, \|r_k(x^k)\|^{1+\varsigma}\right\}$*.*

Proof Pick any x^k satisfying (8). We proceed the proof of parts (i)-(iii) as follows.
(i) Let $\widehat{y}^k$ be the unique optimal solution to (11). Then $\widehat{y}^k \neq x^k$ (if not, x^k is the optimal solution of (11) and $0 = R_k(x^k) = r_k(x^k)$, which by (42) means that $x^k = \overline{x}^k$ and Algorithm 2 stops at x^k). By the optimality condition of (11), $-\nabla f(x^k) - G_k(\widehat{y}^k - x^k) \in \mathcal{N}_{\Pi_k}(\widehat{y}^k)$, which by the convexity of Π_k and $x^k \in \Pi_k$ implies that $\langle \nabla f(x^k) + G_k(\widehat{y}^k - x^k), \widehat{y}^k - x^k \rangle \leq 0$. Along with the expression of Θ_k, we have $\Theta_k(\widehat{y}^k) - \Theta_k(x^k) \leq -\frac{1}{2}\langle \widehat{y}^k - x^k, G_k(\widehat{y}^k - x^k)\rangle < 0$. Since Θ_k is continuous relative to Π_k, for any $z \in \Pi_k$ sufficiently close to $\widehat{y}^k$, $\Theta_k(z) - \Theta_k(x^k) \leq 0$. From $R_k(\widehat{y}) = 0$ and the continuity of R_k, when y sufficiently close to $\widehat{y}$, $y - \overline{\mu}_k^{-1} R_k(y)$ is close to $\widehat{y}$, which together with $y - \overline{\mu}_k^{-1} R_k(y) \in \Pi_k$ implies that $y - \overline{\mu}_k^{-1} R_k(y)$ satisfies the criterion (13). In addition, from the expression of R_k, for any $y \in \mathbb{R}^n$,

$$0 \in G_k(y - x^k) + \nabla f(x^k) - R_k(y) + \mathcal{N}_{\Pi_k}(y - \overline{\mu}_k^{-1} R_k(y)),$$

which by the expression of Θ_k implies that $\overline{\mu}_k^{-1} G_k R_k(y) + R_k(y) \in \partial \Theta_k(y - \overline{\mu}_k^{-1} R_k(y))$. Hence, $\mathrm{dist}(0, \partial\Theta_k(y - \overline{\mu}_k^{-1} R_k(y))) \leq \|\overline{\mu}_k^{-1} G_k R_k(y) + R_k(y)\|$. Noting that $R_k(\widehat{y}^k) = 0$, we have $\|\overline{\mu}_k^{-1} G_k R_k(\widehat{y}^k) + R_k(\widehat{y}^k)\| = 0 < \frac{\min\{\overline{\mu}_k^{-1}, 1\}}{2} \min\big\{\|\overline{\mu}_k(x^k - \overline{x}^k)\|, \|\overline{\mu}_k(x^k - \overline{x}^k)\|^{1+\varsigma}\big\}$. From the continuity of the function $y \mapsto \|\overline{\mu}_k^{-1} G_k R_k(y) + R_k(y)\|$, we conclude that for any y sufficiently close to $\widehat{y}^k$, $y - \overline{\mu}_k^{-1} R_k(y)$ satisfies the inexact criterion (14).
(ii) By (16) and the descent lemma (Bertsekas, 1997, Proposition A.24), for any $\alpha \in (0, 1]$,

$$\begin{aligned}
f(x^k + \alpha d^k) - f(x^k) - \varrho\alpha\langle \nabla f(x^k), d^k\rangle &\leq (1-\varrho)\alpha\langle \nabla f(x^k), d^k\rangle + \frac{L_1\alpha^2}{2}\|d^k\|^2 \\
&\leq -\frac{(1-\varrho)\alpha b_1}{2}\|\overline{\mu}_k(x^k - \overline{x}^k)\|^{\sigma}\|d^k\|^2 + \frac{L_1\alpha^2}{2}\|d^k\|^2 \\
&= \Big(-\frac{(1-\varrho)b_1}{2}\|\overline{\mu}_k(x^k - \overline{x}^k)\|^{\sigma} + \frac{L_1\alpha}{2}\Big)\alpha\|d^k\|^2,
\end{aligned}$$

where the second inequality uses (40). Therefore, when the nonnegative integer t is such that $\beta^t \leq \min\big\{1, \frac{(1-\varrho)b_1}{L_1}\|\overline{\mu}_k(x^k - \overline{x}^k)\|^{\sigma}\big\}$, the line search in (39) holds, which implies that the smallest nonnegative integer t_k should satisfy $\alpha_k = \beta^{t_k} \geq \min\big\{1, \frac{(1-\varrho)b_1\beta}{L_1}\|\overline{\mu}_k(x^k - \overline{x}^k)\|^{\sigma}\big\}$.
(iii) Let $\zeta^k \in \partial\Theta_k(y^k)$ be such that $\|\zeta^k\| = \mathrm{dist}(0, \partial\Theta_k(y^k))$. From $\zeta^k \in \partial\Theta_k(y^k)$ and the expression of Θ_k, we have $y^k = \mathrm{proj}_{\Pi_k}(y^k + \zeta^k - (G_k(y^k - x^k) + \nabla f(x^k)))$. Along with $y^k = \mathrm{proj}_{\Pi_k}(y^k)$ and the nonexpansiveness of proj_{Π_k}, $\|y^k - \mathrm{proj}_{\Pi_k}(y^k - (G_k(y^k - x^k) + \nabla f(x^k)))\| \leq \|\zeta^k\|$. Consequently,

$$\mathrm{dist}(0, \partial\Theta_k(y^k)) \geq \|y^k - \mathrm{proj}_{\Pi_k}(y^k - (G_k(y^k - x^k) + \nabla f(x^k)))\| \geq \min\{\overline{\mu}_k^{-1}, 1\}\|R_k(y^k)\|,$$

where the second inequality follows Lemma 4 of Sra (2012) and the expression of R_k. Combining the last inequality with (14) and (42) leads to the desired inequality. ■

When $\overline{\mu}_k = 1$, the condition that $\|R_k(y^k)\| \leq \frac{1}{2}\min\big\{\|r_k(x^k)\|, \|r_k(x^k)\|^{1+\varsigma}\big\}$ is a special case of the inexact condition in (Yue et al., 2019, Equa (6a)) or the inexact condition in (Mordukhovich et al., 2023, Equa (14)), which along with Lemma 13 (iii) shows that criterion (14) with $\overline{\mu}_k = 1$ is stronger than the ones adopted in these literature.

To analyze the convergence of Algorithm 2 with $\epsilon = 0$, henceforth we assume $x^k \neq \overline{x}^k$ for all k (if not, Algorithm 2 will produce an L-stationary point within finite number of steps, and its convergence holds automatically). From the iteration steps of Algorithm 2, we see that the sequence $\{x^k\}_{k\in\mathbb{N}}$ consists of two parts, $\{x^k\}_{k\in\mathcal{K}_1}$ and $\{x^k\}_{k\in\mathcal{K}_2}$, where

$$\mathcal{K}_1 := \mathbb{N}\backslash\mathcal{K}_2 \ \text{ with } \ \mathcal{K}_2 := \big\{k\in\mathbb{N} \mid \mathrm{supp}(Bx^k) = \mathrm{supp}(B\overline{x}^k),\ \mathrm{supp}(x^k) = \mathrm{supp}(\overline{x}^k)\big\}.$$

Obviously, $\mathcal{K}_1$ consists of those k's with x^{k+1} from the PG step, while $\mathcal{K}_2$ consists of those k's with x^{k+1} from the Newton step.

To close this section, we provide some properties of the sequences $\{x^k\}_{k\in\mathbb{N}}$ and $\{\overline{x}^k\}_{k\in\mathbb{N}}$.

Proposition 14 *The following assertions are true.*

(i) *The sequence $\{F(x^k)\}_{k\in\mathbb{N}}$ is descent and convergent.*

(ii) *There exists $\nu > 0$ such that $|B\overline{x}^k|_{\min} \ge \nu$ and $|\overline{x}^k|_{\min} \ge \nu$ for all $k \in \mathbb{N}$.*

(iii) *There exist $c_1, c_2 > 0$ such that $c_1\|r_k(x^k)\| \le \|d^k\| \le c_2\|r_k(x^k)\|^{1-\sigma}$ for all $k \in \mathcal{K}_2$.*

Proof **(i)** For each $k \in \mathbb{N}$, when $k \in \mathcal{K}_1$, by the line search in step (1a), $F(x^{k+1}) < F(x^k)$, and when $k \in \mathcal{K}_2$, from (39) and (40), it follows that $f(x^{k+1}) < f(x^k)$, which along with $g(x^{k+1}) \le g(x^k)$ by Remark 12 (ii) implies that $F(x^{k+1}) < F(x^k)$. Hence, $\{F(x^k)\}_{k\in\mathbb{N}}$ is a descent sequence. Recall that F is lower bounded on Ω, so $\{F(x^k)\}_{k\in\mathbb{N}}$ is convergent.
(ii) By the definition of $\overline{\mu}_k$ and Remark 12 (iv), $\overline{\mu}_k \in [\mu_{\min}, \widetilde{\mu}]$ for all $k \in \mathbb{N}$. Note that $\{x^k\}_{k\in\mathbb{N}} \subset \Omega$, so the sequence $\{x^k - \overline{\mu}_k^{-1}\nabla f(x^k)\}_{k\in\mathbb{N}}$ is bounded and is contained in a compact set, says, Ξ. By invoking Lemma 8 with such Ξ and $\underline{\mu} = \mu_{\min}, \overline{\mu} = \widetilde{\mu}$, there exists $\nu > 0$ (depending on Ξ, $\mu_{\min}$ and $\widetilde{\mu}$) such that $|[B; I]\overline{x}^k|_{\min} > \nu$. The desired result then follows by noting that $|B\overline{x}^k|_{\min} \ge |[B; I]\overline{x}^k|_{\min}$ and $|\overline{x}^k|_{\min} \ge |[B; I]\overline{x}^k|_{\min}$.
(iii) From the definition of G_k, the continuity of $\nabla^2 f$, $\{x^k\}_{k\in\mathbb{N}} \subset \Omega$, $\{\overline{x}^k\}_{k\in\mathbb{N}} \subset \Omega$ and Remark 12 (iv), there exists $\overline{c} > 0$ such that

$$\|G_k\|_2 \le \overline{c} \ \text{ for all } k \in \mathcal{K}_2. \tag{43}$$

Fix any $k \in \mathcal{K}_2$. By Lemma 13 (iii), $\|R_k(y^k)\| \le \frac{1}{2}\|r_k(x^k)\|$. Then, it holds that

$$\begin{aligned}
&\frac{1}{2}\|r_k(x^k)\| \le \|r_k(x^k)\| - \|R_k(y^k)\| \le \|r_k(x^k) - R_k(y^k)\| \\
&= \overline{\mu}_k\|x^k - \mathrm{proj}_{\Pi_k}(x^k - \overline{\mu}_k^{-1}\nabla f(x^k)) - y^k + \mathrm{proj}_{\Pi_k}(y^k - \overline{\mu}_k^{-1}(G_k(y^k - x^k) + \nabla f(x^k)))\| \\
&\le (2\overline{\mu}_k + \|G_k\|_2)\|y^k - x^k\| \le (2\widetilde{\mu} + \overline{c})\|d^k\|,
\end{aligned}$$

where the third inequality is using the nonexpansiveness of proj_{Π_k}, and the last one is due to (43) and $d^k = y^k - x^k$. Therefore, $c_1\|r_k(x^k)\| \le \|d^k\|$ with $c_1 := 1/(4\widetilde{\mu} + 2\overline{c})$. For the second inequality, it follows from the definitions of $r_k(\cdot)$ and $R_k(\cdot)$ that $R_k(y^k) - \nabla f(x^k) - G_k d^k \in \mathcal{N}_{\Pi_k}(y^k - \overline{\mu}_k^{-1}R_k(y^k))$ and $r_k(x^k) - \nabla f(x^k) \in \mathcal{N}_{\Pi_k}(x^k - \overline{\mu}_k^{-1}r_k(x^k))$, which together with the monotonicity of the set-valued mapping $\mathcal{N}_{\Pi_k}(\cdot)$ implies that

$$\begin{aligned}
\langle d^k, G_k d^k\rangle &\le \langle R_k(y^k) - r_k(x^k), d^k\rangle - \overline{\mu}_k^{-1}\|R_k(y^k) - r_k(x^k)\|^2 - \overline{\mu}_k^{-1}\langle G_k d^k, -R_k(y^k) + r_k(x^k)\rangle \\
&\le \langle (I + \overline{\mu}_k^{-1}G_k)d^k, R_k(y^k) - r_k(x^k)\rangle.
\end{aligned}$$

Combining this inequality with equations (12), (42) and Lemma 13 (iii) leads to

$$b_1\|r_k(x^k)\|^\sigma\|d^k\|^2 \le (1+\overline{\mu}_k^{-1}\|G_k\|_2)(\|R_k(y^k)\|+\|r_k(x^k)\|)\|d^k\| \le (3/2)(1+\overline{\mu}_k^{-1}\|G_k\|_2)\|r_k(x^k)\|\|d^k\|, \tag{44}$$

which along with (43) and $\mu_k \ge \mu_{\min}$ implies that $\|d^k\| \le \frac{3}{2}(1+\mu_{\min}^{-1}\overline{c})b_1^{-1}\|r_k(x^k)\|^{1-\sigma}$. Then, $\|d^k\| \le c_2\|r_k(x^k)\|^{1-\sigma}$ holds with $c_2 := \frac{3}{2}(1+\mu_{\min}^{-1}\overline{c})b_1^{-1}$. The proof is completed. ■

5. Convergence Analysis

Before analyzing the convergence of Algorithm 2, we show that it finally reduces to an inexact projected regularized Newton method for seeking a stationary point of a problem to minimize a smooth function over a polyhedral set. This requires the following lemma.

Lemma 15 *For the sequences $\{x^k\}_{k\in\mathbb{N}}$ and $\{\overline{x}^k\}_{k\in\mathbb{N}}$ generated by Algorithm 2, the following assertions are true.*

(i) *There exists a constant $\gamma > 0$ such that for each $k \in \mathbb{N}$,*

$$F(x^{k+1}) - F(x^k) \le \begin{cases} -\gamma\|x^k-\overline{x}^k\|^2 & \text{if } k\in\mathcal{K}_1,\\ -\gamma\|x^k-\overline{x}^k\|^{2+\sigma} & \text{if } k\in\mathcal{K}_2,\ \alpha_k=1,\\ -\gamma\|x^k-\overline{x}^k\|^{2+2\sigma} & \text{if } k\in\mathcal{K}_2,\ \alpha_k\ne 1.\end{cases} \tag{45}$$

(ii) $\lim_{k\to\infty}\|x^k-\overline{x}^k\| = 0$ *and* $\lim_{\mathcal{K}_2\ni k\to\infty}\|d^k\| = 0$.

(iii) *The accumulation point set of $\{x^k\}_{k\in\mathbb{N}}$, denoted by $\Gamma(x^0)$, is nonempty and compact, and every element of $\Gamma(x^0)$ is an L-stationary point of problem* (1).

Proof **(i)** Fix any $k \in \mathcal{K}_2$. From inequalities (39)-(40), Proposition 14 (iii) and (42),

$$f(x^{k+1}) - f(x^k) \le -\frac{\varrho b_1\alpha_k}{2}\|\overline{\mu}_k(x^k-\overline{x}^k)\|^\sigma\|d^k\|^2 \le -\frac{\varrho c_1^2 b_1\alpha_k}{2}\|\overline{\mu}_k(x^k-\overline{x}^k)\|^{2+\sigma} \le -\frac{\varrho c_1^2 b_1\alpha_k\mu_{\min}^{2+\sigma}}{2}\|x^k-\overline{x}^k\|^{2+\sigma} \tag{46}$$

which, along with $g(x^{k+1}) \le g(x^k)$ by Remark 12 (ii), implies that $F(x^{k+1}) - F(x^k) \le f(x^{k+1}) - f(x^k)$. Together with (46), we have $F(x^{k+1}) - F(x^k) \le -\frac{\varrho c_1^2 b_1\alpha_k\mu_{\min}^{2+\sigma}}{2}\|x^k-\overline{x}^k\|^{2+\sigma}$. Recall that $F(x^{k+1}) - F(x^k) \le \frac{\alpha}{2}\|x^k-\overline{x}^k\|^2$ for $k \in \mathcal{K}_1$. By using Lemma 13 (ii), the desired result then follows with $\gamma := \min\big\{\frac{\alpha}{2}, \frac{\varrho c_1^2 b_1\mu_{\min}^{2+\sigma}}{2}, \frac{\beta(1-\varrho)\varrho c_1^2 b_1^2\mu_{\min}^{2+2\sigma}}{2L_1}\big\}$.

(ii) Let $\widetilde{\mathcal{K}}_2 := \{k \in \mathcal{K}_2 \mid \alpha_k = 1\}$. Doing summation for inequality (45) from $i = 1$ to any $j \in \mathbb{N}$ yields that

$$\sum_{i\in\mathcal{K}_1\cap[j]}\gamma\|x^i-\overline{x}^i\|^2 + \sum_{i\in\widetilde{\mathcal{K}}_2\cap[j]}\gamma\|x^i-\overline{x}^i\|^{2+\sigma} + \sum_{i\in(\mathcal{K}_2\backslash\widetilde{\mathcal{K}}_2)\cap[j]}\gamma\|x^i-\overline{x}^i\|^{2+2\sigma} \le \sum_{i=1}^{j}\big[F(x^i)-F(x^{i+1})\big] = F(x^1)-F(x^{j+1}),$$

which by the lower boundedness of F on the set Ω implies that

$$\sum_{i\in\mathcal{K}_1}\|x^i-\overline{x}^i\|^2+\sum_{i\in\widetilde{\mathcal{K}}_2}\gamma\|x^i-\overline{x}^i\|^{2+\sigma}+\sum_{i\in\mathcal{K}_2\backslash\widetilde{\mathcal{K}}_2}\gamma\|x^i-\overline{x}^i\|^{2+2\sigma}<\infty.$$

Thus, we obtain $\lim_{k\to\infty}\|x^k-\overline{x}^k\|=0$. Together with (42), Proposition 14 (iii) and Remark 12 (iv), it follows that $\lim_{\mathcal{K}_2\ni k\to\infty}\|d^k\|=0$.

(iii) Recall that $\{x^k\}_{k\in\mathbb{N}}\subset\Omega$, so its accumulation point set $\Gamma(x^0)$ is nonempty. Pick any $x^*\in\Gamma(x^0)$. Then, there exists an index set $K\subset\mathbb{N}$ such that $\lim_{K\ni k\to\infty}x^k=x^*$. From part (ii), $\lim_{K\ni k\to\infty}\overline{x}^k=x^*$. By step (1a) and Remark 12 (iv), for each $k\in K$, $\overline{x}^k\in\mathrm{prox}_{\overline{\mu}_k^{-1}g}\big(x^k-\overline{\mu}_k^{-1}\nabla f(x^k)\big)$ with $\overline{\mu}_k\in[\mu_{\min},\widetilde{\mu}]$, and consequently,

$$0\in\overline{\mu}_k(\overline{x}^k-(x^k-\overline{\mu}_k^{-1}\nabla f(x^k)))+\partial g(\overline{x}^k).\tag{47}$$

We claim that $g(\overline{x}^k)\to g(x^*)$ as $K\ni k\to\infty$. Indeed, by the definition of $\overline{x}^k$, we have

$$\frac{\overline{\mu}_k}{2}\big\|\overline{x}^k-(x^k-\overline{\mu}_k^{-1}\nabla f(x^k))\big\|^2+g(\overline{x}^k)\le\frac{\overline{\mu}_k}{2}\big\|x^*-(x^k-\overline{\mu}_k^{-1}\nabla f(x^k))\big\|^2+g(x^*)\quad\forall k\in K.$$

Recall that $\overline{\mu}_k\in[\mu_{\min},\widetilde{\mu}]$ for each k. If necessary by taking a subsequence, we assume that $\overline{\mu}_k\to\overline{\mu}\in[\mu_{\min},\widetilde{\mu}]$ as $K\ni k\to\infty$. Passing $K\ni k\to\infty$ to the above inequality leads to

$$\begin{aligned}\limsup_{K\ni k\to\infty}g(\overline{x}^k)&\le\limsup_{K\ni k\to\infty}\Big[\frac{\overline{\mu}_k}{2}\big\|x^*-(x^k-\overline{\mu}_k^{-1}\nabla f(x^k))\big\|^2+g(x^*)\Big]\\&\quad+\limsup_{K\ni k\to\infty}\Big[-\frac{\overline{\mu}_k}{2}\big\|\overline{x}^k-(x^k-\overline{\mu}_k^{-1}\nabla f(x^k))\big\|^2\Big]=g(x^*),\end{aligned}$$

while $\liminf_{K\ni k\to\infty}g(\overline{x}^k)\ge g(x^*)$ follows from the lower semicontinuity of g. Thus, the claimed limit $\lim_{K\ni k\to\infty}g(\overline{x}^k)=g(x^*)$ holds. Now from the above inclusion (47), it follows that $0\in\nabla f(x^*)+\partial g(x^*)$. By Lemma 7, we know that x^* is an L-stationary point of (1). ■

Next we apply Lemma 15 (ii) to show that, after a finite number of iterations, the switch condition in (8) always holds and the Newton step is executed. To this end, define

$$T_k:=\mathrm{supp}(Bx^k),\ \overline{T}_k:=\mathrm{supp}(B\overline{x}^k),\ S_k:=\mathrm{supp}(x^k)\ \text{ and }\ \overline{S}_k:=\mathrm{supp}(\overline{x}^k).\tag{48}$$

Proposition 16 *For the index sets defined in* (48)*, there exist index sets* $T\subset[p],S\subset[n]$ *and* $\overline{k}\in\mathbb{N}$ *such that for all* $k>\overline{k}$*,* $T_k=\overline{T}_k=T$ *and* $S_k=\overline{S}_k=S$*, which means that* $k\in\mathcal{K}_2$ *for all* $k>\overline{k}$*. Moreover, for each* $x^*\in\Gamma(x^0)$*,* $\mathrm{supp}(Bx^*)=T,\mathrm{supp}(x^*)=S$ *and* $F(x^*)=\lim_{k\to\infty}F(x^k):=F^*$*, where* $\Gamma(x^0)$ *is defined in Lemma 15 (iii).*

Proof We complete the proof of the conclusion via the following three claims:

Claim 1: There exists $\overline{k}\in\mathbb{N}$ such that for $k>\overline{k}$, $|Bx^k|_{\min}\ge\frac{\nu}{2}$, where ν is the same as the one in Proposition 14 (ii). Indeed, for each $k-1\in\mathcal{K}_1$, $x^k=\overline{x}^{k-1}$, and $|Bx^k|_{\min}=|B\overline{x}^{k-1}|_{\min}\ge\nu>\frac{\nu}{2}$ follows by Proposition 14 (ii). Hence, it suffices to consider that $k-1\in\mathcal{K}_2$. By Lemma 15 (ii), there exists $\overline{k}\in\mathbb{N}$ such that for all $k\ge\overline{k}$, $\|x^{k-1}-\overline{x}^{k-1}\|<\frac{\nu}{4\|B\|_2}$, and for all $\mathcal{K}_2\ni k-1>\overline{k}-1$, $\|d^{k-1}\|<\frac{\nu}{4\|B\|_2}$, which implies that for $\mathcal{K}_2\ni k-1>\overline{k}-1$,

$\|Bx^{k-1} - B\overline{x}^{k-1}\| < \frac{\nu}{4}$ and $\|Bd^{k-1}\| < \frac{\nu}{4}$. For each $\mathcal{K}_2 \ni k-1 > \overline{k}-1$, let $i_k \in [p]$ be such that $|(Bx^{k-1})_{i_k}| = |Bx^{k-1}|_{\min}$. Since condition (8) implies that $\mathrm{supp}(Bx^{k-1}) = \mathrm{supp}(B\overline{x}^{k-1})$ for each $k-1 \in \mathcal{K}_2$, we have $|(B\overline{x}^{k-1})_{i_k}| \geq |B\overline{x}^{k-1}|_{\min}$. Thus, for each $\mathcal{K}_2 \ni k-1 > \overline{k}-1$,

$$\begin{aligned}\|Bx^{k-1} - B\overline{x}^{k-1}\| &\geq |(Bx^{k-1})_{i_k} - (B\overline{x}^{k-1})_{i_k}| \geq |(B\overline{x}^{k-1})_{i_k}| - |(Bx^{k-1})_{i_k}| \\ &\geq |B\overline{x}^{k-1}|_{\min} - |Bx^{k-1}|_{\min}.\end{aligned}$$

Recall that $|B\overline{x}^{k-1}|_{\min} \geq \nu$ for all $k \in \mathbb{N}$ by Proposition 14 (ii). Together with the last inequality and $\|Bx^{k-1} - B\overline{x}^{k-1}\| < \frac{\nu}{4}$, for each $\mathcal{K}_2 \ni k-1 > \overline{k}-1$, we have $|Bx^{k-1}|_{\min} \geq \frac{3\nu}{4}$. For each $\mathcal{K}_2 \ni k-1 > \overline{k}-1$, let $j_k \in [p]$ be such that $|(Bx^k)_{j_k}| = |Bx^k|_{\min}$. By Remark 12 (ii), $\mathrm{supp}(Bx^k) \subset \mathrm{supp}(Bx^{k-1})$ for each $k-1 \in \mathcal{K}_2$, which along with $j_k \in \mathrm{supp}(Bx^k)$ implies that $|(Bx^{k-1})_{j_k}| \geq |Bx^{k-1}|_{\min}$. Thus, for each $\mathcal{K}_2 \ni k-1 > \overline{k}-1$,

$$\begin{aligned}\|Bd^{k-1}\| = \frac{1}{\alpha_k}\|Bx^k - Bx^{k-1}\| &\geq \|Bx^k - Bx^{k-1}\| \geq |(Bx^{k-1})_{j_k} - (Bx^k)_{j_k}| \\ &\geq |(Bx^{k-1})_{j_k}| - |(Bx^k)_{j_k}| \geq |Bx^{k-1}|_{\min} - |Bx^k|_{\min},\end{aligned}$$

which together with $\|Bd^{k-1}\| \leq \frac{\nu}{4}$ and $|Bx^{k-1}|_{\min} \geq \frac{3\nu}{4}$ implies that $|Bx^k|_{\min} \geq \frac{\nu}{2}$.

Claim 2: $T_k = \overline{T}_k$ for $k > \overline{k}$. From the above arguments, $\|Bx^k - B\overline{x}^k\| \leq \frac{\nu}{4}$ for $k > \overline{k}$. If $i \in T_k$, then $|(B\overline{x}^k)_i| \geq |(Bx^k)_i| - \frac{\nu}{4} \geq \frac{\nu}{4}$, where the second inequality is using $|Bx^k|_{\min} > \frac{\nu}{2}$ by **Claim 1**. This means that $i \in \overline{T}_k$, so $T_k \subset \overline{T}_k$. Conversely, if $i \in \overline{T}_k$, then $|(Bx^k)_i| \geq |(B\overline{x}^k)_i| - \frac{\nu}{4} \geq \frac{3\nu}{4}$, so $i \in T_k$ and $\overline{T}_k \subset T_k$. Thus, $T_k = \overline{T}_k$ for $k > \overline{k}$.

Claim 3: $T_k = T_{k+1}$ for $k > \overline{k}$. If $k \in \mathcal{K}_1$, the result follows directly by the result in **Claim 2** because $\overline{T}_k = \mathrm{supp}(B\overline{x}^k) = \mathrm{supp}(Bx^{k+1}) = T_{k+1}$. If $k \in \mathcal{K}_2$, from the proof of **Claim 1**, $\|Bx^k - Bx^{k+1}\| \leq \|Bd^k\| \leq \frac{\nu}{4}$ for all $k > \overline{k}$. Then, if $i \in T_k$, $|(Bx^{k+1})_i| \geq |(Bx^k)_i| - \frac{\nu}{4} \geq \frac{\nu}{4}$, where the second inequality is using $|Bx^k|_{\min} > \frac{\nu}{2}$ by **Claim 1**. This implies that $i \in T_{k+1}$ and $T_k \subset T_{k+1}$. Conversely, if $i \in T_{k+1}$, then $|(Bx^k)_i| \geq |(Bx^{k+1})_i| - \frac{\nu}{4} \geq \frac{\nu}{4}$. Hence, $i \in T_k$ and $T_{k+1} \subset T_k$.

From **Claim 2** and **Claim 3**, there exists $T \subset [p]$ such that $T_k = \overline{T}_k = T$ for $k > \overline{k}$. Using the similar arguments can prove the existence of $S \subset [n]$ such that $S_k = \overline{S}_k = S$ for all $k > \overline{k}$ (if necessary increasing $\overline{k}$).

Pick any $x^* \in \Gamma(x^0)$. Let $\{x^k\}_{k\in K}$ be a subsequence such that $\lim_{K\ni k\to\infty} x^k = x^*$. By the above proof, for all sufficiently large $k \in K$, $|Bx^k|_{\min} \geq \frac{\nu}{2}$ and $|x^k|_{\min} \geq \frac{\nu}{2}$, which implies that $|Bx^*|_{\min} \geq \frac{\nu}{2}$ and $|x^*|_{\min} \geq \frac{\nu}{2}$. The results $\mathrm{supp}(Bx^*) = T$ and $\mathrm{supp}(x^*) = S$ can be obtained by a proof similar to **Claim 3**. From $x^* \in \Gamma(x^0)$, there exists an index set $K \subset \mathbb{N}$ such that $\lim_{K\ni k\to\infty} x^k = x^*$. From the above arguments, $g(x^k) = g(\overline{x}^k) = \lambda_1|T| + \lambda_2|S|$ for all $K \ni k \geq \overline{k}$. By the proof of Lemma 15 (iii), $\limsup_{K\ni k\to\infty} g(\overline{x}^k) \leq g(x^*)$, so that

$$\begin{aligned}F^* &= \limsup_{K\ni k\to\infty} F(x^k) = \limsup_{K\ni k\to\infty}[f(x^k) + g(x^k)] \\ &\leq f(x^*) + \limsup_{K\ni k\to\infty} g(x^k) = f(x^*) + \limsup_{K\ni k\to\infty} g(\overline{x}^k) \leq F(x^*).\end{aligned}$$

On the other hand, by the lower semicontinuity of F, we have $F^* \geq F(x^*)$. The two sides imply that $F(x^*) = F^*$. The proof is completed. ■

By Proposition 16, all $k > \overline{k}$ belong to $\mathcal{K}_2$, i.e., the sequence $\{x^{k+1}\}_{k>\overline{k}}$ is generated by the Newton step. This means that $\{x^{k+1}\}_{k>\overline{k}}$ is identical to the one generated by the inexact projected regularized Newton method starting from $x^{\overline{k}+1}$. Also, since $\Pi_k = \Pi_{\overline{k}+1}$ for all $k > \overline{k}$, Algorithm 2 finally reduces to the inexact projected regularized Newton method for solving

$$\min_{x\in\mathbb{R}^n} \ \phi(x) := f(x) + \delta_{\Pi_*}(x) \quad \text{with} \ \ \Pi_* := \Pi_{\overline{k}+1}, \tag{49}$$

which is a minimization problem of function f over the polytope Π_*, much simpler than the original problem (1). Consequently, the global convergence and local convergence rate analysis of PGiPN for model (1) boils down to analyzing those of the inexact projected regularized Newton method for (49). The rest of this section is devoted to this.

Unless otherwise stated, the notation $\overline{k}$ in the sequel is always that of Proposition 16 plus one. In addition, we require the assumption that $\nabla^2 f$ is locally Lipschitz continuous on $\Gamma(x^0)$, where $\Gamma(x^0)$ is defined in Lemma 15 (iii).

Assumption 1 *$\nabla^2 f$ is locally Lipschitz continuous on an open set containing $\Gamma(x^0)$.*

Assumption 1 is very standard when analyzing the convergence behavior of Newton-type method. The following lemma reveals that under this assumption, the step size α_k in Newton step takes 1 when k is sufficiently large. Since the proof is similar to that of Lemma B.1 of the arxiv version of Liu et al. (2024), the details are omitted here.

Lemma 17 *Suppose that Assumption 1 holds. Then $\alpha_k = 1$ for sufficiently large k.*

Notice that Π_* is a polytope, which can be expressed as

$$\Pi_* = \big\{x \in \mathbb{R}^n \mid B_{T^c_{\overline{k}+1}\cdot}x = 0,\ x_{S^c_{\overline{k}+1}} = 0,\ x \geq l,\ -x \geq -u\big\}. \tag{50}$$

For any $x \in \mathbb{R}^n$, we define multifunction $\mathcal{A} : \mathbb{R}^n \rightrightarrows [2n]$ as

$$\mathcal{A}(x) := \{i \mid x_i = l_i\} \cup \{i + n \mid x_i = u_i\}.$$

Clearly, for $x \in \Pi_*$, $\mathcal{A}(x)$ is the index set of those active constraints involved in Π_* at x. To prove the global convergence for PGiPN, we first show that $\mathcal{A}(x^k)$ keeps unchanged for sufficiently large k under the following non-degeneracy assumption.

Assumption 2 *For all $x^* \in \Gamma(x^0)$, $0 \in \nabla f(x^*) + \mathrm{ri}(\mathcal{N}_{\Pi_*}(x^*))$.*

It follows from Proposition 3 and Lemma 15 (iii) that for each $x^* \in \Gamma(x^0)$, x^* is a stationary point of F, which together with Proposition 16 and Lemma 4 (i) yields that $0 \in \nabla f(x^*) + \mathcal{N}_{\Pi_*}(x^*)$, so that Assumption 2 substantially requires that $-\nabla f(x^*)$ does not belong to the relative boundary[2] of $\mathcal{N}_{\Pi_*}(x^*)$. In the next lemma, we prove that under Assumptions 1-2, $\mathcal{A}(x^k) = \mathcal{A}(x^{k+1})$ for sufficiently large k.

2. For convex set $\Xi \subset \mathbb{R}^n$, the set difference $\mathrm{cl}(\Xi)\backslash\mathrm{ri}(\Xi)$ is called the relative boundary of Ξ, see (Rockafellar, 1970, p. 44).

Lemma 18 *Let $\{x^k\}_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 2. Suppose that Assumptions 1-2 hold. Then, there exist $\mathcal{A}^*\subset[2n]$ and a closed and convex cone $\mathcal{N}^*\subset\mathbb{R}^n$ such that $\mathcal{A}(x^k)=\mathcal{A}^*$ and $\mathcal{N}_{\Pi_*}(x^k)=\mathcal{N}^*$ for sufficiently large k.*

Proof We complete the proof via the following two claims.

Claim 1: $\lim_{k\to\infty}\|\mathrm{proj}_{\mathcal{T}_{\Pi_*}(x^k)}(-\nabla f(x^k))\|=0$. Since Π_* is polyhedral, for any $x\in\Pi_*$, $\mathcal{T}_{\Pi_*}(x)$ and $\mathcal{N}_{\Pi_*}(x)$ are closed and convex cones, and $\mathcal{T}_{\Pi_*}(x)$ is polar to $\mathcal{N}_{\Pi_*}(x)$, which implies that when k is sufficiently large, $z=\mathrm{proj}_{\mathcal{T}_{\Pi_*}(x^k)}(z)+\mathrm{proj}_{\mathcal{N}_{\Pi_*}(x^k)}(z)$ holds for any $z\in\mathbb{R}^n$. Then, for all sufficiently large k,

$$\begin{aligned}\|\mathrm{proj}_{\mathcal{T}_{\Pi_*}(x^k)}(-\nabla f(x^k))\| &= \|-\nabla f(x^k)-\mathrm{proj}_{\mathcal{N}_{\Pi_*}(x^k)}(-\nabla f(x^k))\|\\ &= \mathrm{dist}(0,\partial\phi(x^k))=\mathrm{dist}(0,\partial\phi(x^{k-1}+d^{k-1})),\end{aligned}$$

where the third equality is due to Lemma 17. Thus, it suffices to prove that

$$\lim_{k\to\infty}\mathrm{dist}(0,\partial\phi(x^k+d^k))=0.$$

For each $k\in\mathcal{K}_2$, by equation (14), there exists $\zeta_k\in\partial\Theta_k(y^k)=\partial\Theta_k(x^k+d^k)$ or equivalently $0\in\nabla f(x^k)+G_kd^k-\zeta_k+\mathcal{N}_{\Pi_k}(x^k+d^k)$ such that $\|\zeta_k\|$ is not more than the right hand side of (14). Invoking Remark 12 (iv) and Lemma 15 (ii) yields that $\lim_{k\to\infty}\|\zeta_k\|=0$. Moreover, from Proposition 16, for $k>\overline{k}$, the inclusion $0\in\nabla f(x^k)+G_kd^k-\zeta_k+\mathcal{N}_{\Pi_k}(x^k+d^k)$ is equivalent to $0\in\nabla f(x^k)+G_kd^k-\zeta_k+\mathcal{N}_{\Pi_*}(x^k+d^k)$. Note that $\partial\phi(x^k+d^k)=\nabla f(x^k+d^k)+\mathcal{N}_{\Pi_*}(x^k+d^k)$ for each $k>\overline{k}$. Then, $\nabla f(x^k+d^k)-\nabla f(x^k)-G_kd^k+\zeta_k\in\partial\phi(x^k+d^k)$ for each $k>\overline{k}$. This, by the continuity of ∇f, equation (43), Lemma 15 (ii), and $\lim_{k\to\infty}\|\zeta_k\|=0$, implies the desired limit $\lim_{k\to\infty}\mathrm{dist}(0,\partial\phi(x^k+d^k))=0$.

Claim 2: $\mathcal{A}(x^k)\subset\mathcal{A}(x^{k+1})$ for sufficiently large k. If not, there exists an infinite index set $K\subset\mathbb{N}$ such that $\mathcal{A}(x^k)\not\subset\mathcal{A}(x^{k+1})$ for all $k\in K$. If necessary taking a subsequence, we assume that $\{x^k\}_{k\in K}$ converges to x^*. By Lemma 15 (ii), $\{x^{k+1}\}_{k\in K}$ converges to x^*. In addition, from **Claim 1**, $\lim_{k\to\infty}\|\mathrm{proj}_{\mathcal{T}_{\Pi_*}(x^{k+1})}(-\nabla f(x^{k+1}))\|=0$. The two sides along with Assumption 2 and (Burke and Moré, 1988, Corollary 3.6) yields that $\mathcal{A}(x^{k+1})=\mathcal{A}(x^*)$ for all sufficiently large $k\in K$, contradicting to $\mathcal{A}(x^k)\not\subset\mathcal{A}(x^{k+1})$ for all $k\in K$. The claimed inclusion holds for sufficiently large k.

From $\mathcal{A}(x^k)\subset\mathcal{A}(x^{k+1})$ for sufficiently large k, $\{\mathcal{A}(x^k)\}_{k\in\mathbb{N}}$ converges to for some $\mathcal{A}^*\subset[2n]$ in the sense of Painlevé-Kuratowski[3]. From the discreteness of $\mathcal{A}^*$, we conclude that $\mathcal{A}(x^k)=\mathcal{A}^*$ for sufficiently large k. From the expression of Π_* in (50) and $\mathcal{A}(x^k)=\mathcal{A}^*$ for sufficiently large k, we have $\mathcal{N}_{\Pi_*}(x^k)=\mathcal{N}^*$ for sufficiently large k. ■

The global convergence of PGiPN additionally requires the following assumption.

Assumption 3 *For every sufficiently large k, there exists $\xi_k\in\mathcal{N}_{\Pi_*}(x^k)$ such that*

$$\liminf_{k\to\infty}\frac{-\langle\nabla f(x^k)+\xi_k,d^k\rangle}{\|\nabla f(x^k)+\xi_k\|\|d^k\|}>0.$$

3. A sequence of sets $\{C^k\}_{k\in\mathbb{N}}$ with $C^k\subset\mathbb{R}^n$ is said to converge in the sense of Painlevé-Kuratowski if its outer limit set $\limsup_{k\to\infty}C^k$ coincides with its inner limit set $\liminf_{k\to\infty}C^k$. On the definition of $\limsup_{k\to\infty}C^k$ and $\liminf_{k\to\infty}C^k$, see (Rockafellar and Wets, 2009, Definition 4.1).

This assumption essentially requires for every sufficiently large k the existence of one element $\xi_k \in \mathcal{N}_{\Pi_*}(x^k)$ such that the angle between $\nabla f(x^k) + \xi_k$ and d^k is uniformly larger than $\pi/2$. For sufficiently large k, since $x^k + \alpha d^k \in \Pi_*$ for all $\alpha \in [0, 1]$, we have $d^k \in \mathcal{T}_{\Pi^*}(x^k)$, which implies that $\langle \xi^k, d^k \rangle \leq 0$. Together with (40), for sufficiently large k, the angle between $\nabla f(x^k) + \xi_k$ and d^k is larger than $\pi/2$. This means that it is highly possible for Assumption 3 to hold. When $n = 1$, it automatically holds.

Next, we show that if ϕ is a KL function and Assumptions 1-3 hold, the sequence generated by PGiPN is Cauchy and converges to an L-stationary point.

Theorem 19 *Let $\{x^k\}_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 2. Suppose that Assumptions 1-3 hold, and that ϕ is a KL function. Then, $\sum_{k=1}^{\infty} \|x^{k+1} - x^k\| < \infty$, and consequently $\{x^k\}_{k\in\mathbb{N}}$ converges to an L-stationary point of* (1).

Proof By Proposition 16 and the expressions of F and ϕ, we have $F(x^k) = \phi(x^k) + \lambda_1|T| + \lambda_2|S|$ for all $k > \overline{k}$. Along with Lemma 15 (i), the sequence $\{\phi(x^k)\}_{k>\overline{k}}$ is nonincreasing. If there exists $\widetilde{k} > \overline{k}$ such that $\phi(x^{\widetilde{k}}) = \phi(x^{\widetilde{k}+1})$, then $F(x^{\widetilde{k}}) = F(x^{\widetilde{k}+1})$, which along with Lemma 15 (i) leads to $x^{\widetilde{k}} = \overline{x}^{\widetilde{k}}$. Then, $x^{\widetilde{k}}$ meets the termination condition of Algorithm 2, so $\{x^k\}_{k\in\mathbb{N}}$ converges to an L-stationary point of (1) within a finite number of steps. Thus, we only need to consider the case that $\phi(x^k) > \phi(x^{k+1})$ for all $k > \overline{k}$. By Proposition 16, for any $x \in \Gamma(x^0)$, $F^* = F(x) = \phi(x) + \lambda_1|T| + \lambda_2|S|$ or equivalently $\phi(x) = \phi^* := F^* - \lambda_1|T| - \lambda_2|S|$. By (Bolte et al., 2014, Lemma 6), there exist $\varepsilon > 0, \eta > 0$ and a continuous concave function $\varphi \in \Upsilon_\eta$ such that for all $\overline{x} \in \Gamma(x^0)$ and $x \in \{z \in \mathbb{R}^n \mid \mathrm{dist}(z, \Gamma(x^0)) < \varepsilon\} \cap [\phi^* < \phi < \phi^* + \eta]$, $\varphi'(\phi(x) - \phi^*)\mathrm{dist}(0, \partial\phi(x)) \geq 1$ where Υ_η is defined in Definition 6. Then, for $k > \overline{k}$ (if necessary by increasing $\overline{k}$), $x^k \in \{z \in \mathbb{R}^n \mid \mathrm{dist}(z, \Gamma(x^0)) < \varepsilon\} \cap [\phi^* < \phi < \phi^* + \eta]$, so

$$\varphi'(\phi(x^k) - \phi^*)\mathrm{dist}(0, \partial\phi(x^k)) \geq 1. \tag{51}$$

By Assumption 3, there exist $c > 0$ and $\xi_k \in \mathcal{N}_{\Pi_*}(x^k)$ such that for suffciently large k,

$$-\langle \nabla f(x^k) + \xi_k, d^k \rangle > c\|\nabla f(x^k) + \xi_k\|\|d^k\|. \tag{52}$$

From Lemma 18, $\mathcal{N}_{\Pi_*}(x^k) = \mathcal{N}_{\Pi_*}(x^{k+1})$ for all $k > \overline{k}$ (by possibly enlarging $\overline{k}$), which implies that $\xi_k \in \mathcal{N}_{\Pi_*}(x^{k+1})$. Together with (39), (52) and Lemma 17, for all $k > \overline{k}$ (if necessary enlarging $\overline{k}$), it holds that

$$\frac{\phi(x^k) - \phi(x^{k+1})}{\mathrm{dist}(0, \partial\phi(x^k))} \geq \frac{-\varrho\langle \nabla f(x^k) + \xi_k, d^k \rangle}{\mathrm{dist}(0, \partial\phi(x^k))} \geq \frac{\varrho c\|\nabla f(x^k) + \xi_k\|\|d^k\|}{\|\nabla f(x^k) + \xi_k\|} = \varrho c\|x^{k+1} - x^k\|, \tag{53}$$

where the second inequality follows by $\nabla f(x^k) + \xi_k \in \partial\phi(x^k)$ and (52). For each $k > \overline{k}$, let $\Delta_k := \varphi(\phi(x^k) - \phi^*)$. From (51), (53) and the concavity of φ on $[0, \eta)$, for all $k > \overline{k}$,

$$\begin{aligned}
\Delta_k - \Delta_{k+1} = \phi(x^k) - \phi(x^{k+1}) &\geq \varphi'(\phi(x^k) - \phi^*)(\phi(x^k) - \phi(x^{k+1})) \\
&\geq \frac{\phi(x^k) - \phi(x^{k+1})}{\mathrm{dist}(0, \partial\phi(x^k))} \geq \varrho c\|x^{k+1} - x^k\|.
\end{aligned}$$

Summing this inequality from $\overline{k}$ to any $k > \overline{k}$ and using $\Delta_k \geq 0$ yields that

$$\sum_{j=\overline{k}}^{k} \|x^{j+1} - x^j\| \leq \frac{1}{\varrho c}\sum_{j=\overline{k}}^{k}(\Delta_j - \Delta_{j+1}) = \frac{1}{\varrho c}(\Delta_{\overline{k}} - \Delta_{k+1}) \leq \frac{1}{\varrho c}\Delta_{\overline{k}}.$$

Passing the limit $k \to \infty$ leads to $\sum_{j=\overline{k}}^{\infty} \|x^{j+1} - x^j\| < \infty$. Thus, $\{x^k\}_{k\in\mathbb{N}}$ is a Cauchy sequence and converges to x^*. It follows from Lemma 15 (iii) that x^* is an L-stationary point of model (1). The proof is completed. ■

Remark 20 *Since Π_* is a semi-algebraic set, the function δ_{Π_*} is semi-algebraic. According to the comments in Section 2.2, the function ϕ is necessarily a KL function whenever f is definable in an o-minimal structure over the real field; for example, the least-squares loss function f in Section 6.2, the logarithmic loss function f in Section 6.3, the logistic regression loss, and the high order portfolio loss function (Zhou and Palomar (2021)) are all definable in an o-minimal structure over the real field.*

Next we focus on the superlinear rate analysis of PGiPN. For this purpose, define

$$\mathcal{X}^* := \big\{x \in \Gamma(x^0) \mid 0 \in \nabla f(x) + \mathcal{N}_{\Pi_*}(x),\ \nabla^2 f(x) \succeq 0\big\},$$

which is called the set of second-order stationary points of (49). By Lemma 15 (iii) and Proposition 3, the set $\mathcal{X}^*$ is generally smaller than the set of stationary points of (1). We assume that a local Hölderian error bound condition holds with respect to (w.r.t.) $\mathcal{X}^*$ in Assumption 4. For more introduction on the Hölderian error bound condition, we refer the interested readers to Mordukhovich et al. (2023) and Liu et al. (2024).

Assumption 4 *The mapping $\mathbb{R}^n \ni x \mapsto r(x) := x - \mathrm{proj}_{\Pi_*}(x - \nabla f(x))$ has the q-subregularity with $q \in (0, 1]$ at any $\overline{x} \in \Gamma(x^0)$ for the origin w.r.t. the set $\mathcal{X}^*$, i.e., for every $\overline{x} \in \Gamma(x^0)$, there exist $\varepsilon > 0$ and $\kappa > 0$ such that for all $x \in \mathbb{B}(\overline{x}, \varepsilon)$, $\mathrm{dist}(x, \mathcal{X}^*) \le \kappa \|r(x)\|^q$.*

Recently, Liu et al. (2024) proposed an inexact regularized proximal Newton method (IRPNM) for solving the composite problem, the minimization of the sum of a twice continuously differentiable function and an extended real-valued convex function, which includes (49) as a special case. They established the superlinear convergence rate of IRPNM under Assumption 1, and Assumption 4 with proj_{Π_*} replaced by the proximal mapping of the convex function. By (Sra, 2012, Lemma 4) and $\overline{\mu}_k \in [\mu_{\min}, \widetilde{\mu}]$, $\|r(x^k)\| = O(\|r_k(x^k)\|)$ for sufficiently large k. This together with Assumption 4 implies that for every $\overline{x} \in \Gamma(x^0)$, there exist $\varepsilon > 0$ and $\widehat{\kappa} > 0$ such that for sufficiently large k with $x^k \in \mathbb{B}(\overline{x}, \varepsilon)$,

$$\mathrm{dist}(x^k, \mathcal{X}^*) \le \widehat{\kappa} \|r_k(x^k)\|^q. \tag{54}$$

Recall that PGiPN finally reduces to an inexact projected regularized Newton method for solving (49). From Lemma 13 (iii) and Lemma 17, for sufficiently large k,

$$\Theta_k(x^{k+1}) - \Theta_k(x^k) \le 0 \ \text{ and } \ \|R_k(x^{k+1})\| \le \frac{1}{2} \min\{\|r_k(x^k)\|, \|r_k(x^k)\|^{1+\varsigma}\}. \tag{55}$$

Let $\Lambda_k^i := G_k^i - \nabla^2 f(x^k) - b_1 \|\overline{\mu}_k(x^k - \overline{x}^k)\|^\sigma I$ with G_k^i given by (35)-(37). Under Assumption 4, from (Wu et al., 2023, Lemma 4.8), (Liu et al., 2024, Lemma 4.4), and the fact that $G_k^1 - G_k^2 \succeq 0$, it holds that for sufficiently large k,

$$\max\big\{\|\Lambda_k^1\|_2, \|\Lambda_k^2\|_2, \|\Lambda_k^3\|_2\big\} = O(\mathrm{dist}(x^k, \mathcal{X}^*)). \tag{56}$$

In the rest of this section, for completeness, we provide the proof of the superlinear convergence of PGiPN under Assumptions 1 and 4 though it is implied by that of Liu et al. (2024). To this end, for each $k \in \mathcal{K}_2$, define $\widetilde{x}^k$, $\widehat{x}^k$ and f_k as follows.

$$f_k(x) := f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{2}(x - x^k)^\top G_k (x - x^k);$$
$$\widetilde{x}^k : \text{the exact solution to problem (11)}; \quad \widehat{x}^k \in \mathrm{proj}_{\mathcal{X}^*}(x^k).$$

We first bound the gap between y^k and $\widetilde{x}^k$ from above in terms of $\|r_k(x^k)\|$.

Lemma 21 *There exist $\gamma_1 > 0$ and $\gamma_2 > 0$ such that for every $k \in \mathcal{K}_2$, $\|y^k - \widetilde{x}^k\| \le \gamma_1 \|r_k(x^k)\|^{1+\varsigma} + \gamma_2 \|r_k(x^k)\|^{1+\varsigma-\sigma}$.*

Proof Fix any $k \in \mathcal{K}_2$. Recall that $R_k(y^k) = \overline{\mu}_k[y^k - \mathrm{proj}_{\Pi_*}(y^k - \overline{\mu}_k^{-1}\nabla f_k(y^k))]$. Invoking the relation $\mathrm{proj}_{\Pi_*} = (\mathcal{I} + \mathcal{N}_{\Pi_*})^{-1}$ by the convexity of Π_*, where $\mathcal{I}$ is the identity mapping, we have $R_k(y^k) - \nabla f_k(y^k) \in \mathcal{N}_{\Pi_*}(y^k - \overline{\mu}_k^{-1} R_k(y^k))$. Along with $\Theta_k = f_k + \delta_{\Pi_*}$, it holds

$$R_k(y^k) + \nabla f_k(y^k - \overline{\mu}_k^{-1} R_k(y^k)) - \nabla f_k(y^k) \in \partial \Theta_k(y^k - \overline{\mu}_k^{-1} R_k(y^k)).$$

Note that $\nabla f_k(x) = \nabla f(x^k) + G_k(x - x^k)$. The above inclusion can be simplified as

$$(I - \overline{\mu}_k^{-1} G_k) R_k(y^k) \in \partial \Theta_k(y^k - \overline{\mu}_k^{-1} R_k(y^k)).$$

On the other hand, from the definition of $\widetilde{x}^k$, we have $0 \in \partial\Theta_k(\widetilde{x}^k)$. Together with the above inclusion and the strong monotoncity of $\partial\Theta_k$ with model $b_1\|r_k(x^k)\|^\sigma$, it follows that

$$\left\langle (I - \overline{\mu}_k^{-1} G_k) R_k(y^k), y^k - \overline{\mu}_k^{-1} R_k(y^k) - \widetilde{x}^k \right\rangle \ge b_1 \|r_k(x^k)\|^\sigma \|y^k - \overline{\mu}_k^{-1} R_k(y^k) - \widetilde{x}^k\|^2.$$

Using the Cauchy-Schwarz inequality leads to

$$\begin{aligned} &\|y^k - \overline{\mu}_k^{-1} R_k(y^k) - \widetilde{x}^k\| \le (b_1^{-1} \|r_k(x^k)\|^{-\sigma}) \|(I - \overline{\mu}_k^{-1} G_k) R_k(y^k)\| \\ &\le \frac{1}{2b_1 \|r_k(x^k)\|^\sigma}(1 + \mu_{\min}^{-1}\|G_k\|_2)\|r_k(x^k)\|^{1+\varsigma} \le \frac{(1+\mu_{\min}^{-1}\overline{c})}{2b_1}\|r_k(x^k)\|^{1+\varsigma-\sigma}, \end{aligned}$$

where the second inequality is due to (55) and $\overline{\mu}_k \ge \mu_{\min}$, and the third is by (43). Note that $y^k = x^{k+1}$ by Lemma 17. From the above inequality and the second one of (55),

$$\|y^k - \widetilde{x}^k\| \le \frac{1}{2\mu_{\min}}\|r_k(x^k)\|^{1+\varsigma} + \frac{(1+\mu_{\min}^{-1}\overline{c})}{2b_1}\|r_k(x^k)\|^{1+\varsigma-\sigma},$$

and the desired result holds with $\gamma_1 := \frac{1}{2\mu_{\min}}$ and $\gamma_2 := \frac{(1+\mu_{\min}^{-1}\overline{c})}{2b_1}$. ∎

The following lemma bounds the gap between x^k and $\widetilde{x}^k$ by following the similar line of (Liu et al., 2024, Lemma 6).

Lemma 22 *Consider any $\overline{x} \in \Gamma(x^0)$. Under Assumptions 1 and 4, there exist $\epsilon_1 > 0$ and $L_2 > 0$ such that for all $x^k \in \mathbb{B}(\overline{x}, \epsilon_1)$,*

$$\|x^k - \widetilde{x}^k\| \le \left(\frac{L_2 \mathrm{dist}(x^k, \mathcal{X}^*)}{2b_1\|r_k(x^k)\|^\sigma} + \frac{\|\Lambda_k\|_2}{b_1\|r_k(x^k)\|^\sigma} + 2\right)\mathrm{dist}(x^k, \mathcal{X}^*).$$

Proof From Assumption 1, there exist $\epsilon_0 > 0$ and $L_2 > 0$ such that for any $x, x' \in \mathbb{B}(\overline{x}, \epsilon_0)$,

$$\|\nabla^2 f(x) - \nabla^2 f(x')\| \le L_2 \|x - x'\|. \tag{57}$$

From Assumption 4, $\overline{x} \in \mathcal{X}^*$. Recall that $\widehat{x}^k \in \mathrm{proj}_{\mathcal{X}^*}(x^k)$. By taking $\epsilon_1 = \epsilon_0/2$, for $x^k \in \mathbb{B}(\overline{x}, \epsilon_1)$, it holds $\|\widehat{x}^k - \overline{x}\| \le \|x^k - \widehat{x}^k\| + \|x^k - \overline{x}\| \le 2\|x^k - \overline{x}\| \le \epsilon_0$. Therefore, for $x^k \in \mathbb{B}(\overline{x}, \epsilon_1)$, we deduce from (57) that

$$\begin{aligned} &\|\nabla f(\widehat{x}^k) - \nabla f(x^k) - \nabla^2 f(x^k)(\widehat{x}^k - x^k)\| \\ &= \left\| \int_0^1 [\nabla^2 f(x^k + t(\widehat{x}^k - x^k)) - \nabla^2 f(x^k)](\widehat{x}^k - x^k) dt \right\| \le \frac{L_2}{2} \|\widehat{x}^k - x^k\|^2. \end{aligned} \tag{58}$$

By the definition of $\widetilde{x}^k$, $0 \in \nabla f(x^k) + G_k(\widetilde{x}^k - x^k) + \mathcal{N}_{\Pi_*}(\widetilde{x}^k)$; while by the definition of $\widehat{x}^k$, $0 \in \nabla f(\widehat{x}^k) + \mathcal{N}_{\Pi_*}(\widehat{x}^k)$. Using the monotoncity of $\mathcal{N}_{\Pi_*}$ results in

$$\begin{aligned} 0 &\le \langle \nabla f(x^k) + G_k(\widetilde{x}^k - x^k) - \nabla f(\widehat{x}^k), \widehat{x}^k - \widetilde{x}^k \rangle \\ &= \langle \nabla f(x^k) + G_k(\widehat{x}^k - x^k) - \nabla f(\widehat{x}^k), \widehat{x}^k - \widetilde{x}^k \rangle - \langle G_k(\widehat{x}^k - \widetilde{x}^k), \widehat{x}^k - \widetilde{x}^k \rangle. \end{aligned}$$

This implies that

$$\begin{aligned} &b_1 \|r_k(x^k)\|^\sigma \|\widehat{x}^k - \widetilde{x}^k\| \le \lambda_{\min}(G_k) \|\widehat{x}^k - \widetilde{x}^k\| \le \|\nabla f(\widehat{x}^k) - \nabla f(x^k) - G_k(\widehat{x}^k - x^k)\| \\ &\le \|\nabla f(\widehat{x}^k) - \nabla f(x^k) - \nabla^2 f(x^k)(\widehat{x}^k - x^k)\| + \|\Lambda_k\|_2 \|\widehat{x}^k - x^k\| + b_1 \|r_k(x^k)\|^\sigma \|\widehat{x}^k - x^k\| \\ &\le \frac{L_2}{2} \|\widehat{x}^k - x^k\|^2 + \|\Lambda_k\|_2 \|\widehat{x}^k - x^k\| + b_1 \|r_k(x^k)\|^\sigma \|\widehat{x}^k - x^k\|, \end{aligned}$$

where the first inequality is by the expression of G_k and (42), and the fourth follows (58). Rearranging the above inequality, we obtain $\|\widehat{x}^k - \widetilde{x}^k\| \le \frac{L_2 \|\widehat{x}^k - x^k\|^2}{2b_1 \|r_k(x^k)\|^\sigma} + \frac{\|\Lambda_k\|_2 \|\widehat{x}^k - x^k\|}{b_1 \|r_k(x^k)\|^\sigma} + \|\widehat{x}^k - x^k\|$. Then, the desired result holds by the triangle inequality and $\|x^k - \widehat{x}^k\| = \mathrm{dist}(x^k, \mathcal{X}^*)$. ■

Now we are ready to establish the supelinear convergence rate of the sequence. It is noted that the proof is similar to that of (Liu et al., 2024, Theorem 6).

Theorem 23 *Fix any $\overline{x} \in \Gamma(x^0)$. Suppose that Assumption 1 holds, and Assumption 4 holds with $q \in (\frac{1}{1+\sigma}, 1]$. Then, the sequence $\{x^k\}_{k \in \mathbb{N}}$ converges to $\overline{x}$ with the Q-superlinear convergence rate of order $q(1+\sigma)$.*

Proof If necessary enlarging $\overline{k}$, we assume that $x^k \in \mathbb{B}(\overline{x}, \epsilon_1)$ for $k > \overline{k}$, where ϵ_1 is the one in Lemma 22. From the definition of r_k, we have $r_k(\widehat{x}^k) = 0$ for $k > \overline{k}$. This together with the nonexpansive property of proj_{Π_*} yields that

$$\begin{aligned} \|r_k(x^k)\| &= \overline{\mu}_k \|x^k - \mathrm{proj}_{\Pi_*}(x^k - \overline{\mu}_k^{-1} \nabla f(x^k)) - \widehat{x}^k + \mathrm{proj}_{\Pi_*}(\widehat{x}^k - \overline{\mu}_k^{-1} \nabla f(\widehat{x}^k))\| \\ &\le (2\overline{\mu}_k + L_1) \mathrm{dist}(x^k, \mathcal{X}^*) \le (2\widetilde{\mu} + L_1) \mathrm{dist}(x^k, \mathcal{X}^*). \end{aligned} \tag{59}$$

In view of equation (56), if necessary enlarging $\overline{k}$, there exists $\gamma_3 > 0$ such that for $k > \overline{k}$,

$$\|\Lambda_k\|_2 \le \gamma_3 \mathrm{dist}(x^k, \mathcal{X}^*). \tag{60}$$

From $\|d^k\| = \|y^k - x^k\| \le \|y^k - \widetilde{x}^k\| + \|\widetilde{x}^k - x^k\|$, Lemmas 21- 22, Assumption 4 and equations (59)-(60), if necessary enlarging $\overline{k}$, there exists $\gamma_4 > 0$ such that for $k > \overline{k}$,

$$\|d^k\| \le \gamma_4 \mathrm{dist}(x^k, \mathcal{X}^*). \tag{61}$$

In addition, by virtue of equation (54) and Lemma 17, we obtain

$$\begin{aligned}\mathrm{dist}(x^{k+1}, \mathcal{X}^*) &\le \widehat{\kappa}\|r_k(x^{k+1})\|^q = \widehat{\kappa}\left[\|r_k(x^{k+1})\| - \|R_k(x^{k+1})\| + \|R_k(y^k)\|\right]^q \\ &\le \widehat{\kappa}\left[\|r_k(x^{k+1})\| - \|R_k(x^{k+1})\| + \frac{1}{2}\|r_k(x^k)\|^{1+\varsigma}\right]^q \\ &\le \widehat{\kappa}\left[\|r_k(x^{k+1})\| - \|R_k(x^{k+1})\| + \frac{1}{2}(2\widetilde{\mu} + L_1)^{1+\varsigma}\mathrm{dist}(x^k, \mathcal{X}^*)^{1+\varsigma}\right]^q,\end{aligned} \tag{62}$$

where the third inequality is due to (59). Next we bound the term $|\|r_k(x^{k+1})\| - \|R_k(x^{k+1})\||$. If necessary enlarging $\overline{k}$, we have for $k > \overline{k}$,

$$\begin{aligned}&\Big|\|r_k(x^{k+1})\| - \|R_k(x^{k+1})\|\Big| \le \|\nabla f(x^{k+1}) - \nabla f(x^k) - G_k(x^{k+1} - x^k)\| \\ &\le \frac{L_2}{2}\|x^{k+1} - x^k\|^2 + \|\Lambda_k\|_2\|x^{k+1} - x^k\| + b_1\|r_k(x^k)\|^\sigma\|x^{k+1} - x^k\| \\ &\le \frac{L_2}{2}\|d^k\|^2 + \gamma_3\|d^k\|\mathrm{dist}(x^k, \mathcal{X}^*) + b_1(2\widetilde{\mu} + L_1)^\sigma\|d^k\|\mathrm{dist}(x^k, \mathcal{X}^*)^\sigma \\ &\le \left(\frac{L_2\gamma_4^2}{2} + \gamma_3\gamma_4\right)\mathrm{dist}(x^k, \mathcal{X}^*)^2 + b_1\gamma_4(2\widetilde{\mu} + L_1)^\sigma\mathrm{dist}(x^k, \mathcal{X}^*)^{1+\sigma},\end{aligned}$$

where the first inequality is by the definitions of r_k and R_k and the nonexpansive property of $\mathrm{proj}_{\Pi_k} = \mathrm{proj}_{\Pi_*}$, the second one follows Assumption 1 and similar arguments for (58), the third one follows equations (59)-(60), and the fourth is by (61). By combining the above inequality and (62) and letting $\gamma_5 := \frac{L_2\gamma_4^2}{2} + \gamma_3\gamma_4$, $\gamma_6 := b_1\gamma_4(2\widetilde{\mu} + L_1)^\sigma$ and $\gamma_7 := \frac{1}{2}(2\widetilde{\mu} + L_1)^{1+\varsigma}$, it holds that for $k > \overline{k}$ (if necessary enlarging $\overline{k}$),

$$\begin{aligned}\mathrm{dist}(x^{k+1}, \mathcal{X}^*) &\le \widehat{\kappa}\left[\gamma_5\mathrm{dist}(x^k, \mathcal{X}^*)^2 + \gamma_6\mathrm{dist}(x^k, \mathcal{X}^*)^{1+\sigma} + \gamma_7\mathrm{dist}(x^k, \mathcal{X}^*)^{1+\varsigma}\right]^q \\ &\le \widehat{\kappa}(\gamma_5 + \gamma_6 + \gamma_7)^q\mathrm{dist}(x^k, \mathcal{X}^*)^{q(1+\sigma)},\end{aligned} \tag{63}$$

where the last inequality follows by $\lim_{k\to\infty}\mathrm{dist}(x^k, \mathcal{X}^*) = 0$ and $\sigma \le \varsigma \le 1$. The proof for the result that $\{x^k\}_{k\in\mathbb{N}}$ converges to $\overline{x}$ at a superlinear convergence rate is similar to the proof of (Liu et al., 2024, Theorem 6), and the details are omitted here. ■

Remark 24 *When f is convex, $\mathcal{X}^*$ reduces to the set of L-stationary points of* (49)*. In this case, by Lemma 7, the local Hölderian error bound with $q = 1$ in Assumption 4 is precisely the metric subregularity of the residual mapping r at x^* for* 0*, which is equivalent to that of $\partial\phi$ at x^* for* 0 *by (Liu et al., 2024, Lemma 1). Due to the polyhedrality of* Π_**, the latter holds when $f(\cdot) = h(A\,\cdot)$ for some $A \in \mathbb{R}^{m\times n}$ and a continuously differentiable strictly convex h by following the same arguments as those for (Zhou and So, 2017, Theorem*

2). Thus, when $h(u) = \frac{1}{2}\|u\|^2$ or $h(u) = \frac{1}{m}\sum_{i=1}^m \log(1+\exp(-b_iu_i))$ for $u \in \mathbb{R}^m$, i.e., f is the popular least-squares function or logistic regression function, Assumption 4 holds automatically. In addition, when f is a piece-wise linear quadratic function, since $\partial\phi$ is a polyhedral multifunction, the error bound condition automatically holds by (Robinson, 1981, Proposition 1). Such loss functions, covering the Huber loss, the ℓ_1-norm loss, the MCP and SCAD loss, are often used to deal with outliers or heavy-tailed noise.

6. Numerical experiments

This section focuses on the numerical experiments of several variants of PGiPN for solving a fused ℓ_0-norms regularization problem with a box constraint. We first describe the implementation of Algorithm 2 in Section 6.1. In Section 6.2, we make comparison between model (1) with the least-squares loss function f and the fused Lasso model (5) by using PGiPN to solve the former and SSNAL (Li et al. (2018)) to solve the latter, to highlight the advantages and disadvantages of our proposed fused ℓ_0-norms regularization. Among others, the code of SSNAL is available at (https://github.com/MatOpt/SuiteLasso). Finally, in Section 6.3, we present some numerical results toward the comparison among several variants of PGiPN and ZeroFPR and PG method for (1) in terms of efficiency and the quality of the output. The MATLAB code of PGiPN is available at (https://github.com/yuqiawu/PGiPN).

6.1 Implementation of Algorithm 2

6.1.1 Computation of subproblem (11)

Suppose that $\emptyset \neq S_k^c := [n]\backslash S_k$. Based on the fact that every $x \in \Pi_k$ satisfies $x_{S_k^c} = 0$, we can obtain an approximate solution to (11) by solving a problem in a lower dimension. Specifically, for each $k \in \mathcal{K}_2$, write

$$H_k := (G_k)_{S_kS_k},\ v^k := x^k_{S_k},\ \nabla f_{S_k}(v^k) = [\nabla f(x^k)]_{S_k},\ \widehat{\Pi}_k := \{v \in \mathbb{R}^{|S_k|} \mid \widetilde{B}_kv = 0,\ l_{S_k} \le v \le u_{S_k}\},$$

where $\widetilde{B}_k$ is the matrix obtained by removing the rows of $B_{T_k^cS_k}$ whose elements are all zero. We turn to consider the following strongly convex optimization problem,

$$\widehat{v}^k \approx \mathop{\arg\min}_{v\in\mathbb{R}^{|S_k|}} \Big\{\theta_k(v) := f(I_{\cdot S_k}v^k) + \langle \nabla f_{S_k}(v^k), v - v^k\rangle + \frac{1}{2}(v-v^k)^\top H_k(v-v^k) + \delta_{\widehat{\Pi}_k}(v)\Big\}. \quad (64)$$

The following lemma gives a way to find y^k satisfying (13)-(14) by inexactly solving problem (64), whose dimension is much smaller than that of (11) if $|S_k| \ll n$.

Lemma 25 *Let $y^k_{S_k} = \widehat{v}^k$ and $y^k_{S_k^c} = 0$. Then, $\Theta_k(y^k) = \theta_k(\widehat{v}^k)$ and* $\mathrm{dist}(0, \partial\Theta_k(y^k)) = \mathrm{dist}(0, \partial\theta_k(\widehat{v}^k))$. *Consequently, the vector $\widehat{v}^k$ satisfies*

$$\theta_k(\widehat{v}^k) - \theta_k(v^k) \le 0,\ \mathrm{dist}(0, \partial\theta_k(\widehat{v}^k)) \le \frac{\min\{\overline{\mu}_k^{-1}, 1\}}{2} \min\Big\{\|\overline{\mu}_k(x^k - \overline{x}^k)\|, \|\overline{\mu}_k(x^k - \overline{x}^k)\|^{1+\varsigma}\Big\},$$

if and only if the vector y^k satisfies the inexact conditions in (13)-(14).

Proof The first part is straightforward. We consider the second part. By the definition of Θ_k, $\mathrm{dist}(0, \partial\Theta_k(y^k)) = \mathrm{dist}(0, \nabla f(x^k) + G_k(y^k - x^k) + \mathcal{N}_{\Pi_k}(y^k))$. Recall that $\Pi_k = \{x \in \Omega \mid B_{T_k^c\cdot}x = 0, x_{S_k^c} = 0\}$. Then, $\mathcal{N}_{\Pi_k}(y^k) = \mathrm{Range}(B_{T_k^c\cdot}^\top) + \mathrm{Range}(I_{S_k^c\cdot}^\top) + \mathcal{N}_\Omega(y^k)$, and

$$\begin{aligned}\mathrm{dist}(0, \partial\Theta_k(y^k)) &= \mathrm{dist}\big(0, \nabla f(x^k) + G_k(y^k - x^k) + \mathrm{Range}(B_{T_k^c\cdot}^\top) + \mathrm{Range}(I_{S_k^c\cdot}^\top) + \mathcal{N}_\Omega(y^k)\big)\\ &= \mathrm{dist}\big(0, \nabla f_{S_k}(v^k) + H_k(\widehat{v}^k - v^k) + \mathrm{Range}(B_{T_k^c S_k}^\top) + \mathcal{N}_{[l_{S_k}, u_{S_k}]}(\widehat{v}^k)\big)\\ &= \mathrm{dist}(0, \nabla f_{S_k}(v^k) + H_k(\widehat{v}^k - v^k) + \mathcal{N}_{\widehat{\Pi}_k}(\widehat{v}^k)) = \mathrm{dist}(0, \theta_k(\widehat{v}^k)),\end{aligned}$$

where the second equality is using $\mathrm{Range}(I_{S_k^c\cdot}^\top) = \{z \in \mathbb{R}^n \mid z_{S_k} = 0\}$. ■

From the above discussions, we see that the computation of subproblem (11) involves the projection onto Π_k. Next we provide a method for computing it. Fix any $k \in \mathcal{K}_2$. Given $z \in \mathbb{R}^n$, we consider the minimization problem on the projection onto Π_k:

$$\min_{x\in\mathbb{R}^n} \ \frac{1}{2}\|x - z\|^2 \ \text{ s.t. } \widehat{B}_{T_k^c\cdot}x = 0, \ x_{S_k^c} = 0, \ l \le x \le u. \tag{65}$$

We provide a toy example to illustrate how to solve (65). Let $x^k = (1, 1, 2, 3, 3, 0, 0, 0)^\top \in \mathbb{R}^8$. Since $T_k^c = \{1, 4, 6, 7\}$ and $S_k^c = \{6, 7, 8\}$, problem (65) can be written as

$$\min_{x\in\mathbb{R}^8} \frac{1}{2}\|x - z\|^2 \text{ s.t. } x_1 = x_2, \ x_4 = x_5, \ x_6 = x_7 = x_8 = 0, \ l \le x \le u,$$

which can be separated into the following four lower dimensional problems:

$$\begin{aligned}&\min_{x_{1:2}\in\mathbb{R}^2}(1/2)\|x_{1:2} - z_{1:2}\|^2 \text{ s.t. } x_1 = x_2, \ l_{1:2} \le x_{1:3} \le u_{1:2};\\ &\min_{x_3\in\mathbb{R}}(1/2)\|x_3 - z_3\|^2 \text{ s.t. } l_3 \le x_3 \le u_3;\\ &\min_{x_{4:5}\in\mathbb{R}^2}(1/2)\|x_{4:5} - z_{4:5}\|^2 \text{ s.t. } x_4 = x_5, \ l_{4:5} \le x_{4:5} \le u_{4:5};\\ &\min_{x_{6:8}\in\mathbb{R}^3}(1/2)\|x_{6:8} - z_{6:8}\|^2 \text{ s.t. } x_6 = x_7 = x_8 = 0.\end{aligned}$$

Inspired by this toy example, there exists a smallest $\widehat{j} \in \mathbb{N}$ such that the index set T_k^c can be partitioned into $T_k^c = \bigcup_{i\in[\widehat{j}]}[i_1 : i_2]$. Without loss of generality, we assume that these sets are listed in an increasing order according to their left endpoints. Then, problem (65) can be represented as

$$\begin{aligned}&\min_{x\in\mathbb{R}^n} \sum_{i\in[\widehat{j}]} \frac{1}{2}\|x_{i_1:i_2+1} - z_{i_1:i_2+1}\|^2 + \sum_{i\in T_k\setminus(\bigcup_{i\in[\widehat{j}]}\{i_2+1\})} \frac{1}{2}(x_i - z_i)^2\\ &\text{s.t. } x_{S_k^c} = 0; \ l \le x \le u; \ x_{k_1} = x_{k_2} \text{ for } k_1, k_2 \in [i_1 : i_2 + 1], \quad \forall\, i \in [\widehat{j}].\end{aligned} \tag{66}$$

From this equivalent expression, problem (65) can be separated into $\widehat{j} + |T_k\setminus(\bigcup_{i\in[\widehat{j}]}\{i_2+1\})|$ blocks. The following proposition shows that the unique global solution of (65) can be characterized by those of every small block problems.

Proposition 26 *For each $i \in [\widehat{j}]$, if $[i_1 : i_2\!+\!1] \cap S_k^c \neq \emptyset$, let $x^*_{i_1:i_2+1} = 0$; otherwise, let $x^*_{i_1:i_2+1}$ be the unique optimal solution to*

$$\mathop{\arg\min}_{v\in\mathbb{R}^{i_2+2-i_1}} \frac{1}{2}\|v - z_{i_1:i_2+1}\|^2 \ \text{ s.t. } \ l_{i_1:i_2+1} \le v \le u_{i_1:i_2+1},\ v_1 = \cdots = v_{i_2+2-i_1}. \tag{67}$$

For each $i \in T_k\backslash(\bigcup_{i\in[\widehat{j}]}\{i_2+1\})$, if $i \in S_k^c$, let $x_i^* = 0$; otherwise, let x_i^* be the unique optimal solution to

$$\min_{\alpha\in\mathbb{R}} \ \frac{1}{2}(\alpha - z_i)^2 \text{ s.t. } l_i \le \alpha \le u_i. \tag{68}$$

Then, x^* is the unique optimal solution to (65)*.*

An elementary calculation yields the unique solution of (67) as $v^* = \alpha_i^* \mathbf{1}_{i_2+2-i_1}$ with

$$\alpha_i^* = \min\Big\{\max\Big\{\frac{\sum z_{i_1:i_2+1}}{i_2+2-i_1}, \max\{l_{i_1:i_2+1}\}\Big\}, \min\{u_{i_1:i_2+1}\}\Big\},$$

and the unique optimal solution to (68) is $\min\{\max\{z_i, l_i\}\}, u_i\}\}$. Together with Proposition 26, we conclude that the unique optimal solution to (65) is accessible.

6.1.2 Acceleration of Algorithm 2

Generally, when $\|Bx^k\|_0$ or $\|x^k\|_0$ is large, it is difficult for the switch condition in (8) to be satisfied, which will make PGiPN continuously execute PG steps. This phenomenon is evident in the numerical experiment of the restoration of blurred images in Section 6.3.2. To accelerate the iterations of Algorithm 2 or make its iterations enter in Newton steps earlier, we introduce the following relaxed switch condition:

$$\||\text{sign}(Bx^k)| - |\text{sign}(B\overline{x}^k)|\|_1 \le \frac{\eta_1 n}{k} \ \text{ and } \ \||\text{sign}(x^k)| - |\text{sign}(\overline{x}^k)|\|_1 \le \frac{\eta_2 n}{k}, \tag{69}$$

where $\eta_1 \ge 0$ and $\eta_2 \ge 0$ are two given constants. By following the arguments similar to those for Lemma 13, Algorithm 2 equipped with (69) is also well defined. Obviously, when $\frac{\eta_i n}{k} \ge 1$, condition (69) allows the supports of Bx^k and $B\overline{x}^k$ and x^k and $\overline{x}^k$ have some difference; when $\frac{\eta_i n}{k} < 1\,(i=1,2)$, condition (69) is identical to (8). This means that as k grows, Algorithm 2 with relaxed switch condition (69) will finally reduce to the one with (8). Since our convergence analysis does not specify the initial point, the asymptotic convergence results also hold for Algorithm 2 with condition (69).

6.1.3 Choice of parameters in Algorithm 2

We will test the performance of PGiPN with $G_k = G_k^2$ given by (36), and PGiPN(r) that is PGiPN with the relaxed switch condition (69). We apply Gurobi to solve subproblem (11) with such G_k under inexact conditions (13) and (14) controlled by options params.Cutoff and params.OptimalityTol, respectively. Also, we test PGilbfgs that is the same as PGiPN except that the limited-memory BFGS (lbfgs) is used to construct G_k, i.e., to form $G_k = B_k + b_1\|\overline{\mu}_k(x^k - \overline{x}^k)\|^\sigma$ with B_k given by lbfgs. For solving (11) with such G_k, we use the method introduced in Kanzow and Lechner (2022). The parameters of all the variants of PGiPN are chosen as $\alpha = 10^{-8}$, $\sigma = \frac{1}{2}$, $\varrho = 10^{-4}$, $\beta = \frac{1}{2}$, $\varsigma = \frac{2}{3}$, and $b_1 = 10^{-3}$ is used for PGiPN and PGiPN(r), and $b_1 = 10^{-8}$ for PGilbfgs.

We compare the numerical performance of PGiPNs with that of ZeroFPR (Themelis et al. (2018)) and the PG method (Wright et al. (2009)). Among others, ZeroFPR uses the lbfgs to minimize the forward-backward envelope of the objective, and its code package can be downloaded from (http://github.com/kul-forbes/ForBES). We run it with the default setting. In addition, the iteration steps of PG are the same as those of PGiPN without the Newton steps, so that we can check the effect of the additional second-order step on PGiPN. For this reason, the parameters of PG are chosen to be the same as those involved in PG Step of PGiPN. We also observe that the sparsity of the output is very sensitive to μ_k in Algorithm 2. To be fair, as the default setting in ZeroFPR, in all variants of PGiPN and PG, we set $\mu_k = 0.95^{-1}\|A\|_2^2$ for all $k \in \mathbb{N}$ with $\|A\|_2$ computed by the MATLAB sentences: opt.issym = 1; opt.tol = 0.001; ATAmap = @(x) A'*A*x; L = eigs(ATAmap,n,1,'LM',opt)

For each solver, we set $x^0 = 0$ and terminate at the iterate x^k whenever $k \geq 5000$ or $\overline{\mu}_k\|x^k - \mathrm{prox}_{\overline{\mu}_k^{-1}g}(x^k - \overline{\mu}_k^{-1}\nabla f(x^k))\|_\infty < 10^{-4}$. All the numerical tests in this section are conducted on a desktop running on 64-bit Windows System with an Intel(R) Core(TM) i7-10700 CPU 2.90GHz and 32.0 GB RAM.

6.2 Model comparison with the fused Lasso

This subsection is devoted to examining the superiority and shortcoming of model (1) with $f(\cdot) = \frac{1}{2}\|A\cdot - b\|^2$ and $B = \widehat{B}$, i.e. the fused ℓ_0-norms regularization problem with a box constraint (FZNS), compared with the fused Lasso (5). We apply PGiPN to solve FZNS, and SSNAL to solve (5). Considering that the models to solve are different, we only compare the quality of solutions returned by PGiPN and SSNAL, but do not do their running time.

Our first empirical study focuses on the ability of regression via a commonly used dataset, prostate data. There are 97 observations and 9 features included in this dataset. This data was used in Jiang et al. (2021) to check the performance of square root fused Lasso. We randomly select 50 observations to form the training set, and obtain the training data matrix $A \in \mathbb{R}^{50\times 8}$. The corresponding responses are represented by $b \in \mathbb{R}^{50}$. The rest 47 observations are left for testing set, which forms $(\bar{A}, \bar{b})$ with $\bar{A} \in \mathbb{R}^{47\times 8}$ and $\bar{b} \in \mathbb{R}^{47}$. We employ PGiPN to solve FZNS, and SSNAL (Li et al. (2018)) to solve the fused Lasso (5), with (A, b) given above, and $[l, u] = 1000[-\mathbf{1}, \mathbf{1}]$. For each solver, we select 10 groups of $(\lambda_1, \lambda_2) \in [0.003, 400] \times [0.0003, 40]$, ensuring that the outputs exhibit different sparsity levels. We record the sparsity and the testing error, where the latter is defined as $\|\bar{A}x^* - \bar{b}\|$ with x^* being the output. The above procedure is repeated for 100 randomly constructed (A, b) and a random (A, b) is tested with 10 groups of (λ_1, λ_2), resulting in a total of 1000 recorded outputs for each model. All the sparsity pairs $(\|\widehat{B}x^*\|_0, \|x^*\|_0)$ from PGiPN and SSNAL are recorded in lines 1, 4 and 7 of Table 1. For every sparsity pair, the average testing errors of $\|\bar{A}x^* - \bar{b}\|$ for PGiPN and SSNAL corresponding to the given pair is recorded in lines 2, 5 and 8 of Table 1, while the standard deviation of the results is recorded in its lines 3, 6, and 9. Considering that the fused Lasso may produce solutions with components being very small but not equal to 0, we define $\|y\|_0 := \min\{k \mid \sum_{i=1}^k |y|_i^{\downarrow} \geq 0.999\|y\|_1\}$ as in Li et al. (2018) for the outputs of the fused Lasso, where $|y|^{\downarrow}$ is the vector obtained by sorting $|y|$ in a nonincreasing order. As shown in Table 1, when $(\|\widehat{B}x^*\|_0, \|x^*\|_0) = (6, 6)$, the average testing error for FZNS is the smallest among all the testing examples. Among the total 20 experiment results, FZNS outperforms the fused Lasso for 13 cases. For the rest 7 cases, there are 6 cases with $\|\widehat{B}x^*\|_0 \geq 4$. This indicates that our model performs better when $\|\widehat{B}x^*\|_0 \leq 3$.

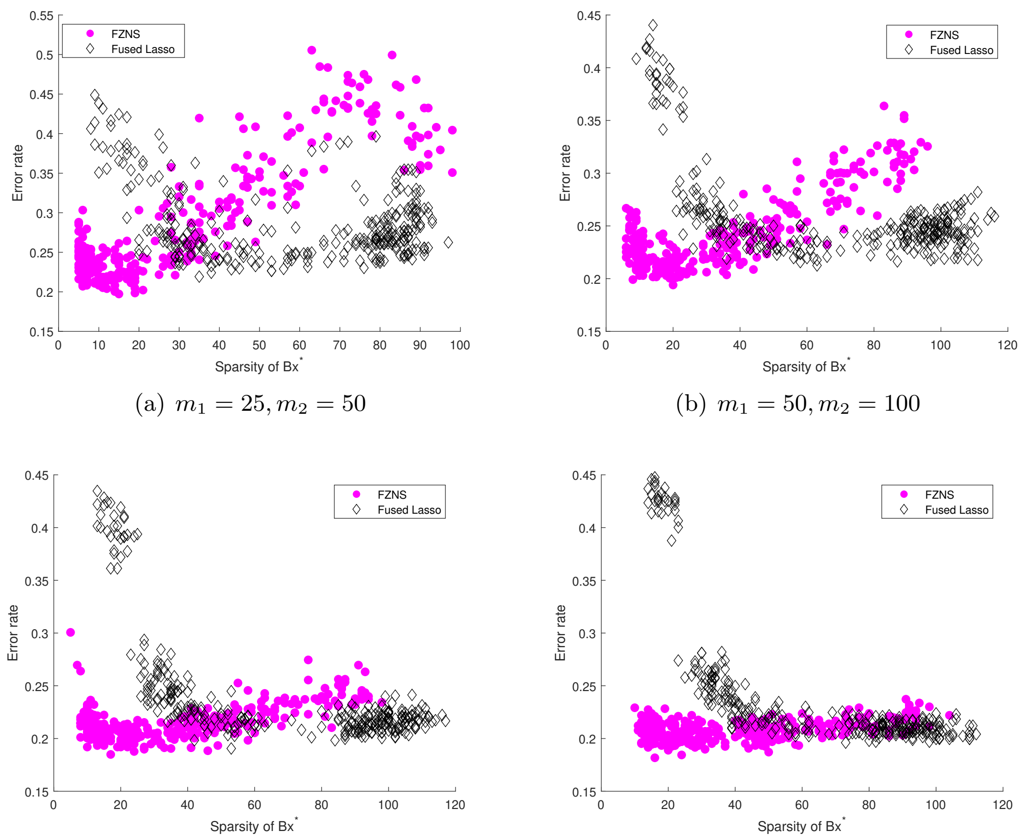
Table 1: Average testing error (FZNS|Fused Lasso) of the outputs.

$(\|\widehat{B}x^*\|_0, \|x^*\|_0)$	(2,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)	(3,8)
Average testing error	8.35|8.54	7.34|7.36	5.44|5.15	5.12|5.74	5.21|6.32	5.08|5.27	5.11|5.70
Standard deviation	0.48|0.37	0.76|0.74	0.27|0.30	1.06|1.10	0.35|0.26	0.30|0.22	0.32|0.24
$(\|\widehat{B}x^*\|_0, \|x^*\|_0)$	(4,5)	(4,6)	(4,7)	(4,8)	(5,5)	(5,6)	(5,7)
Average testing error	5.24|5.86	5.49|4.99	5.25|4.97	5.33|4.78	4.60|5.48	5.60|5.38	5.46|5.58
Standard deviation	1.02|1.15	0.48|0.46	0.28|0.21	0.38|0.29	0.41|1.61	0.61|0.71	0.29|0.40
$(\|\widehat{B}x^*\|_0, \|x^*\|_0)$	(5,8)	(6,6)	(6,7)	(6,8)	(7,7)	(7,8)	
Average testing error	5.35|5.19	4.41|5.26	5.34|4.95	5.20|5.34	5.13|5.22	5.27|5.22	
Standard deviation	0.58|0.43	0.42|1.10	0.69|0.79	0.94|0.73	0.87|1.45	1.52|1.17	

Our second numerical study is to evaluate the classification ability of these two models with the TIMIT database (Acpistoc-Phonetic Continuous Speech Corpus, NTIS, US Dept of Commerce), which consists of 4509 32ms speech frames and each speech frame is represented by 512 samples of 16 KHz rate. The TIMIT database is collected from 437 male speakers. Every speaker provided approximately two speech frames of each of five phonemes, where the phonemes are "sh" as in "she", "dcl" as in "dark", "iy" as the vowel in "she", "aa" as the vowel in "dark", and "ao" as the first vowel in "water". This database is a widely used resource for research in speech recognition. Following the approach described in Land and Friedman (1997), we compute a log-periodogram from each speech frame, which is one of the several widely used methods to generate speech data in a form suitable for speech recognition. Consequently, the dataset comprises 4509 log-periodograms of length 256 (frequency). It was highlighted in Land and Friedman (1997) that distinguishing between "aa" and "ao" is particularly challenging. Our aim is to classify these sounds using FZNS and the fused Lasso with $\lambda_2 = 0$, $l = -\mathbf{1}$ and $u = \mathbf{1}$, or in other words, the zero-order variable fusion (3) plus a box constraint and the first-order variable fusion (4).

In TIMIT, the numbers of phonemes labeled "aa" and "ao" are 695 and 1022, respectively. As in Land and Friedman (1997), we use the first 150 frequencies of the log-periodograms because the remaining 106 frequencies do not appear to contain any information. We randomly select m_1 samples labeled "aa" and m_2 samples labeled "ao" as training set, which together with their labels form $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, with $m = m_1 + m_2$, $n = 150$, where $b_i = 1$ if $A_{i\cdot}$ is labeled as "aa", and $b_i = 2$ otherwise. The rest of dataset is left as the testing set, which forms $\bar{A} \in \mathbb{R}^{(1717-m)\times n}, \bar{b}^{1717-m}$, with $\bar{b}_i = 1$ if $\bar{A}_{i\cdot}$ is labeled as "aa" and $\bar{b}_i = 2$ otherwise. For (A, b), given 10 λ_1's randomly selected within $[2\times 10^{-5}, 300]$ such that the sparsity of the outputs $\|\widehat{B}x^*\|_0$ spans a wide range. If $\bar{A}_{i\cdot}x^* \leq 1.5$, this phoneme is classified as "aa" and hence we set $\hat{b}_i = 1$; otherwise, $\hat{b}_i = 2$. If $\hat{b}_i \neq \bar{b}_i$, $A_{i\cdot}$ is regarded as failure in classification. Then the error rate of classification is given by $\frac{\|\bar{b}-\hat{b}\|_1}{1717-m}$. We record both $\|\widehat{B}x^*\|_0$ and the error rate of classification.

The above procedure is repeated for 30 groups of randomly generated (A, b), resulting in 300 outputs for each solver. The four subfigures in Figure 1 present $\|\widehat{B}x^*\|_0$ and the error rate of each output, with 4 different choices of (m_1, m_2). We see that, for each subfigure the output with the smallest error rate is always achieved by the fused ℓ_0-norms regularization model. It is apparent that FZNS generally performs better than the fused Lasso when $\|\widehat{B}x^*\|_0 \leq 30$, while the average error rate of the fused Lasso is lower than that of FZNS when $\|\widehat{B}x^*\|_0 \geq 60$. This phenomenon is especially evident when m_1 and m_2 are small.



(a) $m_1 = 25, m_2 = 50$

(b) $m_1 = 50, m_2 = 100$

(c) $m_1 = 100, m_2 = 200$

(d) $m_1 = 200, m_2 = 400$

Figure 1: $\|\widehat{B}x^*\|_0$ and the classification error rate for the outputs from FZNS and the fused Lasso under different m_1, m_2.

The numerical results for these two empirical studies show that for prostate database, our model outperforms the fused Lasso when the output is sufficiently sparse, that is, $\|\widehat{B}x^*\|_0 \leq 3$, see the first two lines in Table 1, and for phoneme database, our model performs better when $\|\widehat{B}x^*\|_0 \leq 30$. We also observe that the numerical performance of the fused ℓ_0-norms regularization is not stable if the output is not sparse, especially when the number of observations is small, so when using the fused ℓ_0-norms regularization model, a careful consideration should be given to selecting an appropriate penalty parameter. Moreover, for some optimal solution x^* of the fused Lasso regularization problem, $|\widehat{B}x^*|_{\min}$ and $|x^*|_{\min}$ may be very small but not equal to zero, which leads to a difficulty in interpreting what the outputs mean in the real world application. This also well matches the statements in Land and Friedman (1997) that the ℓ_0-norm variable fusion produces simpler estimated coefficient vectors.

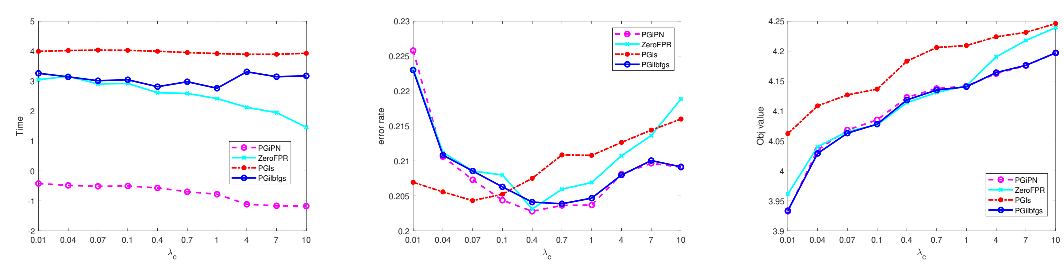
6.3 Comparison with ZeroFPR and PG

This subsection focuses on the numerical comparisons among several variants of PGiPN, ZeroFPR and PG, in terms of the number of iterations, the required CPU time, and the quality of the outputs.

6.3.1 Classification of TIMIT

The experimental data used in this part is the TIMIT dataset, the one in Section 6.2. To test the performance of the algorithms on (1) with nonconvex f, we consider solving model (1) with $f(\cdot) = \sum_{i=1}^{m} \log\big(1 + \frac{(A\cdot - b)_i^2}{\nu}\big)$, $B = \widehat{B}$, $l = -\mathbf{1}$ and $u = \mathbf{1}$, where $A \in \mathbb{R}^{m\times n}$ represents the training data and $b \in \mathbb{R}^m$ is the vector of corresponding labels. It is worth noting that the loss function is nonconvex, and as commented in Aravkin et al. (2012), this loss function is effective to process data denoised by heavy-tailed Student's t-noise.

Following the approach in Section 6.2, we use the first 150 frequencies of the log-periodograms. For the training set, we arbitrarily select 200 samples labeled as "aa" and 400 samples labeled as "ao". These samples, along with their corresponding labels, form the matrices $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, with dimensions $m = 600$ and $n = 150$. The remaining samples are designated as the testing set. Given a group of $\lambda_c > 0$, we set $\lambda_1 = \lambda_c \times 10^{-7}\|A^\top b\|_\infty$ and $\lambda_2 = 0.1\lambda_1$. We employ four solvers, including PGiPN, PGilbfgs, PG and ZeroFPR, to solve model (1) with the above f, and then record the CPU time and the error rate of classification on the testing set. This experimental procedure is repeated for a total of 30 groups of (A, b). Figure 2 plots the average CPU time, error rate and objective value associated with each λ_c, and their standard deviations are reported in Table 2. Motivated by the experiment in Section 6.2, we also plot Figure 3 to show the average $\|\widehat{B}x^*\|_0$ and error rate for all the tested cases produced by four solvers.



(a) λ_c-log(time(seconds)) plot (b) λ_c-error rate plot (c) λ_c-obj value plot

Figure 2: The average CPU time and error rate of 30 examples for four solvers.

We see from Figure 2(a) that in terms of CPU time, PGiPN is always the best one, more than ten times faster than other three solvers. The reason is that other three solvers depend

Table 2: Standard deviation of CPU time, error rate and objective value in Figure 2.

	λ	0.01	0.04	0.07	0.1	0.4	0.7	1	4	7	10
Time	PGiPN	0.12	0.10	0.11	0.10	0.12	0.14	0.11	0.05	0.05	0.04
	ZeroFPR	4.06	5.02	5.12	4.77	3.83	3.41	3.47	2.24	1.91	1.32
	PGls	1.75	2.27	3.03	2.13	3.01	3.51	3.11	3.18	2.75	2.82
	PGilbfgs	6.18	5.77	5.74	5.14	3.73	10.73	9.46	20.92	18.31	16.38
Error rate	PGiPN	0.009	0.008	0.008	0.007	0.008	0.008	0.009	0.009	0.009	0.008
	ZeroFPR	0.010	0.008	0.007	0.007	0.008	0.010	0.010	0.008	0.009	0.010
	PGls	0.006	0.006	0.007	0.008	0.008	0.008	0.008	0.008	0.008	0.009
	PGilbfgs	0.009	0.008	0.007	0.006	0.007	0.008	0.008	0.010	0.008	0.008
Obj	PGiPN	2.67	2.82	3.05	3.17	3.15	2.84	3.13	3.00	3.28	3.38
	ZeroFPR	2.59	2.81	2.86	2.80	3.04	3.00	3.07	3.68	2.88	2.69
	PGls	2.73	2.88	3.04	3.02	2.94	2.97	2.96	3.17	3.25	3.25
	PGilbfgs	2.51	2.74	2.94	2.99	3.00	2.89	3.21	3.00	3.27	3.36

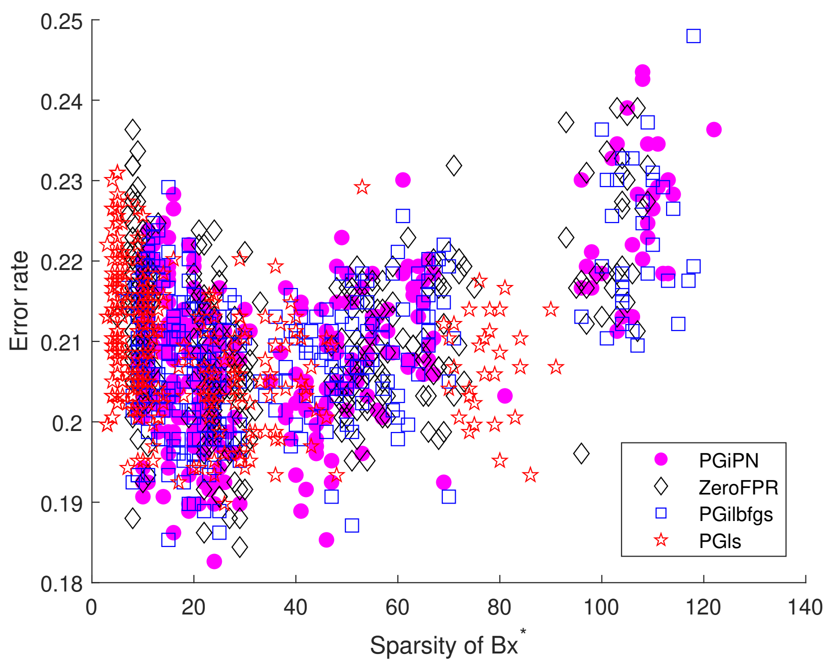


Figure 3: Scatter figure for all tested examples, recording the relationship of sparsity ($\|\widehat{B}x^*\|_0$) and the error rate of classification.

heavily on the proximal mapping of g, and its computation is a little time-consuming. The fact that PGiPN always requires the least CPU time reflects the advantage of the projected regularized Newton steps in PGiPN. From Figure 2(b), when $\lambda_c = 1$, PGiPN attains the smallest average error rate among four solvers for 10 λ_c's. When λ_c is larger, say, $\lambda_c > 0.4$, PGiPN and PGilbfgs tend to outperform ZeroFPR and PG in terms of the average error rate and objective value; when λ_c is smaller, say, $\lambda_c < 0.1$, the solutions returned by PG have the best error rate among four solvers. This is because $\widehat{B}x^*$ returned by PG is sparser than those returned by other three solvers under the same λ_c (see Figure 3), and the solutions given by other three solvers with small λ_c are not sparse, which leads to high error rate.

6.3.2 Recovery of blurred images

Let $\overline{x} \in \mathbb{R}^n$ with $n = 256^2$ be a vector obtained by vectorizing a 256×256 image "cameraman.tif" in MATLAB and then by scaling all the entries to be in $[0, 1]$. Let $A \in \mathbb{R}^{n\times n}$ be a matrix representing a Gaussian blur operator with standard deviation 4 and a filter size of 9, and let $b \in \mathbb{R}^m$ be the vector to represent a blurred image obtained by adding Gauss noise $e \sim \mathcal{N}(0, \epsilon)$ with $\epsilon > 0$ to $A\overline{x}$, i.e., $b = A\overline{x} + e$. We restore the blurred image by using model (1) with $f(\cdot) = \frac{1}{2}\|A \cdot -b\|^2$, $B = \widehat{B}$, $l = 0, u = \mathbf{1}$ and $\lambda_1 = \lambda_2 = 0.0005 \times \|A^\top b\|_\infty$. We test five solvers including PGiPN, PGiPN(r), PGilbfgs, ZeroFPR and PG. For PGiPN(r), the constants η_1 and η_2 in (69) are set to be $\eta_1 = 0.01, \eta_2 = 0.01$. For these five solvers, we compare their performance under different ϵ's in terms of the number of iterations (Iter), cpu time (Time), $F(x^*)$ (Fval), $\|x^*\|_0$ (xNnz), $\|\widehat{B}x^*\|_0$ (BxNnz) and the highest peak signal-to-noise ratio (PSNR), where PSNR $:= 10\log_{10}\left(\frac{n}{\|\overline{x}-x^*\|^2}\right)$. In particular, to check the effect of the Newton step for PGiPN, PGiPN(r) and PGilbfgs, we record the iterations in the form $M(N_f, N_t, N_e)$, where M means the total iterations, N_f means the ordinal number of iterations in which the first Newton step appears, N_t denotes the total number of Newton steps, and N_e denotes the total number of Newton steps in the last 10 iterations of solves. We record the cpu time for these three solvers by $M(N)$, where M is the total time and N represents the time for the Newton steps. PSNR measures the quality of the restored images, and the higher PSNR, the better the quality of restoration. Table 3 reports the numerical results of five solvers, where the number marked in blue means the best one in the same line, whereas the number marked in red means the worst one in the same line.

From Table 3, PGiPN(r) always performs the best in terms of time, which verifies the effectiveness of the acceleration scheme proposed in Section 6.1.2. PGiPN is faster than PGilbfgs, and PGilbfgs is faster than PG, supporting the effective acceleration of the Newton steps. ZeroFPR is the most time-consuming, even worse than PG, a pure first-order method. The reason is that ZeroFPR requires more line-searches, and each line-search involves computing the proximal mapping of g once, which is expensive (2-5 seconds). We observe that PGiPN requires less Newton steps than PGiPN(r). Almost all the Newton steps of PGiPN appear at the end of iterations, while more Newton steps of PGiPN(r) appear along the PG steps. This implies that PGiPN with the relaxed switching condition in (69) lacks the stability.

Despite the superiority of time, the solutions yielded by PGiPN(r) are not good. We also observe that $\|\widehat{B}x^*\|_0$ of PGiPN(r) is a little higher than that of PGiPN, PGilbfgs and PG, because PGiPN(r) runs few PG steps, so that its structured sparsity is not well reduced. Moreover, as the PSNR is closely related to $\|\widehat{B}x^*\|_0$, this leads to the weakest performance of PGiPN(r) in terms of PSNR. Although ZeroFPR always outputs solutions with the smallest objective values, its PSNR is not as good as the objective value. The objective values of the outputs of PGiPN are a litter worse than those of the outputs of PGilbfgs and PG. However, by making a trade-off between the speed and the quality of the outputs, we conclude that PGiPN is a good solver for this test. Finally, we remark that in

Table 3: Numerical comparison of five solvers on recovery of blurred image with $\lambda_1 = \lambda_2 = 0.0005\|A^\top b\|_\infty$.

Noise		PGiPN	PGiPN(r)	PGilbfgs	PG	ZeroFPR
$\epsilon = 0.01$	Iter	119(106,5,4)	66(43,19,9)	444(106,120,7)	796	361
	Time	5.40e2(16.9)	**3.49e2(49.4)**	2.03e3(27.2)	3.46e3	**2.28e4**
	Fval	37.95	**38.06**	37.88	37.88	**37.77**
	xNnz	63805	63637	63858	63858	63717
	BxNnz	5995	6467	5778	5779	5834
	psnr	25.77	**25.47**	25.90	25.90	**25.91**
$\epsilon = 0.02$	Iter	153(144,4,4)	64(50,10,8)	324(144,86,7)	853	286
	Time	6.61e2(8.6)	**3.07e2(28.1)**	1.41e3(22.2)	3.61e3	**1.81e4**
	Fval	46.01	**46.10**	45.98	45.98	**45.83**
	xNnz	63485	63318	63495	63495	63350
	BxNnz	6176	6638	6098	6099	6143
	psnr	25.36	**24.81**	**25.42**	**25.42**	25.33
$\epsilon = 0.03$	Iter	140(135,3,3)	54(42,9,8)	320(135,99,8)	717	332
	Time	5.93e2(6.2)	**2.47e2(19.8)**	1.37e3(22.0)	2.99e3	**1.91e4**
	Fval	60.29	**60.37**	60.25	60.26	**60.02**
	xNnz	62998	62778	63006	63006	62800
	BxNnz	6665	7227	6572	6592	6710
	psnr	24.86	**24.12**	**24.90**	**24.90**	24.76
$\epsilon = 0.04$	Iter	161(153,3,3)	65(41,15,4)	306(155,56,4)	526	230
	Time	6.59e2(9.4)	**3.04e2(41.3)**	1.13e3(11.2)	2.10e3	**1.12e4**
	Fval	77.83	**77.87**	77.81	77.82	**77.44**
	xNnz	62098	61908	62104	62104	61853
	BxNnz	7294	7776	7264	7271	7427
	psnr	24.17	**23.47**	**24.20**	**24.20**	24.00
$\epsilon = 0.05$	Iter	108(101,3,3)	62(46,12,8)	353(101,100,6)	688	168
	Time	4.60e2(6.3)	**2.81e2(28.6)**	1.49e3(31.1)	2.73e3	**5.93e3**
	Fval	99.69	**99.73**	99.65	99.65	**98.93**
	xNnz	61362	61252	61377	61381	60963
	BxNnz	8056	8381	7951	7956	8240
	psnr	23.30	22.85	**23.37**	**23.37**	**22.87**

this experiments, we do not find the case that the Newton steps always performs toward the end of the algorithms for PGiPN, PGiPN(r) and PGilbfgs. That is, some Newton steps are executed along the PG steps.

6.3.3 Numerical validation of Assumption 3

As one reviewer mentioned, due to the highly nonconvexity of model (1), it is not easy to remove Assumption 3 from our global convergence result (see Theorem 19). In this part, we make a numerical study on it. To this end, we introduce a specific choice of ξ_k. Let

$$\xi_k := -\nabla f(x^k) - \mathrm{proj}_{\mathrm{Null}(C_k)}(-\nabla f(x^k)) \ \text{ with } \ C_k = [B_{T_k^c}; I_{S_k^c}] \ \text{ for } \ k \in \mathcal{K}_2.$$

Obviously, for each $k \in \mathcal{K}_2$, $\xi_k \perp \mathrm{Null}(C_k)$, which implies that $\xi_k \in \mathcal{N}_{\mathrm{Null}(C_k)}(x^k) \subset \mathcal{N}_{\Pi_k}(x^k)$. The second inclusion is due to $\Pi_k \subset \mathrm{Null}(C_k)$ and the convexity of Π_k and $\mathrm{Null}(C_k)$.

We are ready to solve the problem in Section 6.3.1 with the termination condition $\overline{\mu}_k \|x^k - \overline{x}^k\|_\infty \le 10^{-8}$. Each test will generate a sequence $\{a_k\}_{k\in\mathcal{K}_2}$ with $a_k := \frac{-\langle \nabla f(x^k)+\xi_k, d^k\rangle}{\|\nabla f(x^k)+\xi_k\|\|d^k\|}$. Since $\{a_k\}_{k\in\mathcal{K}_2}$ is a finite sequence, its infimum limit does not exist. Recall that for a real value infinite sequence $\{b_k\}$, $\liminf_{k\to\infty} b_k = \sup_{l\in\mathbb{N}} \inf_{k\ge l} b_k$. Write the number of elements of $\{a_k\}_{k\in\mathcal{K}_2}$ as t. For each test, we record $\underline{a}$ as follows, as an approximation to the lower limit,

$$\underline{a} := \sup_{l\in[t]} \inf_{k\ge l} a_k.$$

It is not hard to check that $\underline{a} = a_{k'}$, where k' is the maximum element of $\mathcal{K}_2$. We solve the problem for 10 different λ_c's and 10 different groups of (A, b), resulting in 100 $\underline{a}$ for 100 times experiments. We store these 100 $\underline{a}$'s as a MATLAB variable cosinelist, and find that min(cosinelist) = 0.0025, mean(cosinelist) = 0.0761 and std(cosinelist) = 0.0650. This indicates that it is highly possible for Assumption 3 to hold.

7. Conclusions

In this paper, we proposed a hybrid of PG and inexact projected regularized Newton methods for solving the fused ℓ_0-norms regularization problem (1). This hybrid framework fully exploits the advantages of PG method and Newton method, while avoids their disadvantages. We employed the KL property to prove the full convergence of the generated iterate sequence under a curve condition (Assumption 3) on f without assuming the uniformly positive definiteness of the regularized Hessian matrix, and also obtained a superlinear convergence rate under a Hölderian local error bound on the set of the second-order stationary points, without assuming the local optimality of the limit point.

All PGiPN, ZeroFPR and PG have employed the polynomial-time algorithm to compute a point in the proximal mapping of g with $B = \widehat{B}$, which we developed in Section 3.3 of this paper. Numerical tests indicate that our PGiPN not only produces solutions of better quality, but also requires 2-3 times less running time than PG and ZeroFPR, where the latter mainly attributes to our subspace strategy when applying the projected regularized Newton method to solve the problems. It would be an interesting topic to extend the polynomial-time algorithm in Section 3.3 to the case where B is of other special structures.

Acknowledgments

The authors would like to thank the editor and the two anonymous referees for their valuable suggestions, which allowed them to improve the quality of the paper.

The second author's work was supported by the National Natural Science Foundation of China under project No.12371299, and the third author's research was partially supported by Research Grants Council of Hong Kong SAR, P.R. China (PolyU15209921).

References

Masoud Ahookhosh, Andreas Themelis, and Panagiotis Patrinos. A Bregman forward-backward linesearch algorithm for nonconvex composite optimization: superlinear convergence to nonisolated local minima. *SIAM Journal on Optimization*, 31(1):653–685, 2021.

Aleksandr Aravkin, Michael P Friedlander, Felix J Herrmann, and Tristan Van Leeuwen. Robust inversion, dimensionality reduction, and randomized sampling. *Mathematical Programming*, 134:101–125, 2012.

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.

Gilles Bareilles, Franck Iutzeler, and Jérôme Malick. Newton acceleration on manifolds identified by proximal gradient methods. *Mathematical Programming*, 200:37–70, 2023.

Heinz H Bauschke, Jonathan M Borwein, and Wu Li. Strong conical hull intersection property, bounded linear regularity, Jameson's property (g), and error bounds in convex optimization. *Mathematical Programming*, 86:135–160, 1999.

Dimitri P Bertsekas. Projected Newton methods for optimization problems with simple constraints. *SIAM Journal on Control and Optimization*, 20(2):221–246, 1982.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

Wei Bian and Xiaojun Chen. A smoothing proximal gradient algorithm for nonsmooth convex regression with cardinality penalty. *SIAM Journal on Numerical Analysis*, 58(1): 858–883, 2020.

Thomas Blumensath and Mike E Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5):629–654, 2008.

Thomas Blumensath and Mike E Davies. Normalized iterative hard thresholding: Guaranteed stability and performance. *IEEE Journal of selected topics in signal processing*, 4 (2):298–309, 2010.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1): 459–494, 2014.

James V Burke and Jorge J Moré. On the identification of active constraints. *SIAM Journal on Numerical Analysis*, 25(5):1197–1211, 1988.

Harold Davenport and Andrzej Schinzel. A combinatorial problem connected with differential equations. *American Journal of Mathematics*, 87(3):684–694, 1965.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

Felix Friedrich, Angela Kempe, Volkmar Liebscher, and Gerhard Winkler. Complexity penalized m-estimation: Fast computation. *Journal of Computational and Graphical Statistics*, 17(1):201–224, 2008.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2024. URL `https://www.gurobi.com`.

Kyle K Herrity, Anna C Gilbert, and Joel A Tropp. Sparse approximation via iterative thresholding. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 3, pages III–III. IEEE, 2006.

Brad Jackson, Jeffrey D Scargle, David Barnes, Sundararajan Arabhi, Alina Alt, Peter Gioumousis, Elyus Gwin, Paungkaew Sangtrakulcharoen, Linda Tan, and Tun Tao Tsai. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108, 2005.

Sean Jewell and Daniela Witten. Exact spike train inference via ℓ_0 optimization. *The Annals of Applied Statistics*, 12(4):2457–2482, 2018.

Sean W Jewell, Toby Dylan Hocking, Paul Fearnhead, and Daniela M Witten. Fast nonconvex deconvolution of calcium imaging data. *Biostatistics*, 21(4):709–726, 2020.

He Jiang, Shihua Luo, and Yao Dong. Simultaneous feature selection and clustering based on square root optimization. *European Journal of Operational Research*, 289(1):214–231, 2021.

Christian Kanzow and Theresa Lechner. Efficient regularized proximal quasi-Newton methods for large-scale nonconvex composite optimization problems. *arXiv preprint arXiv:2210.07644*, 2022.

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107 (500):1590–1598, 2012.

Stephanie R Land and Jerome H Friedman. Variable fusion: A new adaptive signal regression method. Technical report, Technical Report 656, Department of Statistics, Carnegie Mellon University, 1997.

Jason D Lee, Yuekai Sun, and Michael A Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

Xudong Li, Defeng Sun, and Kim-Chuan Toh. On efficiently solving the subproblems of a level-set method for fused Lasso problems. *SIAM Journal on Optimization*, 28(2): 1842–1866, 2018.

V Liebscher and G Winkler. A potts model for segmentation and jump-detection. In *Proceedings S4G International Conference on Stereology, Spatial Statistics and Stochastic Geometry, Prague June*, volume 21, pages 185–190. Citeseer, 1999.

Jun Liu, Shuiwang Ji, and Jieping Ye. SLEP: Sparse learning with efficient projections. *Arizona State University*, 6(491):7, 2009.

Jun Liu, Lei Yuan, and Jieping Ye. An efficient algorithm for a class of fused Lasso problems. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–332, 2010.

Ruyu Liu, Shaohua Pan, Yuqia Wu, and Xiaoqi Yang. An inexact regularized proximal Newton method for nonconvex and nonsmooth optimization. *Computational Optimization and Applications*, 88:603–641, 2024.

Zhaosong Lu. Iterative hard thresholding methods for ℓ_0 regularized convex cone programming. *Mathematical Programming*, 147(1):125–154, 2014.

Zhaosong Lu and Yong Zhang. Sparse approximation via penalty decomposition methods. *SIAM Journal on Optimization*, 23(4):2448–2478, 2013.

Robert Maidstone, Toby Hocking, Guillem Rigaill, and Paul Fearnhead. On optimal multiple changepoint algorithms for large data. *Statistics and computing*, 27:519–533, 2017.

Cesare Molinari, Jingwei Liang, and Jalal Fadili. Convergence rates of Forward–Douglas–Rachford splitting method. *Journal of Optimization Theory and Applications*, 182:606–639, 2019.

Boris S Mordukhovich, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A globally convergent proximal Newton-type method in nonsmooth convex optimization. *Mathematical Programming*, 198(1):899–936, 2023.

Shaohua Pan, Ling Liang, and Yulan Liu. Local optimality for stationary points of group zero-norm regularized problems and equivalent surrogates. *Optimization*, 72(9):2311–2343, 2023.

René A Poliquin and Ralph Tyrell Rockafellar. A calculus of prox-regularity. *J. Convex Anal*, 17(1):203–210, 2010.

Guillem Rigaill. A pruned dynamic programming algorithm to recover the best segmentations with 1 to $k\_\{max\}$ change-points. *Journal de la Société Française de Statistique*, 156(4):180–205, 2015.

Stephen M Robinson. *Some continuity properties of polyhedral multifunctions.* Springer, 1981.

R Tyrrell Rockafellar. *Convex analysis.* Princeton university press, 1970.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer, 2009.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

Micha Sharir. Davenport-schinzel sequences and their geometric applications. In *Theoretical Foundations of Computer Graphics and CAD*, pages 253–278. Springer, 1995.

Suvrit Sra. Scalable nonconvex inexact proximal splitting. *Advances in Neural Information Processing Systems*, 25, 2012.

Lorenzo Stella, Andreas Themelis, and Panagiotis Patrinos. Forward–backward quasi-Newton methods for nonsmooth optimization problems. *Computational Optimization and Applications*, 67(3):443–487, 2017.

Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM Journal on Optimization*, 28(3):2274–2303, 2018.

Andreas Themelis, Masoud Ahookhosh, and Panagiotis Patrinos. On the acceleration of forward-backward splitting via an inexact Newton method. In *Splitting Algorithms, Modern Operator Theory, and Applications*, pages 363–412. Springer, 2019.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Kenji Ueda and Nobuo Yamashita. Convergence properties of the regularized Newton method for the unconstrained nonconvex optimization. *Applied Mathematics and Optimization*, 62(1):27–46, 2010.

Lou Van den Dries and Chris Miller. Geometric categories and o-minimal structures. *Duke Mathematical Journal*, 84(2), 1996.

Andreas Weinmann, Martin Storath, and Laurent Demaret. The lˆ1-potts functional for robust jump-sparse reconstruction. *SIAM Journal on Numerical Analysis*, 53(1):644–673, 2015.

Stephen J Wright, Robert D Nowak, and Mário AT Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

Fan Wu and Wei Bian. Accelerated iterative hard thresholding algorithm for l_0 regularized regression problem. *Journal of Global Optimization*, 76(4):819–840, 2020.

Yuqia Wu, Shaohua Pan, and Xiaoqi Yang. A regularized Newton method for ℓ_q-norm composite optimization problems. *SIAM Journal on Optimization*, 33(3):1676–1706, 2023.

Man-Chung Yue, Zirui Zhou, and Anthony Man-Cho So. A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo-Tseng error bound property. *Mathematical Programming*, 174(1):327–358, 2019.

Rui Zhou and Daniel P Palomar. Solving high-order portfolios via successive convex approximation algorithms. *IEEE Transactions on Signal Processing*, 69:892–904, 2021.

Shenglong Zhou, Lili Pan, and Naihua Xiu. Newton method for ℓ_0-regularized optimization. *Numerical Algorithms*, 88(4):1541–1570, 2021.

Zirui Zhou and Anthony Man-Cho So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165:689–728, 2017.