Globally Convergent Regularized Newton Methods for Nonconvex Sparse Optimization Problems

Yuqia Wu

PhD

The Hong Kong Polytechnic University

2024

The Hong Kong Polytechnic University

Department of Applied Mathematics

Globally Convergent Regularized Newton Methods for Nonconvex Sparse Optimization Problems

Yuqia Wu

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

March 2024

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

__________________(Signed)

Yuqia Wu (Name of student)

Abstract

Nowadays, many algorithms have been proposed to solve the nonconvex and nonsmooth problems that arise in sparse optimization. Most of these algorithms belong to the first-order type, including the proximal gradient method. First-order methods have several advantages, such as low computational cost in each iteration, weak global convergence conditions, and easy implementation. However, their convergence rate is at most linear, resulting in slow convergence speed when processing large-scale problems. On the other hand, the classical Newton method, which is a second-order method, can achieve a locally superlinear convergence rate. However, the classical Newton method equips with an Armijo line search for minimizing smooth optimization problems can only achieve a subsequence convergence, let alone for nonsmooth sparse optimization. By exploiting the structure of two classes of nonconvex and nonsmooth sparse optimization problems that arise in compressed sensing and machine learning, this thesis presents an efficient hybrid framework that combines a proximal gradient method and a Newton-type method, which takes advantages of these two kinds of optimization algorithms, and simultaneously avoids their disadvantages.

The first part of the thesis designs a hybrid of proximal gradient method and regularized subspace Newton method (HpgSRN) for solving $\ell_q(0<q<1)$-norm regularized minimization problems with a twice continuously differentiable loss function. In the iterates of HpgSRN, we first use the proximal gradient method to find a neighbourhood of a potential stationary point, and then apply a regularized Newton

method in the subspace, at which the objective is locally smooth, to enhance the convergence speed. We show that this hybrid algorithm finally reduces to a regularized Newton method of minimizing a locally smooth function. If the reduced objective function satisfies the Kurdyka-Łojasiewic property and a curve ratio condition holds, the generated sequence converges to an L-stationary point with an arbitrarily picked initial point. Moreover, if we additionally assume that the generated sequence converges to a second-order stationary point, and an error bound condition holds there, we prove a superlinear convergence of the generated sequence, without assuming either the isolatedness or the local minimality of the limit point. Numerical comparison with the proximal gradient method and ZeroFPR, where the later one is an algorithm using limited-memory BFGS method to minimize the forward-backward envelope of the objective function, indicates that our proposed HpgSRN not only converges much faster, but also yields comparable and even better solutions.

The second part of the thesis studies fused zero-norms regularization problems, which are the zero-norm version of the fused Lasso plus a box constraint. We propose a polynomial time algorithm to find an element of the proximal mapping of the fused zero-norms over a box constraint. Based on this, we propose a hybrid of proximal gradient method and inexact projected regularized Newton method for solving the fused zero-norms regularization problems. We prove that the algorithm finally reduces to an inexact projected regularized Newton method for seeking a critical point of a smooth function over a convex constraint. We achieve the convergence of the whole sequence under a nondegeneracy condition, a curve ratio condition and assuming that the reduced objective is a Kurdyka-Łojasiewic function. 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 either the isolatedness or the local optimality of the limit point. Finally, numerical experiments show the features of our considered model, and the superiority of our proposed algorithm.

Acknowledgements

The endeavor of carrying out research is a fascinatingly non-isolated activity. I am grateful to the several individuals who have supported me in various ways during the PhD program and would like to hereby acknowledge their assistance.

First and foremost, I would like to express my deep gratitude to my supervisor, Professor Xiaoqi Yang, for his consistent support and invaluable guidance during my research journey. Professor Yang consistently presented me with challenging problems, and his guidance, encouragement, coupled with the ample time he provided, allowed me to experience the true joy of conducting research. Additionally, I have acquired essential skills in research paper organization and academic writing under his mentorship. This research would not be possible without his selfless support.

I would also like to express my appreciation to Professor Shaohua Pan from South China University of Technology, who was my master supervisor. Professor Pan played an active role in my research, engaging in many invaluable discussions in our regular meetings. Her passion for research has been a consistent source of inspiration for me. I am truly thankful for her involvement and the valuable insights she provided throughout my PhD program.

I thank Professor Guoyin Li from University of New South Wales and Professor Xianfu Wang from University of British Columbia for considering me as a visiting student, despite the fact that I am unable to work with them due to some unforeseen circumstances.

I would like to thank Professor Yaohua Hu for his guidance in my job search, as well as encouraging me to participate in the graduate forum organized by the Operations Research Society of China. His selfless sharing and support have been invaluable to me. I am also thankful to Professor Minghua Li for the financial support for my visit to Chongqing. Furthermore, I am very grateful to my other academic brothers and sisters: Xinlin Hu, Ruyu Liu, Kaiwen Meng, Ying Sun, Pengcheng Wu, Wenfang Yao, Carisa Yu and Chenchen Zu, and my colleagues: Qian Li, Chiyu Ma, Xinyang Yu, Yangzi Zheng, for their support, both in academic and personal matters.

I thanks my girlfriend, Miss Wang Yue, for her care, help, companionship and love, as well as many useful suggestions she provided for my career developments. Last but not the least, I would like to express my heartfelt gratitude to my mom, dad, sister, brother for their love and encouragement. This thesis is dedicated to them.

Contents

List of Figures

List of Tables

List of Notations

$\mathbb{R}$	set of real numbers
$\mathbb{R}^n$	set of n-dimensional real vectors
$\mathbb{R}^{m\times n}$	set of $m \times n$ real matrices
$\mathbb{R}_+$	the set of nonnegative real numbers
$\mathbb{R}_{++}$	the set of positive real numbers
$\langle \cdot, \cdot \rangle$	inner product in Euclidean space
a_+	positive truncation of real number a
$\lceil a \rceil$	ceiling function of real number a
$o(\alpha)$	$o(\alpha)/\alpha \to 0$ as $\alpha \to 0$
$O(\alpha)$	there exists $t > 0$ such that $O(\alpha) \leq t\alpha$
$\|x\|$	Euclidean norm of x
$|x|_{\min}$	smallest nonzero absolute value of x
$\mathrm{supp}(x)$	the set of nonzero indices of x
$\mathrm{sign}(x)$	the signum function of x
$\mathbf{1}$	the ones vector, whose dimension is adaptive
I	the identity matrix, whose dimension is adaptive
$\lambda_{\min}(A)$	smallest eigenvalue of symmetric matrix A
$\|A\|_2$	spectral norm of matrix $A \in \mathbb{R}^{m\times n}$

$A_{S\cdot}$(resp. $A_{\cdot T}$)	the matrix consisting of the rows (resp. columns) of A whose indices correspond to S (resp. T)
$\mathbb{B}(x, \delta)$	the ball centered at x with radius δ
B	the unit ball centered at 0
$\text{dist}(x, X)$	smallest distance between the vector x and the set X
$\delta_X(x)$	indicator function of X at x
$\text{proj}_X(x)$	projection of x onto X
$\mathcal{N}_X(x)$	normal cone to the convex set X at x
$\mathcal{T}_X(x)$	tangent cone to the convex set X at x
Range(A)	the range of A, i.e., $\{Ax \mid x \in \mathbb{R}^n\}$
Null(A)	the null space of A, i.e., $\{x \mid Ax = 0\}$
S^c	complement of the set S
$\mathcal{L}_f(x)$	the level set of f at x^0, i.e., $\{z \mid f(z) \leq f(x)\}$

Chapter 1

Literature Review and Introduction

Over the past two decades, there has been a growing interest in sparse optimization, which is concerned with identifying sparse solutions for loss functions. Sparse optimization has found applications in various fields, including compressed sensing, machine learning, signal processing and so on. In the era of big data, the scale of data and problems is gradually increasing. As a result, researchers are paying more attention to addressing large-scale optimization problems in the context of sparse optimization.

Currently, numerous optimization algorithms have been developed to solve problems in sparse optimization, with many of them falling into the category of first-order methods. First-order methods tend to have low computational requirements per iteration and exhibit good global convergence properties. However, their local convergence rate is typically at most linear, resulting in slow convergence speed when dealing with large-scale problems. In contrast, the classical Newton method can achieve a locally superlinear or even quadratic convergence rate under certain regularity conditions. However, starting from an arbitrary initial point, the classical Newton method equipped with Armijo line search for minimizing smooth optimization problems only achieves subsequence convergence.

To effectively tackle the challenges posed by large-scale problems in sparse optimization, this thesis explores a hybrid framework that combines first-order and Newton-type methods. By leveraging the strengths of both approaches, the hybrid algorithm aims to achieve global performance while maintaining a fast local convergence rate. The primary focus of this thesis lies in designing and analyzing hybrid algorithms for two specific classes of nonconvex sparse optimization problems. The first problem involves $\ell_q(0 < q < 1)$-norm regularized problems, while the second problem pertains to fused ℓ_0-norms regularized optimization, which is a ℓ_0-norm variation of the renowned fused Lasso (Tibshirani et al. (2005)), incorporating a box constraint.

The rest of this chapter will provide a literature review on the topics under study. In Chapter 2, we will cover some necessary preliminaries. The main contents of this thesis are in Chapters 3 and 4, where we will present algorithms, convergence analysis and numerical experiments for the ℓ_q-norm regularized problem and the fused ℓ_0-norms problem, respectively. These two chapters are based on the following published work and preprint, respectively:

- Y. Wu, S. Pan and X. Yang. A Regularized Newton Method for ℓ_q-Norm Composite Optimization Problems. SIAM Journal on Optimization, 33(3):1676–1706, 2023. (Wu et al. (2023b))

- Y. Wu, S. Pan and X. Yang. An Inexact Projected Regularized Newton Method for Fused Zero-norms Regularization Problems. arXiv:2312.15718, 2023. (Wu et al. (2023a))

1.1 $\ell_q(0<q<1)$-Regularization Problems

The formulation of $\ell_q(0<q<1)$-regularization problem is

$$\min_{x\in\mathbb{R}^n} \; f(x)+\lambda\|x\|_q^q, \tag{1.1}$$

where $f: \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function, $\lambda > 0$ is the regularization parameter and $\|x\|_q := (\sum_{i=1}^n |x_i|^q)^{1/q}$ denotes the ℓ_q quasi-norm of x. Here $\|\cdot\|_q$ is not a norm because it does not satisfies the sub-additivity property. When $f(\cdot) = \|A\cdot -b\|^2$ for some matrix $A \in \mathbb{R}^{m\times n}$ and vector $b \in \mathbb{R}^m$, problem (1.1) reduces to the familiar ℓ_q regularized least squares problem studied in the literature (see e.g., Chen et al. (2010)).

Problem (1.1) first appears in statistics as the bridge penalty regression (Frank and Friedman (1993)), and later appears in optimization as a special case of nonsmooth and nonconvex penalty problems studied by (Luo et al. (1996); Huang and Yang (2003); Yang and Huang (2001)) for nonlinear optimization problems. In signal processing, Chartrand (2007) early showed that the $\ell_q\,(0 < q < 1)$ quasi-norm can substantially reduce the number of measurements required by ℓ_1-norm for exact recovery of signals, and Xu et al. (2012) showed that the $\ell_{1/2}$ regularization admits a significantly stronger sparsity promoting capability than the ℓ_1 one in the sense that it allows to obtain a more sparse solution, and predicts a sparse signal from less samplings. These apsects motivate the frequent use of the $\ell_q\,(0<q<1)$ quasi-norm in compressed sensing. Because for any given $x \in \mathbb{R}^n$, $\|x\|_q^q \to \|x\|_0$ as $q \downarrow 0$, where $\|x\|_0$ denotes the zero-norm (cardinality) of x, problem (1.1) is often used as a nonconvex surrogate of the zero-norm regularized problem, and is found to have a wide spectrum of applications in signal and image processing, statistics, and machine learning (see, e.g., Figueiredo et al. (2007); Saab et al. (2008); Nikolova et al. (2008); Wang and Yin (2010); Chen et al. (2012); Bian and Chen (2012a); Xu et al. (2012);

Cao et al. (2013)).

Due to the nonconvexity and non-Lipschitz continuity of the ℓ_q quasi-norm, problem (1.1) is a class of difficult nonconvex and nonsmooth optimization problems. In fact, Ge et al. (2011) showed that finding the global minimum value of the problem (1.1) is strongly NP-hard, while finding one of its local minimum in polynomial time is possible. In the past decade, many first-order methods have been developed for seeking its critical points. For some special q, say $q = 1/2$ or $2/3$, since the proximal mapping of the ℓ_q quasi-norm has a closed-form solution (see Xu et al. (2012); Cao et al. (2013)), the proximal gradient (PG) method becomes a class of popular ones for solving (1.1) with such q. For a general $q \in (0, 1)$, Hu et al. (2017, 2021) also proposed an exact PG method and an inexact PG method for problem (1.1), respectively. When assuming that the limit point is a local minimizer, a linear convergence rate was obtained in (Hu et al. (2017, 2021); Xu et al. (2012)). In addition, a class of PG methods with a nonmonotone line search strategy (called SpaRSA) was proposed (see Wright et al. (2009)). For problem (1.1) with a general $q \in (0, 1)$, the reweighted ℓ_1-minimization method is another class of common first-order methods by solving a sequence of weighted ℓ_1-norm regularized minimization problems (see Candes et al. (2008); Lai and Wang (2011); Lai et al. (2013); Lu (2014b); Chen and Zhou (2014); Wang et al. (2021a, 2023)). The reweighted ℓ_1 minimization combined with extrapolation technique for ℓ_q-regularization problems was also considered in Wang et al. (2022). To overcome the non-Lipschitz difficulty of the ℓ_q quasi-norm, Chen et al. (2010), Chen (2012) and Chen et al. (2013) proposed a class of smoothing method by constructing a smooth approximation of the ℓ_q quasi-norm, and using the steepest descent method, the sequential quadratic programming and trust region Newton method to solve the constructed smooth approximation problems, respectively. The second one is also known as smoothing sequential quadratic programming (Bian and Chen (2012b)). Liu et al. (2019) considered a class of the iterative support shrinking algorithms, which are able to address (1.1) with f being the least square loss function, and obtained the convergence of the whole sequence by virtue of KL property.

1.2 Structured ℓ_0-norms Regularized Problems

Given a matrix $B \in \mathbb{R}^{p\times n}$, $\lambda_1 > 0$, $\lambda_2 > 0$, $l \in \mathbb{R}^n_-$ and $u \in \mathbb{R}^n_+$, the formulation of the structured ℓ_0-norms regularization problem is:

$$\min_{x\in\mathbb{R}^n} \; f(x) + \lambda_1\|Bx\|_0 + \lambda_2\|x\|_0 \;\; \text{s.t.} \;\; l \le x \le u, \tag{1.2}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function, $\|\cdot\|_0$ denotes the ℓ_0-norm (or cardinality) function. Despite the fact that $\|\cdot\|_0$ is not a formal norm as it does not satisfy the absolute homogeneity property, we call it ℓ_0-norm for simplicity. This model encourages sparsity of both variable x and its linear transformation Bx.

It is known that one of the formulations for finding a sparse vector while minimizing f is the following ℓ_0 regularization problem

$$\min_{x\in\mathbb{R}^n} \; f(x) + \lambda_2\|x\|_0, \tag{1.3}$$

where the ℓ_0-norm term shrinks some small coefficients to 0 and identifies a set of influential components. In recent years, many first-order optimization algorithms have been well developed to solve the ℓ_0-norm regularization problems of the form (1.3), which includes the iterative hard thresholding (Herrity et al. (2006); Blumensath and Davies (2008, 2010); Lu (2014a)), the penalty decomposition (Lu and Zhang (2013)), the extrapolation proximal iterative hard-thresholding method (Bao et al. (2016)), mixed integer optimization method (Bertsimas et al. (2016)), the coordinate-wise support optimality method (Beck and Hallak (2018)), the active set Barzilar-Borwein algorithm (Cheng et al. (2020)), the smoothing proximal gradient method (Bian and Chen (2020)), the accelerated iterative hard thresholding (Wu and Bian (2020)). There are also several second-order methods proposed to address problem (1.3) or its special case, such as the PDAS (Ito and Kunisch (2013)), the PDASC (Jiao et al. (2015)), the SDAR (Huang et al. (2018)) and the NL0R (Zhou et al. (2021)). Among others, NL0R employs Newton method to solve a series of stationary equations confined within the subspaces identified by the support of the solution obtained by the proximal mapping of $\lambda_2\|\cdot\|_0$.

However, the ℓ_0-norm penalty only takes the sparsity of x into consideration, but ignores its linear transformation, 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. Moreover, Land and Friedman (1997) studied the phoneme classification on 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. This 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”. Since each phoneme is composed of a series of consecutively sampled points, there is a high chance that each sampled point is close or identical to its neighboring ones. For this reason, Land and Friedman (1997) considered imposing a fused penalty on the coefficients vector x, and proposed the following problems with zero-order variable fusion and first-order variable fusion respectively to train the classifier:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|^2 + \lambda_1\|\widehat{B}x\|_0, \tag{1.4}$$

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|^2 + \lambda_1\|\widehat{B}x\|_1, \tag{1.5}$$

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. If $f(\cdot) = \frac{1}{2}\|A\cdot -b\|^2$ and $B = \widehat{B}$, then we call (1.2) 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

$$\min_{x\in\mathbb{R}^n} \ \frac{1}{2}\|Ax-b\|^2 + \lambda_1\|\widehat{B}x\|_1 + \lambda_2\|x\|_1, \tag{1.6}$$

and presented its nice statistical properties. Friedman et al. (2007) demonstrated that the proximal mapping of $\lambda_1\|\widehat{B}x\|_1 + \lambda_2\|x\|_1$ can be obtained through a process, which is known as "prox-decomposition" later. Based on the accessibility of this proximal mapping, various algorithms can efficiently address model (1.6), 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 (1.6). The numerical results presented in their study indicate that SSNAL is highly efficient.

It was claimed in Land and Friedman (1997) that both (1.4) and (1.5) perform well in signal regression, but the zero-order fusion one produces simpler estimated coefficient vectors. This observation suggests that model (1.2) with $f = \frac{1}{2}\|A\cdot -b\|^2$ and $B = \widehat{B}$ may be able to effectively find a simpler solution while performs well as the fused Lasso does. Compared with regularization problems using ℓ_0-norm, those using $\|Bx\|_0$ regularization remain less explored in terms of algorithm development.

According to Land and Friedman (1997), the global solution of (1.4) cannot be solved exactly. However, one of its approximate critical points can be obtained by numerical method. In fact, Jewell and Witten (2018) and Jewell et al. (2020) have 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 PG method to find a critical point of problem (1.4). However, the highly nonconvex and nonsmooth nature of model (1.2) presents significant challenges in computing the proximal mapping of g when $B = \widehat{B}$ and in developing effective optimization algorithms for solving it. As far as we know, no specific algorithms have yet been designed to solve these challenging problems.

1.3 Newton-type Methods for Composite Optimization Problems

In recent years, many researchers are interested in using second-order methods to solve the following general nonconvex and nonsmooth composite problem

$$\min_{x\in\mathbb{R}^n} \Psi(x) := \psi(x) + \phi(x), \tag{1.7}$$

where $\phi\colon \mathbb{R}^n \to (-\infty, \infty]$ is a closed proper function and ψ is a twice continuously differentiable function on an open subset containing the effective domain of ϕ. The proximal Newton-type method is able to address (1.7) with convex ϕ and convex or weakly convex ψ. In particular, the proximal Newton-type method solves in each iteration the following subproblem

$$\arg\min_{x\in\mathbb{R}^n} \Big\{\psi(x^k) + \langle \nabla\psi(x^k), x - x^k\rangle + \frac{1}{2}\langle x - x^k, G_k(x - x^k)\rangle + \phi(x)\Big\}, \tag{1.8}$$

where G_k is an approximation to $\nabla^2\psi(x^k)$, to calculate a Newton direction, and then a backtracking line search determines the step-size. We remark here that (1.8) can be addressed by using (accelerated) proximal gradient method (Beck and Teboulle (2009)). Bertsekas (1982) considered (1.7) with ψ being convex and ϕ being an indicator function of $\mathbb{R}^n_+$. They proved a local superlinear convergence provided that $G_k = \nabla^2\psi(x^k)$ and $\nabla^2\psi(x^k)$ is uniformly positive definite. For both ψ and ϕ being convex, Lee et al. (2014) proposed an inexact proximal Newton-type method and achieved the local quadratic convergence rate of the iterate sequence under the strong convexity of ψ; Yue et al. (2019) proposed an inexact regularized proximal Newton method and established the local linear, superlinear and quadratic convergence rate of the iterate sequence (by the approximation degree to the Hessian matrix of ψ) under Luo-Tseng error bound; Mordukhovich et al. (2023) proposed a proximal Newton-type method and obtained the superlinear convergence rate of the iterate sequence under the metric p ($> 1/2$)-subregularity of the subdifferential mapping $\partial\Psi$. Liu et al. (2024) proposed an inexact regularized proximal Newton-type method for (1.7) with ψ being weakly convex and ϕ being convex. They achieved the superlinear convergence of the iterate sequence under the metric p ($> 1/2$)-subregularity of a KKT residual function. The inexact proximal Newton-type method in Lee et al. (2014) was also extended by Kanzow and Lechner (2021) to solve problem (1.7) with only a convex ϕ, which essentially belongs to weakly convex optimization. Their global and local superlinear convergence results require the local strong convexity of Ψ around any stationary point.

By following a different line, the forward-backward envelope (FBE), which is proposed in (Patrinos and Bemporad (2013)), has been extensively investigated for designing second-order methods. For ϕ being convex with a cheap computable proximal mapping, Stella et al. (2017) combined a PG method and a quasi-Newton method to minimize the FBE of Ψ and proved the convergence of the whole sequence under the KL property of Ψ and the superlinear convergence rate under the local strong convexity of the FBE of Ψ. For (1.7) with ψ being additionally convex and ϕ just having a cheap computable proximal mapping, Themelis et al. (2019) proposed a hybrid of PG and inexact Newton methods by using FBE of Ψ (named FBTN) and proved that $\mathrm{dist}(x^k, \mathcal{X}^*)$ converges superlinearly to 0 under an assumption without requiring the singleton of the solution set $\mathcal{X}^*$ of (1.7). Themelis et al. (2018) used the FBE of Ψ to develop a hybrid framework of PG and quasi-Newton methods (ZeroFPR), and achieved the global convergence of the iterate sequence by virtue of the KL property of the FBE, and its local superlinear rate under the Dennis-Moré condition and the strong local minimum of the limit point. The convergence rate results in Stella et al. (2017) and Themelis et al. (2018) require the isolatedness of the limit point. Recently, Ahookhosh et al. (2021) utilized the Bregman FBE of Ψ to develop a more general hybrid framework of PG and second-order methods, BELLA. They obtained the global convergence of the iterate sequence for the tame functions ψ and ϕ, and the local superlinear rate of the distance of the iterate sequence to the set of fixed points of the Bregman FBE by assuming that the second-order directions are the superlinear ones with order 1 and KL property of exponent $\theta \in (0, 1)$ of Ψ. Their work greatly improved the results of Stella et al. (2017); Themelis et al. (2018) by removing the isolatedness restriction on local minima and established that the second-order directions are indeed the superlinear ones with order 1 under the assumptions that the limit point is a strong local minimum (also implying the isolatedness) and a Dennis-Moré condition holds. It is unclear what conditions are sufficient for second-order directions to be superlinear without the strong local minimum property.

In addition, for the case $\phi(x) = \lambda\|x\|_0$, Zhou et al. (2021) developed a subspace Newton method by solving the stationary equations restricted in the subspace identified by the proximal mapping of $\lambda\|x\|_0$, and established the local quadratic convergence rate of the iterate sequence under the local strong convexity of ψ around any stationary point. Their subspace Newton method relies on the subspaces identified by a PG method. Recently, Bareilles et al. (2023) considered problem (1.7) where ψ is smooth and ϕ has a cheap computable proximal mapping, and proposed ManAcc-Newton, a hybrid of PG and Newton methods under the framework of manifolds. Their algorithm alternates between a PG step and a Riemannian update on an identified manifold, and was proved to have a quadratic convergence rate under a positive definiteness assumption on the Riemannian Hessian of the objective function at limit points. For the unified analysis on manifold identification of any PG methods, we refer the reader to the work (Sun et al. (2019)).

Chapter 2

Preliminaries

In this chapter, we introduce the notations and some preliminary concepts that will be used in this thesis.

2.1 Notations

Throughout this thesis, $\mathbb{R}^n$ denotes the n-dimensional Euclidean space, equipped with the standard inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\|\cdot\|$. For any $x \in \mathbb{R}^n$ and $\epsilon > 0$, $\mathbb{B}(x, \epsilon) := \{z \in \mathbb{R}^n \mid \|z - x\| \leq \epsilon\}$ denotes the ball centered at x with radius ϵ. Let $\mathbf{B} := \mathbb{B}(0, 1)$. For a closed and convex set $\Xi \subseteq \mathbb{R}^n$, we denote by $\mathcal{N}_\Xi(x)$ and $\mathcal{T}_\Xi(x)$ the normal cone and tangent cone of Ξ at x, respectively. For a closed set $\Xi' \subseteq \mathbb{R}^n$, $\text{dist}(x, \Xi') := \min_{z \in \Xi'} \|z - x\|$, and $\text{proj}_{\Xi'}(x) := \{z \in \Xi' \mid \|z - x\| = \text{dist}(x, \Xi')\}$.

For $t \in \mathbb{R}$, $t_+ := \max\{t, 0\}$. Fix any two nonnegative integers $j < k$, define $[j\!:\!k] := \{j, j\!+\!1, ..., k\}$ and $[k] := [1\!:\!k]$. For an index set $T \subseteq [n]$, write $T^c := [n] \backslash T$ and $|T|$ is the number of the elements of T. Given any $x \in \mathbb{R}^n$, $\text{supp}(x) := \{i \in [n] \mid x_i \neq 0\}$, $\text{sign}(x)$ denotes the vector with $[\text{sign}(x)]_i = \text{sign}(x_i)$, where $\text{sign}(t) = \frac{t}{|t|}$ if $t \neq 0$ and $\text{sign}(0) = 0$. We define $|x|_{\min} := \min_{i \in \text{supp}(x)} |x_i|$ and $x_T \in \mathbb{R}^{|T|}$ is the vector consisting of those x_j's with $j \in T$, and $x_{j:k} := x_{[j:k]}$. $\mathbf{1}$ and I are the vector of ones and the identity matrix, respectively, whose dimensions are adaptive to the context. 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 $S \subseteq [m]$, $A_{S\cdot}$ (resp. $A_{\cdot T}$) denotes the matrix consisting of the rows (resp. the columns) of A whose indices correspond to S (resp. T). We write the range of A by $\text{Range}(A) = \{Ax \mid x \in \mathbb{R}^n\}$ and the null space of A by $\text{Null}(A) = \{y \in \mathbb{R}^n \mid Ay = 0\}$. For another matrix $C \in \mathbb{R}^{m\times p}$, $[A\ C] \in \mathbb{R}^{m\times(n+p)}$ is defined as a matrix composed of two matrices, A and C, placed side by side. For any $D \in \mathbb{R}^{p\times n}$, $[A; D] := [A^\top\ D^\top]^\top$. For function f, we denote $\mathcal{L}_f(x) := \{z \mid f(z) \le f(x)\}$ as the level set of f. Moreover, we denote by $\omega(x^0)$ the set of cluster points of the sequence generated by algorithm with starting point x^0.

2.2 Stationary Point Conditions

We first recall from Rockafellar and Wets (2009) the definitions of several generalized subdifferentials.

Definition 2.1. *(see (Rockafellar and Wets, 2009, Definition 8.3)) Consider a function $h : \mathbb{R}^n \to (-\infty, +\infty]$ and a point x with $h(x)$ finite. The regular (Fréchet) subdifferential of h at x is defined as*

$$\widehat{\partial} h(x) := \left\{ v \in \mathbb{R}^n \;\Big|\; \liminf_{x\neq x'\to x} \frac{h(x') - h(x) - \langle v, x' - x\rangle}{\|x' - x\|} \ge 0 \right\};$$

the basic (limiting or Mordukhovich) subdifferential of h at x is defined as

$$\partial h(x) := \left\{ v \in \mathbb{R}^n \mid \exists\, x^k \underset{h}{\to} x \text{ and } v^k \in \widehat{\partial} h(x^k) \text{ with } v^k \to v \text{ as } k \to \infty \right\},$$

where $x^k \underset{h}{\to} x$ means that $x^k \to x$ and $h(x^k) \to h(x)$; and the horizon subdifferential of h at x is defined as

$$\partial^\infty h(x) := \left\{ v \in \mathbb{R}^n \mid \exists\, x^k \underset{h}{\to} x \text{ and } v^k \in \widehat{\partial} h(x^k) \text{ with } \lambda^k v^k \to v \text{ for some } \lambda^k \downarrow 0 \right\}.$$

For every $x \in \text{dom}h$, the set $\widehat{\partial} h(x)$ are closed and convex, but $\partial h(x)$ is generally nonconvex. The inclusion $\widehat{\partial} h(x) \subseteq \partial h(x)$ always hold, and it may be strict when h is nonconvex. By using (Rockafellar and Wets, 2009, Theorem 10.1), if a proper function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ has a local minimum at x, then $0 \in \widehat{\partial} h(x)$, and hence $0 \in \partial h(x)$.

For a proper lower semicontinuous (lsc) function $h\colon \mathbb{R}^n \to (-\infty, \infty]$, its proximal mapping associated to parameter $t > 0$ is defined by

$$\text{prox}_{th}(x) := \arg\min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2t} \|z - x\|^2 + h(z) \right\} \quad \text{for } x \in \mathbb{R}^n.$$

Consider the following nonsmooth composite optimization problem

$$\min_{x \in \mathbb{R}^n} \ \Psi(x) := \psi(x) + \phi(x), \tag{2.1}$$

where $\phi : \mathbb{R}^n \to \overline{\mathbb{R}} := (-\infty, +\infty]$ is proper lower semicontinuous, and $\psi(x)$ is a continuously differentiable on an open subset of $\mathbb{R}^n$ containing the domain of ϕ. For this class of optimization problems, we define two kinds of stationary points, critical point and L-type stationary point (Beck and Hallak (2019)).

Definition 2.2. *A vector $x \in \mathbb{R}^n$ is called a critical point of problem* (2.1) *if* $0 \in \partial \Psi(x)$, *and we denote by* $\text{crit}\Psi$ *the set of critical points of* Ψ. *A vector $x \in \mathbb{R}^n$ is called an L-type stationary point of problem* (2.1) *if there exists a constant $\mu > 0$ such that* $x \in \text{prox}_{\mu^{-1}\phi}(x - \mu^{-1} \nabla \psi(x))$.

If ϕ is assumed to be directional differentiable, we can define the directional stationary point of Ψ. A vector x is called a directional stationary point if $\Psi'(x; d) \geq 0$ for $\forall d \in \mathcal{T}_{\text{dom}\phi}(x)$. From (Li et al. (2020)) we know that if x is a directional stationary point, it is a critical point. For the reason that the objective function discussed in this thesis may not be directional differentiable, in what follows we only focus on critical point and L-stationary point.

If ϕ is assumed to be convex, then

$$0 \in \partial\Psi(x) \Leftrightarrow 0 \in \mu(x - (x - \mu^{-1}\nabla\psi(x))) + \partial\phi(x) \Leftrightarrow x = \text{prox}_{\mu^{-1}\phi}(x - \mu^{-1}\nabla\psi(x)),$$

which means that for problem (2.1) the L-stationarity of a point x is equivalent to its criticality. To extend this equivalence to a broader class of functions, we recall the definitions of prox-bounded and prox-regularity, where the later one acts as a surrogate of convexity.

Definition 2.3. *(Rockafellar and Wets, 2009, Definition 1.23 & Definition 13.27) A function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is prox-bounded if there exists $\mu > 0$ such that*

$$\inf_{z\in\mathbb{R}^n} \left\{ \frac{\mu}{2}\|z - x\|^2 + h(z) \right\} > -\infty.$$

A function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is prox-regular at a point $\overline{x} \in \text{dom}h$ for $\overline{v} \in \partial h(\overline{x})$ if h is locally lower semicontinuous at $\overline{x}$, and there exist $r \geq 0$ and $\varepsilon > 0$ such that $h(x') \geq h(x) + v^\top(x' - x) - \frac{r}{2}\|x' - x\|^2$ for all $\|x' - \overline{x}\| \leq \varepsilon$, 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 assumption of the prox-regularity of ϕ, the set of L-type stationary points for Ψ coincides with that of its critical points.

Proposition 2.1. *If $\overline{x}$ is an L-stationary point of problem* (2.1)*, then $0 \in \partial\Psi(\overline{x})$. If ϕ is prox-regular at $\overline{x}$ for $-\nabla\psi(\overline{x})$ and prox-bounded, the converse is also true.*

Proof. Pick any $\overline{x}$ from the L-type stationary points of problem (2.1). Then, by definition there exists $\mu > 0$ such that

$$\overline{x} \in \underset{x\in\mathbb{R}^n}{\arg\min} \left\{ \nabla\psi(\overline{x})^\top(x - \overline{x}) + \frac{\mu}{2}\|x - \overline{x}\|^2 + \phi(x) \right\},$$

whose first-order necessary condition is

$$0 \in \nabla\psi(\overline{x}) + \mu(\overline{x} - \overline{x}) + \partial\phi(\overline{x}) = \nabla\psi(\overline{x}) + \partial\phi(\overline{x}) = \partial F(\overline{x}).$$

Therefore, the set of L-type stationary points is contained in that of critical points. Next we argue that the converse inclusion holds. Pick any $\overline{x}$ from the critical points of (3.1). Define $\widetilde{\phi}(y) := \phi(y+\overline{x}) + \langle \nabla\psi(\overline{x}), y+\overline{x}\rangle$ for $y \in \mathbb{R}^n$. Since ϕ is prox-regular at $\overline{x}$ for $-\nabla\psi(\overline{x})$ by, the function $\widetilde{\phi}$ is prox-regular at 0 for 0. Since $\widetilde{\phi}$ is also prox-bounded, by (Rockafellar and Wets, 2009, Proposition 8.46 (f)) the subgradient inequalities in the definition of prox-regularity can be taken to be global. That is, there exists $\gamma_0 > 0$ such that $\widetilde{\phi}(y) > \widetilde{\phi}(0) - \frac{\gamma_0}{2}\|y\|^2$ for all $y \neq 0$, which implies that for all $y \neq 0$ and $\gamma > \gamma_0$,

$$\phi(y+\overline{x}) + \frac{\gamma}{2}\|y+\overline{x} - (\overline{x} - \frac{1}{\gamma}\nabla\psi(\overline{x}))\|^2 > \phi(\overline{x}) + \frac{\gamma}{2}\|\overline{x} - (\overline{x} - \frac{1}{\gamma}\nabla\psi(\overline{x}))\|^2.$$

Therefore, $\overline{x}$ is the unique minimizer of $\phi(\cdot) + \frac{\gamma}{2}\|\cdot - (\overline{x} - \frac{1}{\gamma}\nabla\psi(\overline{x}))\|^2$, which by Definition 2.2 means that $\overline{x}$ is an L-type stationary point of (2.1). Therefore, the inverse inclusion holds, and we obtain the desired result. □

2.3 Kurdyka-Łojasiewicz Property

We first present the definition of Kurdyka-Łojasiewicz (KL) Property.

Definition 2.4. *For any $\eta > 0$, we denote by Υ_η the set consisting of all continuous concave $\varphi\colon [0,\eta) \to \mathbb{R}_+$ that are continuously differentiable on $(0,\eta)$ with $\varphi(0) = 0$ and $\varphi'(s) > 0$ for all $s \in (0,\eta)$. A proper function $h\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to have the KL property at $\overline{x} \in \operatorname{dom}\partial h$ if there exist $\eta \in (0,\infty]$, $\epsilon > 0$ and a function $\varphi \in \Upsilon_\eta$ such that for all $x \in \mathbb{B}(\overline{x},\epsilon) \cap \big[h(\overline{x}) < h < h(\overline{x}) + \eta\big]$,*

$$\varphi'(h(x) - h(\overline{x}))\operatorname{dist}(0, \partial h(x)) \geq 1.$$

If φ can be chosen as $\varphi(s) = cs^{1-\theta}$ for some constant $c > 0$, then h is said to have the KL property of exponent θ at $\overline{x}$. If h has the KL property (of exponent θ) at each point of $\operatorname{dom}\partial h$, then h is called a KL function (of exponent θ).

Remark 2.1. *From (Attouch et al., 2010, Lemma 2.1), a proper lower semicontinuous function* $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ *has the KL property (of exponent* $\theta \in [0, 1]$*) at all noncritical points. Thus, to show that it is a KL function of exponent (of exponent* $\theta \in [0, 1]$*), it suffices to check its KL property (of exponent* $\theta \in [0, 1]$*) at critical points.*

The KL property can be traced back to Lojasiewicz (1963) and Kurdyka (1998), where Lojasiewicz (1963) showed that real-analytical functions satisfy the KL property with $\varphi(s) = s^{1-\theta}$ and $\theta \in [\frac{1}{2}, 1)$, while Kurdyka (1998) extended this result to differentiable functions definable in an o-minimal structure (Definition 2.5). Later, the KL property is extended to nonsmooth functions in the subanalytic and o-minimal settings, see Bolte et al. (2006, 2007a,b, 2008).

In Lojasiewicz (1984), the author proved that a bounded solution of a gradient flow for an analytic cost function converges to a well-defined limit point. Later, the KL property was used in various areas of applied mathematics, including optimization, partial differential equations, and other related fields (see Bolte et al. (2006) and the references therein). Recently, KL property has been a powerful tool in the convergence analysis of various first-order method including gradient-related method (Absil et al. (2005)), proximal algorithm (Attouch and Bolte (2009); Attouch et al. (2010)), proximal alternating linearized minimization algorithm (Bolte et al. (2014)), subgradient method (Noll (2014)), Douglas-Rachford splitting method (Li and Pong (2016)), alternating direction method of multipliers (Li and Pong (2015); Guo et al. (2017); Wang et al. (2019)), and so on. In addition, it is worth mentioning that various abstract convergence theorems via KL property were studied, see Attouch et al. (2013); Ochs et al. (2014); Frankel et al. (2015); Bolte and Pauwels (2016); Ochs (2019); Qian and Pan (2023), which provide guidance for the design of globally convergent optimization algorithms.

On the other hand, the KL property of exponent plays a crucial role in analyzing the convergence rate of the optimization algorithms, see for example, (Attouch et al., 2010, Theorem 3.4). Generally, an exponent $\theta \in (0, 1/2]$ corresponds to a linear convergence rate, while $\theta \in (1/2, 1)$ leads to a sublinear convergence rate. Specifically, as a regularity condition, the KL property with exponent 1/2 has attracted much attention. It was discussed in Bolte et al. (2017); Wang et al. (2021b); Pan and Liu (2018) that for primal lower nice functions, the KL property with exponent 1/2 is usually weaker than the metric subregularity of their subdifferential mapping or the Luo-Tseng error bound, which are the commonly used regularity conditions to achieve the liner convergence rate of the first-order methods (see Luo and Tseng (1992); Wen et al. (2017); Zhou and So (2017)). The calculus of KL exponent has also been an interesting topic. We refer the interested readers to the recent works Li and Pong (2018); Wu et al. (2021); Yu et al. (2022); Wang and Wang (2023); Li et al. (2023).

Next, we aim at discussing in which cases the considered problems in this thesis satisfy the KL property. The tool we use is the o-minimal structure. Introduced in Van den Dries (1998), the o-minimal structures can be seen as an axiomatization of the properties of semi-algebraic sets. Its formal definition is given as follows.

Definition 2.5. *Let $\mathcal{O} = \{\mathcal{O}_n\}_{n\in\mathbb{N}}$ be such that each $\mathcal{O}_n$ is a collection of subsets in $\mathbb{R}^n$. We say $\mathcal{O}$ is an o-minimal structure if the following axioms are met:*

(i) For each n, $\mathcal{O}_n$ is an boolean algebra. That is, $\emptyset \in \mathcal{O}_n$ and for each $A, B \in \mathcal{O}_n$, $A \cup B, A \cap B$ and $\mathbb{R}^n \backslash A$ belong to $\mathcal{O}_n$.

(ii) For all $A \in \mathcal{O}_n$, $A \times \mathbb{R}$ and $\mathbb{R} \times A$ belong to $\mathcal{O}_{n+1}$.

(iii) For all $A \in \mathcal{O}_{n+1}$, $\{(x_1, ..., x_n) \in \mathbb{R}^n \mid (x_1, ..., x_n, x_{n+1}) \in A\}$ belongs to $\mathcal{O}_n$.

(iv) For all $i \neq j$ in $[n]$, $\{(x_1, ..., x_n) \in \mathbb{R}^n \mid x_i = x_j\}$ belongs to $\mathcal{O}_n$.

(v) The set $\{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 < x_2\}$ *belongs to* $\mathcal{O}_2$.

(vi) The elements of $\mathcal{O}_1$ *are exactly finite union of intervals.*

A set A *is said to be definable in* $\mathcal{O}$*, if* A *belongs to* $\mathcal{O}$*. A set-valued mapping* $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ *(resp. a real-extended-valued function* $f : \mathbb{R}^n \to \overline{\mathbb{R}}$*) is said to be definable if its graph is a definable subset of* $\mathbb{R}^n \times \mathbb{R}^m$ *(resp.* $\mathbb{R}^n \times \mathbb{R}$*).*

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, we know from (Attouch et al., 2010, Section 4) that the definable functions have very nice properties, which are presented in the following lemma.

Proposition 2.2. *The following statements are true.*

(i) Finite sums of definable functions are definable;

(ii) Compositions of definable functions or mappings are definable;

(iii) Indicator functions of definable sets are definable;

(iv) Generalized inverses of definable mappings are definable.

As mentioned above, Kurdyka (1998) showed that any differentiable function definable in an o-minimal structure satisfies the KL property. This result was extended to nonsmooth setting as follows in Bolte et al. (2007a).

Theorem 2.1. *Any proper lower semicontinuous function* $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ *that is definable in an o-minimal structure* $\mathcal{O}$ *has the KL property at each point of* $\mathrm{dom}\partial h$.

In the following, we prove that the considered problems in this thesis are KL functions.

Proposition 2.3. *The following assertions are true.*

(i) Problem (1.1) *with* $f(x) = \frac{1}{2}\|Ax - b\|^2$ *or* $f(x) = \sum_{i=1}^m \log(1 + \exp(-b_i(Ax)_i))$ *is a KL function.*

(ii) Problem (4.43) *with* $f(x) = \frac{1}{2}\|Ax - b\|^2$ *or* $f(x) = \sum_{i=1}^m \log\left(1 + \frac{(Ax-b)_i^2}{\nu}\right)$ *for some* $\nu > 0$ *is a KL function.*

Proof. (i) It holds by (Van den Dries and Miller, 1996, Example 2.5) that the exp structure is an o-minimal structure, and that $a^r : \mathbb{R} \to \mathbb{R}$ defined by

$$a \mapsto \begin{cases} a^r, & a > 0 \\ 0, & a \leq 0, \end{cases} \quad \text{where } r \in \mathbb{R},$$

is definable in the exp structure. Since $a \mapsto |a|$ is semi-algebraic, and that semi-algebraic functions are definable in the exp structure (Van den Dries and Miller, 1996, Example 2.5), we obtain from Proposition 2.2 (ii) that $\lambda\|x\|_q^q$ is definable in the exp structure. This along with Proposition 2.2 implies that for any f definable in exp structure (for example, f is semi-algebraic or globally subanalytic), problem (1.1) is definable, hence a KL function by applying Theorem 2.1.

We now consider problem (1.1) with $f(x) = f_1(x) := \frac{1}{2}\|Ax - b\|^2$ and $f(x) = f_2(x) := \sum_{i=1}^m \log(1+\exp(-b_i(Ax)_i))$. It is clear that f_1 is semi-algebraic, and hence F with $f = f_1$ is a KL function. From the definition and Proposition 2.2 (ii) and (iv) we have that f_2 is definable in the exp structure, which implies that F with $f = f_2$ also meets the KL property.

(ii) Notice that Π_* is a polyhedron, hence a semi-algebraic set and definable in the exp structure. Then, it follows by Proposition 2.2 (iii) that δ_{Π_*} is definable in the exp structure. Therefore, from Proposition 2.2 (i) we conclude that for any f definable in the exp structure, problem (4.43) is a KL function. Note that both

$f_1(x) := \frac{1}{2}\|Ax-b\|^2$ and $f_2(x) := \sum_{i=1}^m \log\left(1+\frac{(Ax-b)_i^2}{\nu}\right)$ are definable in the exp structure, we conclude that (4.43) with $f = f_1$ or $f = f_2$ is a KL function. □

2.4 Proximal Gradient Method

In this section, we briefly introduce proximal gradient method for solving the following optimization problem,

$$\min_{x\in\mathbb{R}^n} \ \Psi(x) := \psi(x) + \phi(x), \tag{2.2}$$

where $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper lower semicontinuous function whose proximal mapping is accessible, $\psi : \mathbb{R}^n \to \mathbb{R}$ is a continuous differentiable function and $\nabla\psi$ is globally Lipschitz continuous over domϕ with Lipschitz constant $L > 0$. We assume that $\inf \Psi > -\infty$.

The proximal gradient method is also known as the forward-backward splitting method (Combettes and Wajs (2005)). From another point of view, it can also be interpreted as a majorization-minimization algorithm. In fact, by noting that $\nabla\psi$ is assumed as Lipschitz continuous with Lipschitz constant L on domϕ, from (Bertsekas, 1997, Proposition A.24), there is a quadratic upper bound of ψ given by

$$\psi(x) \le \psi(y) + \nabla\psi(y)^T(x-y) + \frac{\gamma}{2}\|x-y\|^2, \ \forall x, y \in \mathrm{dom}\phi, \ \gamma > L.$$

At current point $y = x^k$, the proximal gradient method obtains the next iterate by minimizing an upper bound of Ψ. That is,

$$\begin{aligned} &x^{k+1} \in \arg\min_{x\in\mathbb{R}^n} \left\{\psi(x^k) + \nabla\psi(x^k)^T(x-x^k) + \frac{\gamma}{2}\|x-x^k\|^2 + \phi(x)\right\}, \\ \Leftrightarrow\ &x^{k+1} \in \mathrm{prox}_{\gamma^{-1}\phi}(x^k - \gamma^{-1}\nabla\psi(x^k)). \end{aligned}$$

The detailed iterates of the proximal gradient method are presented as follows.

Algorithm 1 (Proximal gradient method for problem (2.2))

Initialization: Choose an arbitrary $x^0 \in \text{dom}\phi$, $\gamma > L$. Set $k = 0$.
While the termination condition is not met, solve the subproblem

$$x^{k+1} \in \text{prox}_{\gamma^{-1}\phi}(x^k - \gamma^{-1}\nabla\psi(x^k)), \tag{2.3}$$

and let $k \leftarrow k+1$.
end

In large-scale setting, the Lipschitz constant of ∇f is sometimes hard to compute, for which γ in (2.3) is not available. For this case, one can perform a line search procedure to select a suitable γ such that Ψ has a descent property, see for example, Wright et al. (2009) and Gong et al. (2013). The iterates of the line search version of proximal gradient method are given as follows.

Algorithm 2 (Proximal gradient method with line search for problem (2.2))

Initialization: Choose an arbitrary $x^0 \in \text{dom}\phi$, $\alpha > 0$ and $0 < \mu_{\min} \leq \mu_{\max}$. Set $k = 0$.
While the termination condition is not met
Select $\mu_k \in [\mu_{\min}, \mu_{\max}]$. Let m_k be the smallest nonnegative integer m such that

$$\Psi(x^{k+1}) \leq \Psi(x^k) - \frac{\alpha}{2}\|x^k - x^{k+1}\|^2 \quad \text{with} \quad x^{k+1} \in \text{prox}_{(\mu_k\tau^m)^{-1}\phi}(x^k - (\mu_k\tau^m)^{-1}\nabla\psi(x^k)). \tag{2.4}$$

and let $k \leftarrow k+1$.
end

Remark 2.2. (i) *In numerical experiments,* $\mu_{\min}$ *and* $\mu_{\max}$ *is usually set as* 10^{-20} *and* 10^{20}*, respectively, and* μ_k *is usually given by the Barzilai-Borwein method, i.e.,*

$$\mu_k = \frac{(x^k - x^{k-1})^\top(\nabla f(x^k) - \nabla f(x^{k-1})}{\|x^k - x^{k-1}\|^2}.$$

Moreover, α *is usually set as a small positive constant.*

(ii) *We claim that Algorithm 2 is well defined, i.e., the line search procedure must terminate after a finite number of backtrackings. Recall that* $\nabla\psi$ *is assumed to be Lipschitz continuous over* $\text{dom}\phi$ *with Lipschitz constant* L*. Let* $x^{k,m} \in \text{prox}_{(\mu_k\tau^m)^{-1}\phi}(x^k-$

$(\mu_k\tau^m)^{-1}\nabla\psi(x^k))$ with $\mu_k\tau^m \geq L+\alpha$. It follows from the descent lemma (Bertsekas, 1997, Proposition A.24) that

$$\begin{aligned}
\Psi(x^{k,m}) &\leq \psi(x^k) + \nabla\psi(x^k)^T(x^{k,m}-x^k) + \frac{L}{2}\|x^{k,m}-x^k\|^2 + \phi(x^{k,m}) \\
&\leq \psi(x^k) + \nabla\psi(x^k)^T(x^{k,m}-x^k) + \frac{\mu_k\tau^m}{2}\|x^{k,m}-x^k\|^2 + \phi(x^{k,m}) - \frac{\alpha}{2}\|x^{k,m}-x^k\|^2 \\
&\leq \psi(x^k) + \phi(x^k) - \frac{\alpha}{2}\|x^{k,m}-x^k\|^2 = \Psi(x^k) - \frac{\alpha}{2}\|x^{k,m}-x^k\|^2,
\end{aligned}$$

where the last inequality uses the definition of $x^{k,m}$. Therefore, when $\mu_k\tau^m \geq L+\alpha$, equation (2.4) *holds, from which we deduce that the line search must terminate after a finite number of searchings, and $\mu_k\tau^{m_k} < \tau(L+\alpha)$.*

The following theorem presents the global convergence result of Algorithm 2 by virtue of KL property and (Attouch et al., 2013, Theorem 2.9). Since the analysis for Algorithm 1 is similar, we only consider that of Algorithm 2 here.

Theorem 2.2. *Assume that Ψ is level bounded, and that $\{x^k\}_{k\in\mathbb{N}}$ is generated by Algorithm 2. If Ψ is a KL function, then $\{x^k\}_{k\in\mathbb{N}}$ is a Cauchy sequence, and converges to a critical point of Ψ.*

Proof. From (Attouch et al., 2013, Theorem 2.9), it suffices to prove that there exist $\alpha_1, \alpha_2 > 0$ such that

(i) For each $k \in \mathbb{N}$, $\Psi(x^{k+1}) \leq \Psi(x^k) - \alpha_1\|x^{k+1}-x^k\|^2$;

(ii) For each $k \in \mathbb{N}$, $\mathrm{dist}(0, \partial\Psi(x^{k+1})) \leq \alpha_2\|x^{k+1}-x^k\|$;

(iii) There exists a subsequence $\{x^{k_j}\}_{j\in\mathbb{N}}$ and x^* such that $x^{k_j} \underset{\Psi}{\to} x^*$ as $j \to \infty$.

Property (i) directly holds with $\alpha_1 = \frac{\alpha}{2}$ by equation (2.4) and Remark 2.2. Next we consider property (ii). Indeed, let $\overline{\mu}_k := \mu_k\tau^{m_k}$. It follows from the first-order optimality condition and (Rockafellar and Wets, 2009, Exercise 8.8) that

$0 \in \nabla\psi(x^k) + \overline{\mu}_k(x^{k+1} - x^k) + \partial\phi(x^{k+1})$, which implies that

$$\nabla\psi(x^{k+1}) - \nabla\psi(x^k) - \overline{\mu}_k(x^{k+1} - x^k) \in \nabla\psi(x^{k+1}) + \partial\phi(x^{k+1}) = \partial\Psi(x^{k+1}).$$

Therefore,

$$\begin{aligned}\operatorname{dist}(0, \partial\Psi(x^{k+1})) &\le \|\nabla\psi(x^{k+1}) - \nabla\psi(x^k) - \overline{\mu}_k(x^{k+1} - x^k)\| \\ &\le (L + \tau(L+\alpha))\,\|x^{k+1} - x^k\|,\end{aligned}$$

where the last inequality uses the Lipschitz continuity of $\nabla\psi$ and Remark 2.2. Property (ii) holds with $\alpha_2 = L+\tau(L+\alpha)$. Finally, we prove property (iii). From property (i) we know that $\{x^k\}_{k\in\mathbb{N}} \subseteq \mathcal{L}_\Psi(x^0)$, which is a compact set since Ψ is level bounded, then $\{x^k\}_{k\in\mathbb{N}}$ has at least one accumulation point. Pick any accumulation point x^*. By definition, there exists $\{x^{k_j}\}_{j\in\mathbb{N}}$ such that $\lim_{j\to\infty} x^{k_j} = x^*$. From the definition of x^{k_j}, it holds that

$$\begin{aligned}&\psi(x^{k_j-1}) + \langle \nabla\psi(x^{k_j-1}), x^{k_j} - x^{k_j-1}\rangle + \frac{\overline{\mu}_{k_j-1}}{2}\|x^{k_j} - x^{k_j-1}\|^2 + \phi(x^{k_j}) \\ &\le \psi(x^{k_j-1}) + \langle \nabla\psi(x^{k_j-1}), x^* - x^{k_j-1}\rangle + \frac{\overline{\mu}_{k_j-1}}{2}\|x^* - x^{k_j-1}\|^2 + \phi(x^*).\end{aligned} \tag{2.5}$$

From property (i) and the fact that $\inf\Psi > -\infty$, we have $\lim_{k\to\infty}\|x^{k+1} - x^k\| = 0$, then $\lim_{j\to\infty} x^{k_j-1} = x^*$. Letting $j \to \infty$ in (2.5) and taking upper limit on both sides yield that

$$\limsup_{j\to\infty} \phi(x^{k_j}) \le \phi(x^*),$$

which together with the lower semicontinuousness of ϕ and the continuity of ψ implies that $x^{k_j} \underset{\Psi}{\to} x^*$ as $j \to \infty$. The proof is completed. □

We note here that the above theorem relies on the global Lipschitz continuity of $\nabla\psi$. For the case where $\nabla\psi$ is only assumed to be locally Lipschitz continuous, the global convergence result of proximal gradient method is also available, see the recent papers Bauschke et al. (2017); Bolte et al. (2018); Bello-Cruz et al. (2021); De Marchi and Themelis (2022); Kanzow and Mehlitz (2022); Jia et al. (2023).

Chapter 3

A Regularized Newton Method for ℓ_q-Regularization Problems

In this chapter, we consider the following problem

$$\min_{x\in\mathbb{R}^n} F(x) := f(x) + \lambda\|x\|_q^q, \tag{3.1}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable with $c_f := \inf_{z\in\mathbb{R}^n} f(z) > -\infty$. To simplify the notation, in the rest of this chapter we write $g(x) := \lambda\|x\|_q^q$ for $x \in \mathbb{R}^n$.

In this chapter, we propose a hybrid of proximal gradient (PG) and subspace regularized Newton methods (HpgSRN) for problem (3.1), which takes advantage of PG and Newton-type methods, and avoids their disadvantages. Though problem (3.1) is a special case of (1.7), our HpgSRN is quite different from ZeroFPR (Themelis et al. (2018)) and ManAcc-Newton (Bareilles et al. (2023)) mentioned in Section 1.3; see the discussions in Remark 3.1 (d) and (e).

To describe the working flow of HpgSRN, for any given $S \subseteq \{1, 2, \dots, n\}$ we define

$$F_S(u) := f_S(u) + g_S(u) \ \text{ with } \ f_S(u) := f(I_{\cdot S}u), g_S(u) := \lambda \sum_{i\in[|S|]} |u_i|^q \ \text{ for } u \in \mathbb{R}^{|S|}. \tag{3.2}$$

By Lemma 3.3, for $S = \mathrm{supp}(x)$, such F_S is twice continuously differentiable at x_S.

The main idea of HpgSRN is to use a PG method to seek a good estimate in some neighborhood of a potential critical point, and enhance the convergence speed by using a regularized Newton method in the subspace associated to the support of the iterate generated by the PG method. Specifically, with the current x^k, the PG step yields $\overline{x}^k$ by computing

$$\overline{x}^k \in \mathop{\arg\min}_{x\in\mathbb{R}^n} \Big\{ f(x^k) + \langle \nabla f(x^k), x - x^k\rangle + \frac{\overline{\mu}_k}{2}\|x - x^k\|^2 + g(x) \Big\},$$

where the step-size $\overline{\mu}_k$ depends on the (local) Lipschitz constant of ∇f near x^k. Write $S_k := \mathrm{supp}(x^k)$. If the switch condition are not satisfied, let $x^{k+1} := \overline{x}^k$ and return to the PG step; otherwise switch to a regularized Newton step where the Newton direction d^k has the form $(d^k_{S_k}; 0)$ with

$$d^k_{S_k} := \mathop{\arg\min}_{d_{S_k}\in\mathbb{R}^{|S_k|}} \Big\{ F_{S_k}(u^k) + \langle \nabla F_{S_k}(u^k), d_{S_k}\rangle + \frac{1}{2}\langle G_k d_{S_k}, d_{S_k}\rangle \Big\}, \tag{3.3}$$

where $G_k = \nabla^2 F_{S_k}(u^k) + \xi_k I$, $u^k = x^k_{S_k}$, $\xi_k = b_1\Lambda_k + b_2\|\nabla F_{S_k}(u^k)\|^\sigma$ with $b_1 > 1,\ b_2 > 0$, $\sigma \in (0, \frac{1}{2}]$ and $\Lambda_k = [-\lambda_{\min}(\nabla^2 F_{S_k}(u^k))]_+$. In Newton method with line search, to ensure a sufficient descent in objective, the generalized Hessian G_k is required to be positive definite. Since $\nabla^2 F_{S_k}(u^k)$ may not be positive semidefinite, we add $b_1\Lambda_k$ to ensure that $\nabla^2 F_{S_k}(u^k) + b_1\Lambda_k$ is at least positive semidefinite. On the other hand, to ensure that G_k is positive definite, we add $b_2\|\nabla F_{S_k}(u^k)\|^\sigma I$. When x^k is close to some critical point, $b_2\|\nabla F_{S_k}(u^k)\|^\sigma$ will approach 0, and hence it makes G_k well approach $\nabla^2 F_{S_k}(u^k)$. Under this construction, G_k is positive definite if x^k is not a critical point of F. It is easy to verify that $d^k_{S_k}$ is the unique solution of the system of linear equations

$$G_k d_{S_k} = -\nabla F_{S_k}(u^k).$$

We perform the Armijo line search along the direction d^k to seek an appropriate step-size α_k, set $x^{k+1} := x^k + \alpha_k d^k$, and then return to the PG step to guarantee that the iterate sequence has a global convergence property.

From the above statement, the iterate sequence $\{x^k\}_{k\in\mathbb{N}}$ of HpgSRN consists of two parts: the iterates given by the PG step and those generated by the subspace regularized Newton step. Some switching conditions involving $\text{sign}(x^k) = \text{sign}(\overline{x}^k)$ decide which step the next iterate x^{k+1} enters in.

The main contributions of this chapter include three aspects:

(i) We propose a hybrid of the PG and subspace regularized Newton methods for solving problem (3.1). Different from ZeroFPR and ManAcc-Newton, each iterate of HpgSRN does not necessarily perform a second-order step until sufficiently many steps are performed and the computation of the regularized Newton step fully exploits the subspace structure, which substantially reduces the computation cost. Numerical comparison with ZeroFPR indicates that HpgSRN not only requires much less computing time (especially for those problem with $n \gg m$) but also yields comparable even better sparsity and objective function values.

(ii) For the proposed HpgSRN, we achieve the global convergence of the iterate sequence under the local Lipschitz continuity of $\nabla^2 f$ on $\mathbb{R}^n$ (see Assumption 3.1), the KL property of F, and a curve-ratio condition for the subspace regularized Newton directions (see Assumption 3.2). Both Assumptions 3.1 and 3.2 are commonly used in the convergence analysis of Newton-type methods with line search.

(iii) Under Assumptions 3.1 and 3.2, if the KL property of F is strengthened to be the KL property of exponent 1/2, we establish the R-linear convergence rate of the iterate sequence. If in addition a local error bound condition holds at the limit point, the iterate sequence is shown to converge superlinearly with rate

$1+\sigma$ for $\sigma \in (0, 1/2]$. This not only removes the local optimality of the limit point as required by ZeroFPR and BELLA (Ahookhosh et al. (2021)), but also gets rid of its isolatedness as BELLA does.

The rest of this chapter is organized as follows. Section 3.1 gives some preliminaries, including the subdifferential characterization of F and the equivalence between the KL property of exponent $1/2$ of F and that of F_S. Section 3.2 presents the formal iterate steps of HpgSRN and some auxiliary results. Section 3.3 provides the global and local convergence analysis of HpgSRN. Finally, in section 3.4 we conduct numerical experiments for HpgSRN on ℓ_q quasi-norm regularized linear and logistic regressions on real data and compare its performance with ZeroFPR and the PG method with a monotone line search (PGls).

3.1 Preliminaries on ℓ_q- Regularization Problem

In this section, we present some preliminary results of problem (3.1). For the proximal mapping of g, from (Chen et al., 2010, Theorem 2.1) we have the following lemma.

Lemma 3.1. *Fix any $\mu > 0$ and $y \in \mathbb{R}^n$, if $\overline{x} \in \mathrm{prox}_{\mu g}(x)$, then it holds that $|\overline{x}|_{\min} \geq \big[\mu\lambda q(1-q)\big]^{\frac{1}{2-q}}$.*

Next, we charcterize the generalized subdifferentials of g. Since the results directly follow by Definition 2.1, the details are omitted here.

Lemma 3.2. *Fix any $x \in \mathbb{R}^n$. Then, $\widehat{\partial} g(x) = \partial g(x) = \partial g_1(x_1) \times \cdots \times \partial g_n(x_n)$ with $\partial g_i(0) = \mathbb{R}$ and $\partial g_i(x_i) = \{\lambda q \mathrm{sign}(x_i)|x_i|^{q-1}\}$ if $x_i \neq 0$.*

Recall that f is twice continuously differentiable. By combining Lemma 3.2 and (Rockafellar and Wets, 2009, Exercise 8.8), $\partial F(x) = \nabla f(x) + \partial g(x)$ for all $x \in \mathbb{R}^n$.

Therefore, the set of critical points of (3.1) is $\{x \mid -\nabla f(x) \in \partial g(x)\}$. Since g is prox-regular at $\overline{x}$ for $-\nabla f(\overline{x})$ by (Ochs, 2018, Example 2.3), we know by Proposition 2.1 that the set of critical points of problem (3.1) coincides with that of its L-type stationary points.

Next we state the differential properties of F in a subspace.

Lemma 3.3. *For the objective function F of* (3.1)*, the following statements hold.*

(i) For any given index set $S \subseteq [n]$ and any given $x \in \mathbb{R}^n \backslash \{0\}$ with $\mathrm{supp}(x) = S$*, the function F_S is twice continuously differentiable at x_S with*

$$\nabla F_S(x_S) = I_{\cdot S}^{\top} \nabla f(I_{\cdot S} x_S) + \lambda q \mathrm{sign}(x_S) \circ |x_S|^{q-1}, \tag{3.4a}$$

$$\nabla^2 F_S(x_S) = I_{\cdot S}^{\top} \nabla^2 f(I_{\cdot S} x_S) I_{\cdot S} + \lambda q(q-1)\mathrm{Diag}(|x_S|^{q-2}), \tag{3.4b}$$

and the function g_S is three times continuously differentiable at x_S with

$$D^3 g_S(x_S)(v) = \lambda q(q-1)(q-2)\mathrm{Diag}(\mathrm{sign}(x_S) \circ |x_S|^{q-3} \circ v) \quad \forall v \in \mathbb{R}^{|S|}. \tag{3.5}$$

(ii) For any given bounded set $\Xi \subseteq \mathbb{R}^n$ and any given constant $\kappa > 0$, there exist $\widehat{c}_1 > 0, \widehat{c}_2 > 0$ and $\widehat{c}_3 > 0$ such that for all $x \in \Xi \backslash \{0\}$ with $|x|_{\min} \geq \kappa$,

$$\|\nabla F_{\mathrm{supp}(x)}(x_{\mathrm{supp}(x)})\| \leq \widehat{c}_1, \ \|\nabla^2 F_{\mathrm{supp}(x)}(x_{\mathrm{supp}(x)})\|_2 \leq \widehat{c}_2,$$
$$\|D^3 g_{\mathrm{supp}(x)}(x_{\mathrm{supp}(x)})(v)\|_2 \leq \widehat{c}_3\|v\|, \ \text{for } v \in \mathbb{R}^{|\mathrm{supp}(x)|}.$$

(iii) For any $x \in \mathbb{R}^n \backslash \{0\}$, $\mathrm{dist}(0, \partial F(x)) = \|\nabla F_{\mathrm{supp}(x)}(x_{\mathrm{supp}(x)})\|$.

Proof. **(i)** The first part of (i) is immediate since g_S is continuously differentiable at x_S. To establish the second part, for any sufficiently small $v \in \mathbb{R}^{|S|}$,

$$\begin{aligned}\nabla^2 g_S(x_S + v) - \nabla^2 g_S(x_S) &= q(q-1)\mathrm{Diag}(|x_S + v|^{q-2} - |x_S|^{q-2}) \\ &= q(q-1)(q-2)\mathrm{Diag}(\mathrm{sign}(x_S) \circ |x_S|^{q-3} \circ v).\end{aligned}$$

This, by the definition of differentiability, implies the expression of $D^3 g_S(x_S)(v)$.

(ii) Notice that $g_{\text{supp}(x)}$ is smooth at those $x_{\text{supp}(x)}$ with $x \in \Xi$ and $\min_{i\in\text{supp}(x)}|x_i| \geq \kappa$. The result follows by using formula (3.4a)-(3.4b) and (3.5) and the boundedness of Ξ.

(iii) Fix any $x \in \mathbb{R}^n$. Write $S = \text{supp}(x)$. Fix any $x \in \mathbb{R}^n\backslash\{0\}$. Write $S = \text{supp}(x)$. From Lemma 3.2 and (Rockafellar and Wets, 2009, Exercise 8.8),

$$\partial F(x) = \nabla f(x) + \partial g_1(x_1) \times \cdots \times \partial g_n(x_n).$$

Then, we get $\text{dist}(0, \partial F(x)) = \|I_{\cdot S}^{\top}\nabla f(I_{\cdot S}x_S) + \lambda q\,\text{sign}(x_S) \circ |x_S|^{q-1}\|$. Together with (3.4a), the result follows. □

The following proposition establishes the equivalence between the KL property of exponent $\theta \in (0, 1)$ of F and that of F_S.

Proposition 3.1. *Let $\theta \in (0, 1)$. For any given $\overline{x} \in \mathbb{R}^n\backslash\{0\}$, F has the KL property of exponent θ at $\overline{x}$ if and only if $F_{\overline{S}}$ with $\overline{S} = \text{supp}(\overline{x})$ has the KL property of exponent θ at $\overline{u} = \overline{x}_{\overline{S}}$.*

Proof. From Lemma 3.3 (iii), one can verify that if $\overline{x} \in \mathbb{R}^n\backslash\{0\}$, $\overline{x} \in \text{crit}F$ if and only if $\overline{x}_{\overline{S}} \in \text{crit}F_{\overline{S}}$ for $\overline{S} = \text{supp}(\overline{x})$. Then, by (Attouch et al., 2010, Lemma 2.1), it suffices to consider the case that $\overline{x} \in \text{crit}F\backslash\{0\}$.

Necessity. Since F has the KL property of exponent θ at $\overline{x}$, there exist $\eta > 0, \varepsilon > 0$ and $c > 0$ such that for all $x \in \Gamma(\varepsilon, \eta) := \big\{x \in \mathbb{R}^n \,|\, \|x - \overline{x}\| \leq \varepsilon, F(\overline{x}) < F(x) < F(\overline{x}) + \eta\big\}$,

$$\text{dist}(0, \partial F(x)) \geq c[F(x) - F(\overline{x})]^{\theta}. \tag{3.6}$$

Since $\overline{x}_i \neq 0$ for each $i \in \overline{S}$, there exists $\varepsilon' > 0$ such that for all $z \in \mathbb{B}(\overline{x}, \varepsilon')$, $z_i \neq 0$ with each $i \in \overline{S}$. Set $\widetilde{\varepsilon} := \min\{\varepsilon, \varepsilon'\}$. Pick any $u \in \Gamma_{\overline{S}}(\widetilde{\varepsilon}, \eta) := \big\{u \in \mathbb{R}^{|\overline{S}|} \,|\, \|u - \overline{u}\| \leq \widetilde{\varepsilon}, F_{\overline{S}}(\overline{u}) < F_{\overline{S}}(u) < F_{\overline{S}}(\overline{u}) + \eta\big\}$. Let $x \in \mathbb{R}^n$ with $x_{\overline{S}} = u$ and $x_{\overline{S}^c} = 0$. Clearly, $\text{supp}(x) = \overline{S}$. From Lemma 3.3 (iii), it follows that $\text{dist}(0, \partial F(x)) = \|\nabla F_{\overline{S}}(u)\|$.

Also, from $F_{\overline{S}}(u) = F(x)$ and $F_{\overline{S}}(\overline{u}) = F(\overline{x})$, we have $x \in \Gamma(\varepsilon, \eta)$. Along with (3.6), we get

$$\|\nabla F_{\overline{S}}(u)\| = \text{dist}(0, \partial F(x)) \geq c[F(x) - F(\overline{x})]^{\theta} = c[F_{\overline{S}}(u) - F_{\overline{S}}(\overline{u})]^{\theta}.$$

By the arbitrariness of u in $\Gamma_{\overline{S}}(\varepsilon, \eta)$, $F_{\overline{S}}$ has the KL property of exponent θ at $\overline{u}$.

Sufficiency. Since $F_{\overline{S}}$ has the KL property of exponent θ at $\overline{u}$, there are $\widetilde{\varepsilon} > 0, \widetilde{\eta} > 0, c > 0$ such that for all $u \in \Gamma_{\overline{S}}(\widetilde{\varepsilon}, \widetilde{\eta}) := \{u \in \mathbb{R}^{|\overline{S}|} \,|\, \|u - \overline{u}\| \leq \widetilde{\varepsilon}, F_{\overline{S}}(\overline{u}) < F_{\overline{S}}(u) < F_{\overline{S}}(\overline{u}) + \widetilde{\eta}\}$,

$$\text{dist}(0, \partial F_{\overline{S}}(u)) \geq c[F_{\overline{S}}(u) - F_{\overline{S}}(\overline{u})]^{\theta}.$$

Since every entry of $\overline{u}$ is nonzero, by reducing $\widetilde{\varepsilon}$ if necessary, for any u with $\|u - \overline{u}\| \leq \widetilde{\varepsilon}$, its entries are all nonzero. By Lemma 3.3 (iii), the last inequality can be rewritten as

$$\|\nabla F_{\overline{S}}(u)\| \geq c[F_{\overline{S}}(u) - F_{\overline{S}}(\overline{u})]^{\theta}. \tag{3.7}$$

By continuity, there exists $\varepsilon' > 0$ such that for all $x \in \mathbb{B}(\overline{x}, \varepsilon')$, $\text{supp}(x) \supseteq \overline{S}$. Let $\delta := \max_{\|x - \overline{x}\| \leq 1} \|\nabla f(x)\|_{\infty}$. Set $\varepsilon := \min\{\frac{1}{4}, \widetilde{\varepsilon}, \varepsilon', (\frac{\delta+1}{\lambda q})^{\frac{1}{q-1}}\}$ and $\eta := \frac{1}{2}\min\{\widetilde{\eta}, 1\}$. Let $\Gamma'(\varepsilon, \eta) := \{x \in \Gamma(\varepsilon, \eta) \,|\, \text{supp}(x) = \overline{S}\}$ where $\Gamma(\varepsilon, \eta)$ is defined as above, and $\Gamma''(\varepsilon, \eta) := \Gamma(\varepsilon, \eta) \backslash \Gamma'(\varepsilon, \eta)$. Pick any $x \in \Gamma(\varepsilon, \eta)$. We proceed the proof by two cases.

Case 1: $x \in \Gamma'(\varepsilon, \eta)$. Let $u = x_{\overline{S}}$. We have $u \in \Gamma_{\overline{S}}(\varepsilon, \eta) \subseteq \Gamma_{\overline{S}}(\widetilde{\varepsilon}, \widetilde{\eta})$, where the second inclusion is due to $\varepsilon < \tilde{\varepsilon}$ and $\eta < \tilde{\eta}$. From Lemma 3.3 (iii) and (3.7),

$$\text{dist}(0, \partial F(x)) = \|\nabla F_{\overline{S}}(u)\| \geq c[F_{\overline{S}}(u) - F_{\overline{S}}(\overline{u})]^{\theta} = c[F(x) - F(\overline{x})]^{\theta}.$$

Case 2: $x \in \Gamma''(\varepsilon, \eta)$. Recall that $\text{supp}(x) \supseteq \overline{S}$. By the definition of $\Gamma''(\varepsilon, \eta)$, there exists $i \notin \overline{S}$ such that $0 < |x_i| \leq \varepsilon$. Write $S := \text{supp}(x)$. Since F_S is continuously differentiable at x_S by Lemma 3.3 (i), for all $i \in S \backslash \overline{S}$ it holds that

$$\begin{aligned}\text{dist}(0, \partial F(x)) &\geq |[\nabla F_S(x)]_i| = \left|[\nabla f(x)]_i + \lambda q \text{sign}(x_i)|x_i|^{q-1}\right| \\ &\geq \lambda q |x_i|^{q-1} - |[\nabla f(x)]_i| > \lambda q \varepsilon^{q-1} - \delta \geq 1,\end{aligned} \tag{3.8}$$

where the last inequality follows by the definition of ε. Since $F(\overline{x}) < F(x) < F(\overline{x})+\eta$ and $0 < \eta < 1$, we have $[F(x) - F(\overline{x})]^{\theta} < 1$. Together with (3.8), we have

$$\mathrm{dist}(0, \partial F(x)) > [F(x) - F(\overline{x})]^{\theta}.$$

From the above two cases and the arbitrariness of x in $\Gamma(\varepsilon, \eta)$, the function F has the KL property of exponent θ at $\overline{x}$. Thus, the proof is completed. □

3.2 A Hybrid of PG and Subspace Regularized Newton Methods

In this section, we describe the iterate steps of HpgSRN, a hybrid of PG and subspace regularized Newton methods for solving problem (3.1). The detailed iterates of the algorithm are shown as follows.

Remark 3.1. **(a)** *Algorithm 3 uses $\overline{\mu}_k\|x^k - \overline{x}^k\|_\infty \le \epsilon$ as the stopping rule, which by Definition 2.2 means that the output x^k is an approximate L-type stationary point.*
(b) *Every iterate of Algorithm 3 executes Step 1, but does not necessarily perform Step 2, due to the participation of the switch condition* (3.10)*. In fact, we believe that when the current iterate is far away from the critical point, PG is more cost-to-effective than the Newton method, and the switch condition is to judge whether the current iterate is close to some potential critical point. In sparse optimization, to check whether the signs of x^k and $\overline{x}^k$ are equal is an intuitive choice for switching, while the second criterium in switch condition is for convergence analysis, see Lemma 3.6 (i). Step 1 in Algorithm 3 aims to ensure the convergence of the whole iterate sequence, while Step 2 is a subspace regularized Newton step used to enhance the convergence speed whenever the iterates are stable. When setting $\epsilon = 0$ and Algorithm 3 generates an infinite sequence, we will show in Proposition 3.2 that under Assumption 3.1, after a finite number of iterates, Algorithm 3 reduces to a regularized Newton method to minimize F_{S_*} for some $S_* \subseteq [n]$.*

Algorithm 3 (a hybrid of PG and subspace regularized Newton methods)

Initialization: Choose $\tau > 1, \alpha > 0$, $\mu_{\max} > \mu_{\min} > 0, \sigma \in (0, \frac{1}{2}], \varrho \in (0, \frac{1}{2}), \beta \in (0,1)$, $b_1 > 1$ and $b_2 > 0$. Choose an initial $x^0 \in \mathbb{R}^n$ and a tolerance $\epsilon \geq 0$. Let $k = 0$.

Step 1: proximal gradient step

(1a) Choose an initial step-size $\mu_k \in [\mu_{\min}, \mu_{\max}]$. Let m_k be the smallest integer m such that

$$F(\overline{x}^k) \leq F(x^k) - \frac{\alpha}{2}\|x^k - \overline{x}^k\|^2 \ \text{ with } \ \overline{x}^k \in \mathrm{prox}_{(\mu_k\tau^m)^{-1}g}(x^k - (\mu_k\tau^m)^{-1}\nabla f(x^k)). \tag{3.9}$$

(1b) Let $\overline{\mu}_k = \mu_k\tau^{m_k}$. **If** $\overline{\mu}_k\|x^k - \overline{x}^k\|_\infty \leq \epsilon$, output x^k; **otherwise** go to (1c).

(1c) Let $\overline{\omega}_k = \overline{\mu}_k + \lambda q(q-1)|\overline{x}^k|_{\min}^{q-2}$. **If**

$$\mathrm{sign}(x^k) = \mathrm{sign}(\overline{x}^k) \ \text{ and } \ \overline{\mu}_k + \lambda q(q-1)|x^k|_{\min}^{q-2} \geq \frac{1}{2}\overline{\omega}_k, \tag{3.10}$$

then go to **Step 2**; **otherwise** let $x^{k+1} = \overline{x}^k$ and $k \leftarrow k+1$. Go to **Step 1**.

Step 2: subspace regularized Newton step

(2a) Let $S_k = \mathrm{supp}(x^k)$ and $u^k = x^k_{S_k}$. Seek a subspace Newton direction $d^k_{S_k}$ by solving $G^k d^k_{S_k} = -\nabla F_{S_k}(u^k)$, where $G^k = \nabla^2 F_{S_k}(u^k) + (b_1\Lambda_k + b_2\|\nabla F_{S_k}(u^k)\|^\sigma)I$ with $\Lambda_k = [-\lambda_{\min}(\nabla^2 F_{S_k}(u^k))]_+$. Let $d^k_{S_k^c} = 0$.

(2b) Let t_k be the smallest nonnegative integer t such that

$$F_{S_k}(u^k + \beta^t d^k_{S_k}) \leq F_{S_k}(u^k) + \varrho\beta^t\langle\nabla F_{S_k}(u^k), d^k_{S_k}\rangle. \tag{3.11}$$

(2c) Let $\alpha_k = \beta^{t_k}$ and $x^{k+1} = x^k + \alpha_k d^k$ and $k \leftarrow k+1$. Go to **Step 1**.

(c) *We claim that Algorithm 3 is well defined, i.e., the line search procedures in (1a) and (2b) of Algorithm 3 must hold after a finite number of backtrackings.*

We first argue that the number of backtrackings in (1a) is finite. For this purpose, define $h_\mu(z;x) := \langle\nabla f(x), z - x\rangle + \frac{\mu}{2}\|z - x\|^2 + g(z)$ *for* $z \in \mathbb{R}^n$*. For each* $m \in \mathbb{N}$, *pick* $x^{k,m} \in \mathrm{prox}_{(\mu_k\tau^m)^{-1}g}\big(x^k - (\mu_k\tau^m)^{-1}\nabla f(x)\big)$, *then it follows that*

$$\begin{aligned} h_{\mu_k}(x^{k,m}; x^k) &\le \langle \nabla f(x^k), x^{k,m} - x^k\rangle + \frac{\mu_k\tau^m}{2}\|x^{k,m} - x^k\|^2 + g(x^{k,m}) \\ &\le g(x^k) = h_{\mu_k}(x^k; x^k). \end{aligned} \tag{3.12}$$

Since $h_{\mu_k}(\cdot; x^k)$ *is continuous and coercive, the set*

$$\mathcal{L}_k := \{z \in \mathbb{R}^n \mid h_{\mu_k}(z; x^k) \le h_{\mu_k}(x^k; x^k)\}$$

is compact. Since ∇f *is continuously differentiable, there exists* $L_k > 0$ *such that for any* $y, w \in \mathcal{L}_k$, $\|\nabla f(y) - \nabla f(w)\| \le L_k\|y - w\|$. *When* $\mu_k\tau^m \ge L_k + \alpha$, *from* $x^k, x^{k,m} \in \mathcal{L}_k$ *and the descent lemma (Bertsekas, 1997, Proposition A.24),*

$$\begin{aligned} F(x^{k,m}) &\le f(x^k) + \langle \nabla f(x), x^{k,m} - x^k\rangle + \frac{L_k}{2}\|x^{k,m} - x^k\|^2 + g(x^{k,m}) \\ &\le f(x^k) + \langle \nabla f(x^k), x^{k,m} - x^k\rangle + \frac{\mu_k\tau^m}{2}\|x^{k,m} - x^k\|^2 + g(x^{k,m}) - \frac{\alpha}{2}\|x^{k,m} - x^k\|^2 \\ &\le f(x^k) + g(x^k) - \frac{\alpha}{2}\|x^{k,m} - x^k\|^2 = F(x^k) - \frac{\alpha}{2}\|x^{k,m} - x^k\|^2, \end{aligned} \tag{3.13}$$

where the last inequality is due to (3.12). *This implies that the line search procedure stops in the* m*-th backtracking. The above arguments only use the Lipschitz continuity of* ∇f *on the set* $\mathcal{L}_k$ *rather than its global Lipschitz continuity, and the coercivity of* $h_{\mu_k}(\cdot; x^k)$ *rather than that of* g. *For more discussion on line search of PG methods in a general setting, see also Bello Cruz and Nghia (2016); Salzo (2017) for the convex* f *and De Marchi and Themelis (2022); Kanzow and Mehlitz (2022); Jia et al. (2023) for the nonconvex* f.

Next we argue that the line search in (2b) will terminate after a finite number of backtrackings. We see that when the iteration goes from Step 1 to Step 2, it is necessary that $S_k \ne \emptyset$. *As in this case* (3.10) *is satisfied,* $x^k \ne 0$ *must hold. If not, by* (3.10), $\overline{x}^k = 0$. *So the termination condition in (1b) is satisfied and the algorithm* *stops. By Lemma 3.3 (i), F_{S_k} is continuously differentiable at u^k, which along with $G^k \succ 0$ implies that*

$$\langle \nabla F_{S_k}(u^k), d^k_{S_k} \rangle = -\langle G^k d^k_{S_k}, d^k_{S_k} \rangle < 0, \tag{3.14}$$

i.e., $d^k_{S_k}$ is a descent direction of F_{S_k} at u^k. In addition, F_{S_k} is bounded from below on $\mathbb{R}^{|S_k|}$ because f is bounded from below on $\mathbb{R}^n$. By following the same arguments as those for (Nocedal and Wright, 2006, Lemma 3.1), the smallest nonnegative integer t_k satisfying (3.11) *exists. Therefore, combining the last part, we conclude that Algorithm 3 is well defined.*

From the iterate steps of Algorithm 3, the sequence $\{x^k\}_{k\in\mathbb{N}}$ consists of two parts, i.e., $\{x^k\}_{k\in\mathbb{N}} = \{x^k\}_{k\in\mathcal{K}_1} \cup \{x^k\}_{k\in\mathcal{K}_2}$, where

$$\mathcal{K}_1 := \left\{k \in \mathbb{N} \mid x^{k+1} \text{ is generated by Step 1}\right\} \quad \text{and} \quad \mathcal{K}_2 := \mathbb{N}\backslash\mathcal{K}_1. \tag{3.15}$$

It is clear now that for $k \in \mathcal{K}_2$, $S_k \neq \emptyset$, that is, x^k has a nonempty support.

(d) *Although Algorithm 3 is a hybrid of PG and second-order methods, it is not a special case of ZeroFPR (Themelis et al. (2018)) and FBTN (Themelis et al. (2019)) due to the following four aspects. Firstly, each iterate of Algorithm 3 does not necessarily perform Newton step, while each iterate of ZeroFPR and FBTN must execute a second-order step. Secondly, Algorithm 3 is using the Armijo line search, which is different from the ones used in ZeroFPR and FBTN. Let F_γ denote the forward-backward envelope of F associated to $\gamma > 0$, and $\eta > 0$ be a constant related to the (local) Lipschitz constant of ∇f. For* (3.1)*, the line search of ZeroFPR is to seek the smallest nonnegative integer t_k of those t's such that*

$$F_\gamma(\overline{x}^k + \beta^t \overline{d}^k) - F_\gamma(x^k) \leq -\eta\|x^k - \overline{x}^k\|^2.$$

Then set $x^{k+1} = \overline{x}^k + \beta^{t_k}\overline{d}^k$, where $\overline{d}^k$ is a Newton-type direction at $\overline{x}^k$ rather than x^k; and the line search of FBTN is to seek the smallest nonnegative integer t_k of *those t's such that*

$$F_\gamma\left((1-\beta^t)\overline{x}^k+\beta^t(x^k+d^k)\right)-F_\gamma(x^k)\le -\eta\|x^k-\overline{x}^k\|^2,$$

and then set $x^{k+1}=(1-\beta^{t_k})\overline{x}^k+\beta^{t_k}(x^k+d^k)$, *where* d^k *is a second-order direction at* x^k. *We observe that the decrease of the successive iterates for ZeroFPR and FBTN, i.e.* $F_\gamma(x^{k+1})-F_\gamma(x^k)$, *is controlled by* $-\|x^k-\overline{x}^k\|^2$, *while the decrease of the successive iterates for Step 2 of Algorithm 3, i.e.,* $F(x^{k+1})-F(x^k)$, *is controlled by the curve ratio* $\alpha_k\langle\nabla F_{S_k}(u^k),d^k_{S_k}\rangle$. *Thirdly, the line search procedures of ZeroFPR and FBTN involve computing the forward-backward envelope of* F, *which means that prox-gradient evaluations are needed at each backtracking trial and this is not the case for (2b) of Algorithm 3.2. Finally, the global convergence analysis of ZeroFPR requires its second-order direction* d^k *to satisfy*

$$\exists \text{ a constant } \widehat{c}>0 \text{ such that } \|d^k\|\le\widehat{c}\|x^k-\overline{x}^k\| \text{ for all } k, \tag{3.16}$$

but now it is unclear whether the regularized Newton direction in (2a) satisfies (3.16) *or not.*

(e) *Our algorithm is similar to the Newton acceleration framework of the PG method proposed in (Bareilles et al. (2023)), which first uses the PG method to identify the underlying manifold substructure of* (3.1) *and then accelerates it with a Riemannian Newton method. However, our algorithm is not a special case of this framework due to the following facts. Firstly, similar to ZeroFPR and FBTN, the framework in (Bareilles et al. (2023)) executes a Newton step in each iteration. As discussed in part (d), our algorithm adaptively executes a Newton step by condition* (3.10), *which avoids some unnecessary waste in second-order step. Secondly, the Riemannian Hessian was used to yield the Newton directions in (Bareilles et al. (2023)), while a regularized one is used in our algorithm to yield the Newton directions. Thirdly, a quadratic convergence rate of the iterate sequence was established in (Bareilles et al.*

(2023)) by assuming that the Riemannian Hessian is positive definite at the limit point. However, under weaker conditions we show that the generated sequence is convergent and has a superlinear convergence rate; see Theorems 3.1 and 3.2, respectively.

To conduct the convergence analysis of Algorithm 3 with $\epsilon = 0$ in the next section, from now on we assume that $x^k \neq \overline{x}^k$ for all k (if not, Algorithm 3 yields an L-type stationary point within a finite number of steps), i.e., Algorithm 3 generates an infinite sequence $\{x^k\}_{k\in\mathbb{N}}$. The following lemma shows that the sequences $\{x^k\}_{k\in\mathbb{N}}$ and $\{\overline{x}^k\}_{k\in\mathbb{N}}$ are bounded, and the sequence $\{\overline{\mu}_k\}_{k\in\mathbb{N}}$ is upper bounded. The latter will be used to derive a uniform lower bound for $|\overline{x}^k|_{\min}$; see Lemma 3.6 (i) later.

Lemma 3.4. *The following assertions hold for $\{x^k\}_{k\in\mathbb{N}}, \{\overline{x}^k\}_{k\in\mathbb{N}}$ and $\{\overline{\mu}_k\}_{k\in\mathbb{N}}$.*

(i) The sequence $\{F(x^k)\}_{k\in\mathbb{N}}$ is nonincreasing and convergent, and consequently, $\{x^k\}_{k\in\mathbb{N}} \subseteq \mathcal{L}_F(x^0) := \{x \in \mathbb{R}^n \mid F(x) \le F(x^0)\}$ and $\{\overline{x}^k\}_{k\in\mathbb{N}} \subseteq \mathcal{L}_F(x^0)$.

(ii) $\{x^k\}_{k\in\mathbb{N}}$ and $\{\overline{x}^k\}_{k\in\mathbb{N}}$ are bounded, the cluster point set of $\{x^k\}_{k\in\mathbb{N}}$, denoted by $\omega(x^0)$, is nonempty and compact, and F is constant on $\omega(x^0)$.

(iii) For all $k \in \mathbb{N}$, $\overline{\mu}_k < \widetilde{L} := \max\{\mu_{\max} + 1, \tau(2\widehat{L} + \alpha)\}$, where $\widehat{L}$ is the Lipschitz constant of ∇f on the set $\mathcal{L}_F(x^0) + \overline{\tau}\mathbf{B}$ with $\overline{\tau} := \frac{\tau_0 + \sqrt{\tau_0^2 + 2\widetilde{c}_f \mu_{\min}}}{\mu_{\min}}$. Here, $\tau_0 := \max_{x\in\mathcal{L}_F(x^0)} \|\nabla f(x)\|$ and $\widetilde{c}_f = F(x^0) - c_f$.

Proof. **(i)** Fix any $k \in \mathbb{N}$. When $k \in \mathcal{K}_1$, $x^{k+1} = \overline{x}^k$, and by (3.9), $F(x^{k+1}) \le F(x^k)$. When $k \in \mathcal{K}_2$, it follows from (3.11) and (3.14) that

$$F_{S_k}(u^{k+1}) \le F_{S_k}(u^k) + \varrho\beta^{m_k}\langle \nabla F_{S_k}(u^k), d^k_{S_k}\rangle \le F_{S_k}(u^k),$$

which along with $S_{k+1} \subseteq S_k$ implies that $F(x^{k+1}) \le F(x^k)$. The two cases show that $\{F(x^k)\}_{k\in\mathbb{N}}$ is nonincreasing, which along with the lower boundedness of F means that $\{F(x^k)\}_{k\in\mathbb{N}}$ is convergent. The nonincreasing behavior of $\{F(x^k)\}_{k\in\mathbb{N}}$, together with $F(\overline{x}^k) \le F(x^k)$ for each $k \in \mathbb{N}$, implies that $F(\overline{x}^k) \le F(x^k) \le F(x^0)$ for each $k \in \mathbb{N}$, and consequently, $\{x^k\}_{k\in\mathbb{N}} \subseteq \mathcal{L}_F(x^0)$ and $\{\overline{x}^k\}_{k\in\mathbb{N}} \subseteq \mathcal{L}_F(x^0)$.

(ii) Since g is coercive and f is lower bounded, the level set $\mathcal{L}_F(x^0)$ is compact. By part (i), $\{x^k\}_{k\in\mathbb{N}}$ and $\{\overline{x}^k\}_{k\in\mathbb{N}}$ are bounded, so the set $\omega(x^0)$ is nonempty. Using the same arguments as in (Bolte et al., 2014, Lemma 5 (iii)) yields the compactness of $\omega(x^0)$. Pick any $x^* \in \omega(x^0)$. There exists a subsequence $\{x^{k_j}\}_{j\in\mathbb{N}}$ such that $\lim_{j\to\infty} x^{k_j} = x^*$. By the continuity of F and the convergence of $\{F(x^k)\}_{k\in\mathbb{N}}$, we have $F(x^*) = \lim_{j\to\infty} F(x^{k_j}) = F^*$, where F^* is the limit of $\{F(x^k)\}_{k\in\mathbb{N}}$. This means that F is constant on the set $\omega(x^0)$.

(iii) Define $\mathcal{K} := \{k \in \mathbb{N} \mid \overline{\mu}_k > \mu_k\}$. If $\mathcal{K}$ is empty, the desired result holds because $\overline{\mu}_k = \mu_k \le \mu_{\max} < \widetilde{L}$ for all $k \in \mathbb{N}$, so we assume that $\mathcal{K} \ne \emptyset$. Write $\widehat{\mu}_k := \overline{\mu}_k/\tau$ and $\widehat{x}^k := \mathrm{prox}_{\widehat{\mu}_k^{-1} g}(x^k - \widehat{\mu}_k^{-1}\nabla f(x^k))$ for each $k \in \mathcal{K}$. We first argue that

$$\|\widehat{x}^k - x^k\| \le \overline{\tau} \ \text{ for each } k \in \mathcal{K}. \tag{3.17}$$

Since $\widehat{\mu}_k < \overline{\mu}_k$, by (3.9) we have $F(\widehat{x}^k) > F(x^k) - \frac{\alpha}{2}\|\widehat{x}^k - x^k\|^2$, which implies that $\widehat{x}^k \ne x^k$ for each $k \in \mathcal{K}$. For each $k \in \mathcal{K}$, from the definition of $\widehat{x}^k$, we have

$$\langle \nabla f(x^k), \widehat{x}^k - x^k \rangle + \frac{\widehat{\mu}_k}{2}\|\widehat{x}^k - x^k\|^2 + g(\widehat{x}^k) - g(x^k) \le 0. \tag{3.18}$$

By using Cauchy-Schwarz inequality and the nonnegativity of g, it follows that

$$\begin{aligned}
\frac{\widehat{\mu}_k}{2}\|\widehat{x}^k - x^k\|^2 &\le \|\nabla f(x^k)\|\|\widehat{x}^k - x^k\| + g(x^k) - g(\widehat{x}^k) \\
&\le \|\nabla f(x^k)\|\|\widehat{x}^k - x^k\| + F(x^k) - f(x^k) \\
&\le \|\nabla f(x^k)\|\|\widehat{x}^k - x^k\| + F(x^0) - c_f \le \tau_0\|\widehat{x}^k - x^k\| + \widetilde{c}_f,
\end{aligned}$$

where the third inequality is due to $F(x^k) \le F(x^0)$ and $f(x^k) \ge c_f$, and the last one is by the definitions of τ_0 and $\widetilde{c}_f$. For each $k \in \mathcal{K}$, since $\widehat{\mu}_k \ge \mu_k \ge \mu_{\min}$, from the last inequality,

$$\frac{\mu_{\min}}{2}\|\widehat{x}^k - x^k\|^2 - \tau_0\|\widehat{x}^k - x^k\| - \widetilde{c}_f \leq 0.$$

This, by the definition of $\overline{\tau}$, implies that inequality (3.17) holds. Now for each $k \in \mathcal{K}$, by the mean-value theorem, there exists z^k on the line segment connecting x^k and $\widehat{x}^k$ such that

$$F(\widehat{x}^k) - F(x^k) = \langle \nabla f(z^k), \widehat{x}^k - x^k \rangle + g(\widehat{x}^k) - g(x^k).$$

Substituting this equality into (3.18) and using $F(\widehat{x}^k) - F(x^k) > -\frac{\alpha}{2}\|x^k - \widehat{x}^k\|^2$ yields that

$$\begin{aligned}\frac{\widehat{\mu}_k - \alpha}{2}\|x^k - \widehat{x}^k\|^2 &< \langle \nabla f(\xi^k) - \nabla f(x^k), \widehat{x}^k - x^k \rangle \\ &\leq \|\nabla f(x^k) - \nabla f(\xi^k)\|\|\widehat{x}^k - x^k\|.\end{aligned}$$

From part (i) and (3.17), $\{x^k\}_{k\in\mathcal{K}} \subseteq \mathcal{L}_F(x^0)$ and $\{\widehat{x}^k\}_{k\in\mathcal{K}} \subseteq \mathcal{L}_F(x^0) + \overline{\tau}\mathbf{B}$. Hence, $\{z^k\}_{k\in\mathcal{K}} \subseteq \mathcal{L}_F(x^0) + \overline{\tau}\mathbf{B}$. From the last inequality, for each $k \in \mathcal{K}$,

$$\frac{\widehat{\mu}_k - \alpha}{2}\|x^k - \widehat{x}^k\| < \|\nabla f(x^k) - \nabla f(\xi^k)\| \leq \widehat{L}\|x^k - \xi^k\| \leq \widehat{L}\|x^k - \widehat{x}^k\|.$$

Thus, $\widehat{\mu}_k < 2\widehat{L} + \alpha$ and $\overline{\mu}_k < \tau(2\widehat{L} + \alpha)$ for each $k \in \mathcal{K}$. The proof is completed. □

For any given $\gamma > 0, s \in \mathbb{R}$, define a real-valued function $h : \mathbb{R} \to \mathbb{R}$ by

$$h_{\gamma,s}(t) := \frac{\gamma}{2}(t - s)^2 + \lambda|t|^q \quad \text{for } t \in \mathbb{R}. \tag{3.19}$$

It is easy to see that $t = 0$ is always a local minimizer of $h_{\gamma,s}$ and that the absolute value of another possible local minimizer is greater than $\overline{\nu}$, where $\overline{\nu} := \left(\frac{\lambda q(1-q)}{\gamma}\right)^{\frac{1}{2-q}}$. In next lemma, we will establish the existence of a uniform lower bound ϖ of $h''_{\gamma,s}$ at its nonzero local minimizer for any $\gamma > 0$ and $s \in \mathbb{R}$. We will show that the existence of such ϖ will ultimately lead to the validity of the second condition of (3.10) for some k in any large enough interval and hence, together with the validity of the first condition of (3.10), the infinite cardinality of $\mathcal{K}_2$. Indeed, if for all the integers k in any large enough interval, x^{k+1} is produced by Step 1, then the sufficient decrease property in (1a) of Step 1 implies that

$$F(x^k) - F(x^{k+1}) \geq \frac{\alpha}{2}\|x^k - x^{k+1}\|^2 \text{ (with } x^{k+1} = \overline{x}^k).$$

Summing this up for all such integers, it follows from the lower boundedness of F that $\sum \|x^k - x^{k+1}\|^2$ is bounded. Thus, for some integer k, $\|x^k - x^{k+1}\|$ should be sufficiently small. By using an integral mean-value theorem, $|x^k|_{\min}^{q-2} - |\overline{x}^k|_{\min}^{q-2}$ is bounded by $\|x^k - \overline{x}^k\|$. Therefore $|x^k|_{\min}^{q-2} - |\overline{x}^k|_{\min}^{q-2}$ should be sufficiently small. If so, it is true that $\frac{\varpi}{2} + \lambda q(q-1)(|x^k|_{\min}^{q-2} - |\overline{x}^k|_{\min}^{q-2}) \geq 0$, which implies that the second condition of (3.10) holds for some integer k.

Lemma 3.5. *For any given* $0 < \upsilon < M < \infty$*, there exists a constant* $\varpi > 0$ *such that for any* $\gamma > 0$ *and* $s \in \mathbb{R}$ *with* $|\bar{t}(\gamma, s)| \in [\upsilon, M]$*,*

$$h''_{\gamma,s}(\bar{t}(\gamma, s)) = \gamma + \lambda q(q-1)|\bar{t}(\gamma, s)|^{q-2} \geq \varpi.$$

Proof. Suppose that the conclusion does not hold. Then, there exist sequences $\{\gamma_k\}_{k\in\mathbb{N}} \subseteq \mathbb{R}_{++}$ and $\{s_k\}_{k\in\mathbb{N}} \subseteq \mathbb{R}$ with $|\bar{t}(\gamma_k, s_k)| \in [\upsilon, M]$ such that $h''_{\gamma_k,s_k}(\bar{t}(\gamma_k, s_k)) \leq \frac{1}{k}$ for all $k \in \mathbb{N}$. For each $k \in \mathbb{N}$, write $\bar{t}_k := \bar{t}(\gamma_k, s_k)$ and $\vartheta_k := h_{\gamma_k,s_k}$. Clearly, there exists $\widetilde{k} \in \mathbb{N}$ such that for all $k > \widetilde{k}$, $\vartheta''_k(\bar{t}_k) < \frac{\kappa\upsilon}{10} := \varepsilon$, where $\kappa := \lambda q(q–1)(q–2)M^{q-3}$. By the expression of ϑ_k, for any t with $|t| \in (0, M]$, the following inequality holds:

$$|\vartheta'''_k(t)| = \lambda q(q-1)(q-2)|t|^{q-3} \geq \kappa. \tag{3.20}$$

Fix any $k > \widetilde{k}$. We proceed the arguments by $\bar{t}_k \in [\upsilon, M]$ and $\bar{t}_k \in [-M, -\upsilon]$.

Case 1: $\bar{t}_k \in [\upsilon, M]$**.** Since $\vartheta''_k(\bar{t}_k) < \varepsilon$ and $\vartheta'''_k(t) > \kappa$ for $t \in (0, M]$, by the integral mean-value theorem, $\vartheta''_k(\bar{t}_k) > \vartheta''_k(\bar{t}_k - \frac{\varepsilon}{\kappa}) + \varepsilon$, which by $\vartheta''_k(\bar{t}_k) < \varepsilon$ implies that $\vartheta_k''(\bar{t}_k - \frac{\varepsilon}{\kappa}) < 0$. Together with $\vartheta_k''(\bar{t}_k) > 0$ (see (Hu et al., 2017, Lemma 14)), there exists $0 < \delta < \frac{\varepsilon}{\kappa}$ such that $\vartheta_k''(\bar{t}_k - \delta) = 0$. Recall that $\vartheta_k'''(t) > \kappa > 0$ for all $t \in (0, M]$. Then,

$$\vartheta_k''(t) < 0 \quad \text{for } t \in (0, \bar{t}_k - \delta) \quad \text{and} \quad \vartheta_k''(t) > 0 \quad \text{for } t \in (\bar{t}_k - \delta, M]. \tag{3.21}$$

Note that $\vartheta_k'(\bar{t}_k) = 0$. This, along with the second inequality in (3.21), implies that $\vartheta_k'(\bar{t}_k - \delta) < 0$. Also, since $0 < \vartheta_k''(t) < \varepsilon$ for all $t \in (\bar{t}_k - \delta, \bar{t}_k)$, from the integral mean-value theorem, $\vartheta_k'(\bar{t}_k - \delta) > \vartheta_k'(\bar{t}_k) - \varepsilon\delta = -\varepsilon\delta$, and then $\vartheta_k'(\bar{t}_k - \delta) \in (-\varepsilon\delta, 0)$. Next we argue that there exists a point $\tilde{t}_k \in (\bar{t}_k - \delta - \sqrt{2\varepsilon\delta/\kappa}, \bar{t}_k - \delta)$ such that $\vartheta_k'(\tilde{t}_k) = 0$, which along with the first inequality in (3.21) implies that

$$\vartheta_k'(t) > 0 \text{ for } t \in (0, \tilde{t}_k) \quad \text{and} \quad \vartheta_k'(t) < 0 \text{ for } t \in (\tilde{t}_k, \bar{t}_k - \delta). \tag{3.22}$$

Indeed, for any $t \in (0, \bar{t}_k - \delta)$, using $\vartheta_k''(\bar{t}_k - \delta) = 0$ and inequality (3.20) yields that

$$\begin{aligned} -\varepsilon\delta < \vartheta_k'(\bar{t}_k - \delta) &= \vartheta_k'(t) + \int_t^{\bar{t}_k - \delta} \vartheta_k''(s) ds = \vartheta_k'(t) + \int_t^{\bar{t}_k - \delta} \left[\vartheta_k''(s) - \vartheta_k''(\bar{t}_k - \delta)\right] ds \\ &\leq \vartheta_k'(t) + \int_t^{\bar{t}_k - \delta} \kappa(s - \bar{t}_k + \delta) ds = \vartheta_k'(t) - \frac{\kappa}{2}(t - \bar{t}_k + \delta)^2, \end{aligned}$$

which implies that $\vartheta_k'(t) > 0$ for all $t \leq \bar{t}_k - \delta - \sqrt{2\varepsilon\delta/\kappa}$. Along with $\vartheta_k'(\bar{t}_k - \delta) < 0$, there exists $\tilde{t}_k \in (\bar{t}_k - \delta - \sqrt{\frac{2\varepsilon\delta}{\kappa}}, \bar{t}_k - \delta)$ such that $\vartheta_k'(\tilde{t}_k) = 0$.

From (3.21) we deduce that ϑ_k' is decreasing in $(\tilde{t}_k, \bar{t}_k - \delta)$ and is increasing in $(\bar{t}_k - \delta, \bar{t}_k)$, which means that $\vartheta_k'(t) \geq \vartheta_k'(\bar{t}_k - \delta) > -\varepsilon\delta$ for all $t \in (\tilde{t}_k, \bar{t}_k)$. Then,

$$\begin{aligned} \vartheta_k(\bar{t}_k) - \vartheta_k(\tilde{t}_k) &= \int_{\tilde{t}_k}^{\bar{t}_k} \vartheta_k'(s) ds > -\varepsilon\delta(\bar{t}_k - \tilde{t}_k) > -\varepsilon\delta\left(\delta + \sqrt{\frac{2\varepsilon\delta}{\kappa}}\right) \qquad (3.23) \\ &> -\left(\frac{\varepsilon^3}{\kappa^2} + \sqrt{2}\frac{\varepsilon^3}{\kappa^2}\right) > -3\frac{\varepsilon^3}{\kappa^2} = -\frac{3\kappa}{1000}\upsilon^3, \end{aligned}$$

where the third inequality is due to $0 < \delta < \frac{\varepsilon}{\kappa}$. On the other hand, we have

$$\begin{aligned}\vartheta_k(\tilde{t}_k) - \vartheta_k(0) &= \int_0^{\tilde{t}_k}\int_{\tilde{t}_k}^{s} \vartheta_k''(\tau)d\tau ds = \int_0^{\tilde{t}_k}\int_{\tilde{t}_k}^{s}[\vartheta_k''(\tau) - \vartheta_k''(\bar{t}_k - \delta)]d\tau ds \\ &\geq \int_0^{\tilde{t}_k}\int_{\tilde{t}_k}^{s} \kappa(\tau - \bar{t}_k + \delta)d\tau ds = \int_0^{\tilde{t}_k} \frac{\kappa}{2}s^2 - \frac{\kappa}{2}\tilde{t}_k^2 + \kappa(\bar{t}_k - \delta)(\tilde{t}_k - s)ds \\ &= \frac{\kappa}{6}\tilde{t}_k^3 - \frac{\kappa}{2}\tilde{t}_k^3 + \frac{\kappa\tilde{t}_k^2}{2}(\bar{t}_k - \delta) \geq \frac{\kappa}{6}\tilde{t}_k^3 \geq \frac{\kappa}{6}\Big(\bar{t}_k - \delta - \sqrt{\frac{2\varepsilon\delta}{\kappa}}\Big)^3 \geq \frac{\kappa}{6}\Big(\upsilon - \frac{3\varepsilon}{\kappa}\Big)^3,\end{aligned} \tag{3.24}$$

where the first equality is due to $\vartheta_k'(\tilde{t}_k) = 0$, the second one is using $\vartheta_k''(\bar{t}_k - \delta) = 0$, the first inequality is using (3.20) and the last inequality is due to $0 < \delta < \frac{\varepsilon}{\kappa}$ and $\bar{t}_k \geq \upsilon$. Thus, from (3.23) and (3.24) and $\varepsilon := \frac{\kappa\upsilon}{10}$, we have $\vartheta_k(\bar{t}_k) - \vartheta_k(0) > \frac{465\kappa}{6000}\upsilon^3 > 0$, contradicting that $\bar{t}_k$ is a global minimizer of $\vartheta_k = h_{\gamma_k, s_k}$. The conclusion then holds.

Case 2: $\bar{t}_k \in [-M, -\upsilon]$. By using the similar arguments to those for Case 1, one can verify that the conclusion holds. Here, the details are omitted. □

To provide a sufficient condition for the switching condition (3.10), we introduce the following notation that will be used in the subsequent analysis:

$$\overline{S}_k := \mathrm{supp}(\overline{x}^k) \;\text{ and }\; \overline{u}^k := \overline{x}^k_{\overline{S}_k} \;\text{ for each } k \in \mathbb{N}.$$

Lemma 3.6. *Let $\{x^k\}_{k\in\mathbb{N}}$ and $\{\overline{x}^k\}_{k\in\mathbb{N}}$ be generated by Algorithm 3, and write $\nu := [\widetilde{L}^{-1}\lambda q(1-q)]^{\frac{1}{2-q}}$, with $\widetilde{L}$ being the one in Lemma 3.4 (iii). Then, the following statements hold.*

(i) $|\overline{x}^k|_{\min} > \nu$ for all $k \in \mathbb{N}$, and $|x^k|_{\min} > \nu$ for all $k \in \mathcal{K}_2$.

(ii) $\overline{\omega}_k \geq \varpi$ for all $k \in \mathbb{N}$, where $\overline{\omega}_k$ is the one in (1c) of Algorithm 3, and ϖ is the one in Lemma 3.5 with $\upsilon = \nu$ and $M = \big(\frac{F(x^0) - c_f}{\lambda}\big)^{\frac{1}{q}}$.

(iii) For each $k \in \mathbb{N}$, if $|x^k|_{\min} > \frac{\nu}{2}$ and $\|x^k - \overline{x}^k\| \le \min\big\{\frac{\nu}{3}, \frac{2^{q-3}\varpi}{2\lambda q(1-q)(2-q)\nu^{q-3}}\big\}$, then condition (3.10) *holds.*

Proof. **(i)** By using Lemma 3.1 with $\mu = \overline{\mu}_k^{-1}$ and $y = x^k - \overline{\mu}_k^{-1}\nabla f(x^k)$ for each $k \in \mathbb{N}$ and noting that $\mu_{\min} \le \overline{\mu}_k < \widetilde{L}$ from Lemma 3.4 (iii), we have $|\overline{x}^k|_{\min} > \nu$ for all k. To argue that $|x^k|_{\min} > \nu$ for all $k \in \mathcal{K}_2$, we only need to prove that $|x^k|_{\min} > \nu$ if x^k satisfies condition (3.10). Indeed, the second condition in (3.10) is equivalent to $|x^k|_{\min}^{q-2} \le \frac{\overline{\mu}_k}{2\lambda q(1-q)} + \frac{1}{2}|\overline{x}^k|_{\min}^{q-2}$, which by $\overline{\mu}_k < \widetilde{L}$ and the definition of ν means that

$$|x^k|_{\min}^{q-2} < \frac{\widetilde{L}}{2\lambda q(1-q)} + \frac{1}{2}\nu^{q-2} = \nu^{q-2},$$

where the equality is using the expression of ν. Thus, $|x^k|_{\min} > \nu$ for all $k \in \mathcal{K}_2$.
(ii) From (3.9), $F(\overline{x}^k) \le F(x^k)$ for each $k \in \mathbb{N}$. Then,

$$c_f + \lambda\|\overline{x}^k\|_q^q \le f(\overline{x}^k) + \lambda\|\overline{x}^k\|_q^q = F(\overline{x}^k) \le F(x^k) \le F(x^0),$$

which implies that $\|\overline{x}^k\|_q^q \le \lambda^{-1}(F(x^0) - c_f)$, and then $|\overline{x}_i^k| \le \big(\frac{F(x^0)-c_f}{\lambda}\big)^{1/q}$ for each $i \in \overline{S}_k$. In addition, from part (i), $|\overline{x}_i^k| > \nu$ for each $i \in \overline{S}_k$. For each k, let $y^k := x^k - \overline{\mu}_k^{-1}\nabla f(x^k)$. Then, $\overline{x}_i^k \in \arg\min_{t\in\mathbb{R}} h_{\overline{\mu}_k, y_i^k}(t)$ for each $i \in \overline{S}_k$, where $h_{\overline{\mu}_k, y_i^k}$ is defined by (3.19). Now by invoking Lemma 3.5 with $\upsilon = \nu$, $M = \big(\frac{F(x^0)-c_f}{\lambda}\big)^{\frac{1}{q}}$ and $\overline{t}(\overline{\mu}_k, y_i^k) = \overline{x}_i^k$ for all $i \in \overline{S}_k$, we obtain $\overline{\omega}_k = \overline{\mu}_k + \lambda q(q-1)|\overline{x}^k|_{\min}^{q-2} \ge \varpi$.
(iii) Fix any $k \in \mathbb{N}$. We first prove that the equality in (3.10) holds. From part (i), $|\overline{x}^k|_{\min} > \nu$, while from the given condition, $|x^k|_{\min} > \frac{\nu}{2}$. If there exists an index $i \in [n]$ such that $\mathrm{sign}(x_i^k) \ne \mathrm{sign}(\overline{x}_i^k)$, then $\|x^k - \overline{x}^k\| \ge |x_i^k - \overline{x}_i^k| > \frac{\nu}{2}$, which is a contradiction to $\|x^k - \overline{x}^k\| < \nu/3$. Thus, $\mathrm{sign}(x^k) = \mathrm{sign}(\overline{x}^k)$, and hence $S_k = \overline{S}_k$. For the inequality in (3.10), from part (ii), it suffices to argue that $\frac{\varpi}{2} + \lambda q(q-1)(|x^k|_{\min}^{q-2} - |\overline{x}^k|_{\min}^{q-2}) \ge 0$ or $|x^k|_{\min}^{q-2} - |\overline{x}^k|_{\min}^{q-2} \le \frac{\varpi}{2\lambda q(1-q)}$. Indeed, by invoking the

integral mean-value theorem,

$$
\begin{aligned}
&|x^k|_{\min}^{q-2} - |\overline{x}^k|_{\min}^{q-2} = \int_{|x^k|_{\min}}^{|\overline{x}^k|_{\min}} (2-q)t^{q-3}dt \\
&\leq (2-q)\big(\min\{|x^k|_{\min}, |\overline{x}^k|_{\min}\}\big)^{q-3}\big||x^k|_{\min} - |\overline{x}^k|_{\min}\big| \\
&< (2-q)(\nu/2)^{q-3}\big||x^k|_{\min} - |\overline{x}^k|_{\min}\big| \leq (2-q)(\nu/2)^{q-3}\|x^k - \overline{x}^k\| \leq \frac{\varpi}{2\lambda q(1-q)},
\end{aligned}
$$

where the second inequality is by $|\overline{x}^k|_{\min} > \nu$ and $|x^k|_{\min} > \frac{\nu}{2}$, the third one is due to $S_k = \overline{S}_k$, and the last one is using $\|x^k - \overline{x}^k\| < \frac{2^{q-3}\varpi}{2\lambda q(1-q)(2-q)\nu^{q-3}}$. □

From Lemma 3.6, we obtain the following corollary, stating that $\mathcal{K}_2$ contains infinite indices, so HpgSRN is different from PG method. In the next section, we will improve this result so that after a finite number of steps, the iterates of Algorithm 3 always enter into Step 2.

Corollary 3.1. *There exists $\widetilde{k} \in \mathbb{N}$ such that for any $k_1, k_2 \in \mathbb{N}$ with $k_2 - k_1 > \widetilde{k}$, $[k_1 : k_2] \cap \mathcal{K}_2 \neq \emptyset$, so $\mathcal{K}_2$ is an infinite set and Algorithm 3 is different from PG method.*

Proof. Let $\delta = \min\{\frac{\nu}{3}, \frac{2^{q-3}\varpi}{2\lambda q(1-q)(2-q)\nu^{q-3}}\}$ and $\widetilde{k} = \lceil \frac{2(F(x^0)-c_f)}{\alpha\delta^2} \rceil$. We argue by contradiction that the result holds. If not, there must exist $\widehat{k}_1, \widehat{k}_2 \in \mathbb{N}$ with $\widehat{k}_2 - \widehat{k}_1 > \widetilde{k}$ such that $[\widehat{k}_1 : \widehat{k}_2] \cap \mathcal{K}_2 = \emptyset$. Clearly, $[\widehat{k}_1 : \widehat{k}_2] \subseteq \mathcal{K}_1$. By the definition of $\mathcal{K}_1$, for every $k-1 \in [\widehat{k}_1 : \widehat{k}_2 - 1]$, x^k is obtained by the PG step, which by Lemma 3.6 (i) implies that $|x^k|_{\min} > \nu$ and then $\|\overline{x}^k - x^k\| \geq \delta$ must hold (if not, by Lemma 3.6 (iii), $[\widehat{k}_1 + 1 : \widehat{k}_2]$ would contain an index of $\mathcal{K}_2$). For every $k \in [\widehat{k}_1 : \widehat{k}_2] \subset \mathcal{K}_1$, we also have $x^{k+1} = \overline{x}^k$. By (3.9), for every $k \in [\widehat{k}_1 : \widehat{k}_2]$, $F(x^{k+1}) \leq F(x^k) - \frac{\alpha}{2}\|\overline{x}^k - x^k\|^2$, and then

$$
\frac{2\big(F(x^{\widehat{k}_1+1}) - c_f\big)}{\alpha} \geq \frac{2\big(F(x^{\widehat{k}_1+1}) - F(x^{\widehat{k}_2+1})\big)}{\alpha} \geq \sum_{i=\widehat{k}_1+1}^{\widehat{k}_2} \|\overline{x}^k - x^k\|^2 \geq (\widehat{k}_2 - \widehat{k}_1)\delta^2,
$$

where the last inequality is due to $\|\overline{x}^k - x^k\| \geq \delta$ for every $k \in [\widehat{k}_1{+}1{:}\widehat{k}_2]$. Together with $F(x^{\widehat{k}_1+1}) \leq F(x^0)$, we obtain $\widehat{k}_2 - \widehat{k}_1 \leq \frac{2(F(x^0)-c_f)}{\alpha\delta^2} \leq \widetilde{k}$, a contradiction to the given condition $\widehat{k}_2 - \widehat{k}_1 > \widetilde{k}$. The proof is then completed. $\square$

3.3 Convergence Analysis

In this part, we analyze the convergence rate of the objective function value sequence $\{F(x^k)\}_{k\in\mathbb{N}}$, and establish the global convergence of the iterate sequence $\{x^k\}_{k\in\mathbb{N}}$ and its superlinear convergence rate. Throughout this section, we write

$$r^k := \nabla F_{S_k}(u^k) \ \text{ and } \ H^k := \nabla^2 F_{S_k}(u^k) \ \text{ for each } k \in \mathcal{K}_2.$$

First, we give several technical lemmas that are used for the subsequent convergence analysis. The following lemma states that the subsequences $\{r^k\}_{k\in\mathcal{K}_2}$ and $\{d^k\}_{k\in\mathcal{K}_2}$ are bounded, and the subsequence $\{r^k\}_{k\in\mathcal{K}_2}$ is lower bounded by $\{\|u^k - \overline{u}^k\|\}_{k\in\mathcal{K}_2}$. The latter is crucial to control $F(x^{k+1}) - F(x^k)$ by using $-\|x^k - \overline{x}^k\|^2$; see Lemma 3.9.

Lemma 3.7. *Let $\{x^k\}_{k\in\mathbb{N}}$ be generated by Algorithm 3. The following holds.*

- *(i) There exists a constant $r_{\max} > 0$ such that $\|r^k\| \leq r_{\max}$ and $\|d^k\| \leq b_2^{-1} r_{\max}^{1-\sigma}$ for all $k \in \mathcal{K}_2$, where b_2 is the one in (2a) of Algorithm 3.*
- *(ii) For each $k \in \mathcal{K}_2$, $\|r^k\| \geq \frac{\varpi}{4}\|u^k - \overline{u}^k\|$ where ϖ is the same as in Lemma 3.6 (ii).*

Proof. **(i)** Fix any $k \in \mathcal{K}_2$. By Remark 3.1 (c), we know that $S_k \neq \emptyset$. From Lemma 3.6 (i), $|x_i^k| > \nu$ for all $i \in S_k$. By invoking Lemma 3.3 (ii) with $\kappa = \nu/2$ and $\Xi = \big\{z \in \mathcal{L}_F(x^0) \,|\, |z_i| \geq \nu/2 \text{ for all } i \in S_k\big\}$, there exists $r_{\max} > 0$ (independent of k) such that $\|r^k\| \leq r_{\max}$. Together with $\lambda_{\min}(G^k) \geq b_2\|r^k\|^\sigma$, it follows that

$$\|d^k\| = \|d_{S_k}^k\| \leq \|(G^k)^{-1}\|_2\|r^k\| \leq b_2^{-1}\|r^k\|^{1-\sigma} \leq b_2^{-1} r_{\max}^{1-\sigma}. \tag{3.25}$$

(ii) Fix any $k \in \mathcal{K}_2$. Write $v^k := u^k - \overline{\mu}_k^{-1} I_{\cdot S_k}^{\top} \nabla f(I_{\cdot S_k} u^k) = u^k - \overline{\mu}_k^{-1}[\nabla f(x^k)]_{S_k}$. Let $h_k(u) := \sum_{i=1}^{|S_k|} h_{\overline{\mu}_k, v_i^k}(u_i)$ for $u \in \mathbb{R}^{|S_k|}$, where $h_{\overline{\mu}_k, v_i^k}$ is the function defined in (3.19) with $(\gamma, s) = (\overline{\mu}_k, v_i^k)$. From (3.10), $\text{sign}(x^k) = \text{sign}(\overline{x}^k)$, and then $S_k = \overline{S}_k$. Therefore, we have $\overline{u}^k \in \arg\min_{u \in \mathbb{R}^{|S_k|}} h_k(u)$, whose optimality condition is given by

$$0 = \nabla h_k(\overline{u}^k) = \overline{\mu}_k(\overline{u}^k - v^k) + \lambda q \text{sign}(\overline{u}^k) \circ |\overline{u}^k|^{q-1}. \tag{3.26}$$

In addition, by combining Lemma 3.6 (ii) and the inequality in (3.10), it holds that

$$\varpi/2 \le \overline{\omega}_k/2 \le \overline{\mu}_k + \lambda q(q-1)|x^k|_{\min}^{q-2} = \overline{\mu}_k + \lambda q(q-1)|u^k|_{\min}^{q-2} = h''_{\overline{\mu}_k, v_i^k}(|u^k|_{\min}). \tag{3.27}$$

Define the index sets $\mathcal{I}_1^k := \{i \in [|S_k|] \mid u_i^k > 0\}$ and $\mathcal{I}_2^k := [|S_k|] \backslash \mathcal{I}_1^k$. For each $i \in [|S_k|]$, write $\widetilde{u}_i^k := \text{sign}(u_i^k) \min\{|u_i^k|, |\overline{u}_i^k|\}$. Note that each $h_{\overline{\mu}_k, v_i^k}$ is smooth at any $t \neq 0$, and $h''_{\overline{\mu}_k, v_i^k}$ is nonincreasing at $(-\infty, 0)$ and nondecreasing at $(0, \infty)$. From (3.27) and Lemma 3.6 (ii), it follows that $h''_{\overline{\mu}_k, v_i^k}(\widetilde{u}_i^k) \ge \varpi/2$ for all $i \in [|S_k|]$. Consequently, there exists $\varepsilon > 0$ such that for each $i \in \mathcal{I}_1^k$, $h''_{\overline{\mu}_k, v_i^k}(t) > \frac{\varpi}{4}$ when $t \in (\widetilde{u}_i^k - \varepsilon, \infty)$; and for each $i \in \mathcal{I}_2^k$, $h''_{\overline{\mu}_k, v_i^k}(t) > \frac{\varpi}{4}$ when $t \in (-\infty, \widetilde{u}_i^k + \varepsilon)$. Define

$$\Omega_k := \big\{u \in \mathbb{R}^{|S_k|} \mid u_i > \widetilde{u}_i^k - \varepsilon \text{ for } i \in \mathcal{I}_1^k \text{ and } u_i < -\widetilde{u}_i^k + \varepsilon \text{ for } i \in \mathcal{I}_2^k\big\}.$$

Then, h_k is twice continuously differentiable on the convex set Ω_k with $\nabla^2 h_k(u) \succ \frac{\varpi}{4} I$ for all $u \in \Omega_k$, which implies that $\widetilde{h}_k(u) := h_k(u) - \frac{\varpi}{8}\|u - v^k\|^2$ is strongly convex on the set Ω_k. From (3.26) and the expression of $\widetilde{h}_k$, clearly, $\nabla \widetilde{h}_k(\overline{u}^k) = \frac{\varpi}{4}(v^k - \overline{u}^k)$. Let $\widehat{u}^k := u^k + \frac{4}{\varpi}\nabla \widetilde{h}_k(u^k)$. By the convexity of $\widetilde{h}_k$ on Ω_k and $u^k, \overline{u}^k \in \Omega_k$, we have

$$0 \le \langle \nabla \widetilde{h}_k(\overline{u}^k) - \nabla \widetilde{h}_k(u^k), \overline{u}^k - u^k \rangle = \frac{\varpi}{4}\langle (v^k - \overline{u}^k) - (\widehat{u}^k - u^k), \overline{u}^k - u^k \rangle,$$

which implies that $\|u^k - \overline{u}^k\| \le \|v^k - \widehat{u}^k\| = \|u^k - \overline{\mu}_k^{-1}[\nabla f(x^k)]_{S_k} - \widehat{u}^k\|$. Together with $\frac{\varpi}{4}(\widehat{u}^k - u^k) = \nabla\widetilde{h}_k(u^k) = (\overline{\mu}_k - \frac{\varpi}{4})(u^k - v^k) + \lambda q \text{sign}(u^k) \circ |u^k|^{q-1}$, it follows that

$$
\begin{aligned}
\|r^k\| &= \big\|[\nabla f(x^k)]_{S_k} + \lambda q \text{sign}(u^k) \circ |u^k|^{q-1}\big\| \\
&= \big\|[\nabla f(x^k)]_{S_k} - (\overline{\mu}_k - \frac{\varpi}{4})(u^k - v^k) + \frac{\varpi}{4}(u^k - \widehat{u}^k)\big\| \\
&= \frac{\varpi}{4}\|\overline{\mu}_k^{-1}[\nabla f(x^k)]_{S_k} - u^k + \widehat{u}^k\| \geq \frac{\varpi}{4}\|u^k - \overline{u}^k\|,
\end{aligned}
$$

where the third equality is by the definition of v^k. The proof is completed. □

Assumption 3.1. *$\nabla^2 f$ is locally Lipschitz continuous on $\mathbb{R}^n$.*

Assumption 3.1 is a common one in the convergence analysis of Newton-type methods (see, e.g., Yue et al. (2019); Mordukhovich et al. (2023)). It is satisfied for third differentiable loss functions such as least square function, logistic regression, student's t-loss function (Aravkin et al. (2012)), high-order portfolio loss (Zhou and Palomar (2021)), the nonlinear least square loss $\sum_{i=1}^m (b_i - \phi_i(A_i.x))^2$ with ϕ_i being smooth, and the Log-Cosh dice loss function (Jadon (2020)) etc., while it is violated for the least absolute deviation and Huber loss as they are not twice continuously differentiable.

By the Heine-Borel open covering theorem, one can show that under Assumption 3.1 the Hessian $\nabla^2 f$ is Lipschitz continuous on any compact subset of $\mathbb{R}^n$. We next use this fact to prove that $\{\alpha_k\}_{k\in\mathcal{K}_2}$ has a uniform lower bound, which will be employed to establish the sufficient decrease of $\{F(x^k)\}_{k\in\mathbb{N}}$; see Lemma 3.9.

Lemma 3.8. *Under Assumption 3.1 there is $\underline{\alpha} > 0$ such that for all $k \in \mathcal{K}_2$, $\alpha_k \geq \underline{\alpha}$.*

Proof. Let $\Xi := \mathcal{L}_F(x^0) + \frac{1}{2}\nu\mathbf{B}$. By invoking Assumption 3.1, there exists a constant $L_2 > 0$ such that

$$
\|\nabla^2 f(y) - \nabla^2 f(z)\|_2 \leq L_2\|(y - z)\| \quad \forall y, z \in \Xi. \tag{3.28}
$$

Fix any integer $m \geq 0$ with $\beta^m \leq \min\{1, \frac{1}{2}\nu b_2 r_{\max}^{\sigma-1}\}$, where ν is the same as the one in Lemma 3.6. Fix any $k \in \mathcal{K}_2$. From $d^k_{S_k^c} = 0$, $|x_i^k| > \nu$ for all $i \in S_k$ (Lemma 3.6 (i)) and Lemma 3.7 (i), we have $\mathrm{sign}(x^k+\tau\beta^m d^k) = \mathrm{sign}(x^k)$ and $|x^k+\tau\beta^m d^k|_{\min} > \frac{\nu}{2}$ for all $\tau \in [0,1]$. By Lemma 3.3 (i), F_{S_k} is twice continuously differentiable on an open set containing the line segment between u^k and $u^k+\beta^m d^k_{S_k}$. From the mean-value theorem,

$$\begin{aligned}
&F_{S_k}(u^k+\beta^m d^k_{S^k}) - F_{S_k}(u^k) - \langle r^k, \beta^m d^k_{S_k}\rangle \\
&= \frac{1}{2}\beta^{2m}\langle \nabla^2 F_{S_k}(u^k+\tau_k\beta^m d^k_{S_k})d^k_{S_k}, d^k_{S_k}\rangle \ \text{ for some } \tau_k \in [0,1]. \qquad (3.29)
\end{aligned}$$

Note that $x^k+\tau\beta^m d^k \in \Xi$ for all $\tau \in [0,1]$ by Lemma 3.7 (i). By using Lemma 3.3 (ii) with $\kappa = \nu/2$, there exists a constant $\widehat{c}_3 > 0$ (independent of k) such that

$$\begin{aligned}
\|\nabla^2 g_{S_k}(u^k) - \nabla^2 g_{S_k}(u^k+\tau_k\beta^m d^k_{S_k})\|_2 &\leq \int_0^{\tau_k} \|D^3 g_{S_k}(u^k+t\beta^m d^k_{S_k})(\beta^m d^k_{S_k})\|_2 dt \\
&\leq \tau_k \widehat{c}_3 \beta^m \|d^k_{S_k}\|.
\end{aligned}$$

In addition, since $x^k, x^k+\tau_k\beta^m d^k \in \Xi$, using inequality (3.28) with $y = x^k$ and $z = x^k+\tau_k\beta^m d^k$ and noting that $\mathrm{supp}(x^k) = \mathrm{supp}(x^k+\tau_k\beta^m d^k) = S_k$, we have

$$\|I_{\cdot S_k}^\top \nabla^2 f(I_{\cdot S_k} u^k) I_{\cdot S_k} - I_{\cdot S_k}^\top \nabla^2 f\left(I_{\cdot S_k}(u^k+\tau_k\beta^m d^k_{S_k})\right) I_{\cdot S_k}\|_2 \leq \tau_k L_2 \beta^m \|d^k_{S_k}\|.$$

From the last two inequalities with the expression of $\nabla^2 F_{S_k}$, it follows that

$$\|\nabla^2 F_{S_k}(u^k) - \nabla^2 F_{S_k}(u^k+\tau_k\beta^m d^k_{S_k})\|_2 \leq (L_2+\widehat{c}_3)\beta^m \|d^k_{S_k}\|. \qquad (3.30)$$

Combining (3.29)-(3.30) with (2a) of Algorithm 3 and recalling that $H^k = \nabla^2 F_{S_k}(u^k)$, we obtain

$$
\begin{aligned}
&F_{S_k}(u^k) - F_{S_k}(u^k+\beta^m d^k_{S_k}) + \varrho\beta^m\langle r^k, d^k_{S_k}\rangle \\
&= (1-\varrho)\beta^m \left\langle (H^k+b_1\Lambda_k I+b_2\|r^k\|^\sigma I)d^k_{S_k}, d^k_{S_k}\right\rangle \\
&\quad -\frac{1}{2}\beta^{2m}\langle d^k_{S_k}, \nabla^2 F_{S_k}(u^k+\tau_k\beta^m d^k_{S_k})d^k_{S_k}\rangle \\
&\geq \frac{1}{2}b_2\beta^m\|r^k\|^\sigma\|d^k_{S_k}\|^2 + \frac{1}{2}\beta^{2m}\left\langle\left(H^k-\nabla^2 F_{S_k}(u^k+\tau_k\beta^m d^k_{S_k})\right) d^k_{S_k}, d^k_{S_k}\right\rangle \\
&\geq \frac{1}{2}b_2\beta^m\|r^k\|^\sigma\|d^k_{S_k}\|^2 - \frac{1}{2}(L_2+\widehat{c}_3)\beta^{3m}\|d^k_{S_k}\|^3 \\
&= \frac{1}{2}\beta^m\|d^k_{S_k}\|^3\Big(b_2\frac{\|r^k\|^\sigma}{\|d^k_{S_k}\|} - \widetilde{c}_3\beta^{2m}\Big) \ \text{ with } \widetilde{c}_3 := L_2+\widehat{c}_3,
\end{aligned}
\tag{3.31}
$$

where the first equality is using $r^k = -G^k d^k_{S_k}$ by (2a) of Algorithm 3, and the first inequality is due to $H^k + b_1\Lambda_k I \succeq 0$, $\varrho \in (0, \frac{1}{2}]$ and $\Lambda_k \geq 0$. By the definition of $d^k_{S_k}$ and Lemma 3.7 (i),

$$
\frac{\|d^k_{S_k}\|}{\|r^k\|^\sigma} \leq \frac{\|(G^k)^{-1}\|_2\|r^k\|}{\|r^k\|^\sigma} \leq \frac{\|r^k\|^{1-2\sigma}}{b_2} \leq \frac{r_{\max}^{1-2\sigma}}{b_2}. \tag{3.32}
$$

The above arguments demonstrate that whenever $\beta^m \leq \min\left\{1, \frac{1}{2}\nu b_2 r_{\max}^{\sigma-1}, \frac{b_2}{\sqrt{\widetilde{c}_3 r_{\max}^{1-2\sigma}}}\right\}$,

$$
F_{S_k}(u^k) - F_{S_k}(u^k + \beta^m d^k_{S_k}) + \varrho\beta^m\langle r^k, d^k_{S_k}\rangle \geq 0.
$$

Let $\underline{\alpha} := \beta\min\left\{1, \frac{1}{2}\nu b_2 r_{\max}^{\sigma-1}, \frac{b_2}{\sqrt{\widetilde{c}_3 r_{\max}^{1-2\sigma}}}\right\}$. Then, for all $k \in \mathcal{K}_2$, $\alpha_k \geq \underline{\alpha}$. □

3.3.1 Convergence Rate of Objective Value Sequence

We have achieved the convergence of the sequence $\{F(x^k)\}_{k\in\mathbb{N}}$ in Lemma 3.4 (i). To establish its convergence rate, we need two technical lemmas. Among others, Lemma 3.9 states that $\{F(x^k)\}_{k\in\mathbb{N}}$ is sufficiently decreasing under Assumption 3.1, while Lemma 3.10 reveals that under Assumption 3.1 the subsequence $\{d^k\}_{k\in\mathcal{K}_2}$ converges to 0.

Lemma 3.9. *Let $\{x^k\}_{k\in\mathbb{N}}$ and $\{\overline{x}^k\}_{k\in\mathbb{N}}$ be the sequences yielded by Algorithm 3. Then, under Assumption 3.1, the following assertions hold.*

(i) There exists $\widehat{\gamma}>0$ such that for all $k\in\mathbb{N}$, $F(x^{k+1})\le F(x^k)-\frac{\widehat{\gamma}}{2}\|x^k-\overline{x}^k\|^2$.

(ii) $\lim_{k\to\infty}\|x^k-\overline{x}^k\|=0$.

(iii) Every element of $\omega(x^0)$ is an L-type stationary point of (3.1).

Proof. **(i)-(ii)** By Lemma 3.4 (i), $\{x^k\}_{k\in\mathbb{N}}$ is contained in the compact set $\mathcal{L}_F(x^0)$, while $|x^k|_{\min}>\nu$ for all $k\in\mathcal{K}_2$ by Lemma 3.6 (i). Then, by invoking Lemma 3.3 (ii) with $\Xi=\mathcal{L}_F(x^0)$ and $\kappa=\nu$, there exists $\widehat{c}_2>0$ (independent of k) such that $\|H^k\|_2=\|\nabla^2 F_{S_k}(u^k)\|_2\le\widehat{c}_2$ for all $k\in\mathcal{K}_2$. Together with the expression of G^k in (2a) and Lemma 3.7 (i), for all $k\in\mathcal{K}_2$,

$$G^k\preceq[(1+b_1)\|H^k\|_2+b_2\|r^k\|^{\sigma}]I\preceq[(1+b_1)\widehat{c}_2+b_2r_{\max}^{\sigma}]I. \tag{3.33}$$

From the line search step in (3.11), Lemma 3.7 (ii) and Lemma 3.8, for all $k\in\mathcal{K}_2$,

$$\begin{aligned}F(x^{k+1})-F(x^k)&\le\alpha_k\varrho\langle r^k,d_{S_k}^k\rangle=-\alpha_k\varrho\langle r^k,(G^k)^{-1}r^k\rangle\\ &\le-\frac{\varrho\underline{\alpha}}{(1+b_1)\widehat{c}_2+b_2r_{\max}^{\sigma}}\|r^k\|^2\le-\frac{\varrho\underline{\alpha}\varpi^2}{16[(1+b_1)\widehat{c}_2+b_2r_{\max}^{\sigma}]}\|x^k-\overline{x}^k\|^2,\end{aligned} \tag{3.34}$$

where the last inequality is using $\mathrm{sign}(x^k)=\mathrm{sign}(\overline{x}^k)$ implied by $k\in\mathcal{K}_2$. In addition, by (3.9), $F(x^{k+1})\le F(x^k)-\frac{\alpha}{2}\|x^k-\overline{x}^k\|^2$ for all $k\in\mathcal{K}_1$. Along with the last inequality, part (i) holds with $\widehat{\gamma}=\min\big\{\frac{\varrho\underline{\alpha}\varpi^2}{8[(1+b_1)\widehat{c}_2+b_2r_{\max}^{\sigma}]},\alpha\big\}$. From part (i) and the convergence of $\{F(x^k)\}_{k\in\mathbb{N}}$, we obtain part (ii).

(iii) Pick any $x^*\in\omega(x^0)$. There exists a subsequence $\{x^{k_j}\}_{j\in\mathbb{N}}$ such that $x^{k_j}\to x^*$ as $j\to\infty$. From part (ii), $\lim_{j\to\infty}\overline{x}^{k_j}=x^*$. For each $j\in\mathbb{N}$, from (1a) of Algorithm 3 we have $\overline{x}^{k_j}\in\mathrm{prox}_{\overline{\mu}_{k_j}^{-1}g}(x^{k_j}-\overline{\mu}_{k_j}^{-1}\nabla f(x^{k_j}))$; while by Lemma 3.4 (iii), $\overline{\mu}_{k_j}\in[\mu_{\min},\widetilde{L})$. We assume that $\overline{\mu}_{k_j}\to\overline{\mu}_*$ (if necessary taking a subsequence). Define the mapping

$\mathcal{F}(\mu, x) := \text{prox}_{\mu^{-1}g}(x - \mu^{-1}\nabla f(x))$ for $x \in \mathbb{R}^n$ and $\mu \in [\mu_{\min}, \widetilde{L})$. By (Bonnans and Shapiro, 2013, Proposition 4.4), the mapping $\mathcal{F}$ is upper semicontinuous, so it is outer semicontinuous at $(\overline{\mu}_*, x^*)$ by (Facchinei and Pang, 2003, p. 138-139). Thus, $x^* \in \text{prox}_{\overline{\mu}_*^{-1}g}(x^* - \overline{\mu}_*^{-1}\nabla f(x^*))$, and the result follows. □

Lemma 3.10. *Let $\{x^k\}_{k\in\mathbb{N}}$ and $\{\overline{x}^k\}_{k\in\mathbb{N}}$ be the sequences given by Algorithm 3. Then, under Assumption 3.1 there exists a constant $\widehat{c}_2 > 0$ such that for all $k \in \mathcal{K}_2$,*

$$\text{dist}(0, \partial F(x^k)) \leq \widetilde{c}_2 \|x^k - \overline{x}^k\| \quad \text{with } \widetilde{c}_2 = \widehat{L} + \widetilde{L} + \widehat{c}_2,$$

where $\widetilde{L}$ and $\widehat{L}$ are the ones in Lemma 3.4, and consequently, $\lim_{\mathcal{K}_2 \ni k \to \infty} \|r^k\| = 0$ and $\lim_{\mathcal{K}_2 \ni k \to \infty} \|d^k\| = 0$.

Proof. Fix any $k \in \mathcal{K}_2$. Since $\overline{x}^k \in \text{prox}_{\overline{\mu}_k^{-1}g}\big(x^k - \overline{\mu}_k^{-1}\nabla f(x^k)\big)$, by (Rockafellar and Wets, 2009, Exercise 8.8), we have $0 \in \nabla f(x^k) + \overline{\mu}_k(\overline{x}^k - x^k) + \partial g(\overline{x}^k)$, which implies that

$$\nabla f(\overline{x}^k) - \nabla f(x^k) + \overline{\mu}_k(x^k - \overline{x}^k) \in \partial F(\overline{x}^k).$$

Recall that ∇f is Lipschitz continuous on the compact set $\mathcal{L}_F(x^0)$ with Lipschitz constant not more than $\widehat{L}$, which is the same as the one appearing in the proof of Lemma 3.4 (iii). Then, $\|\nabla f(x^k) - \nabla f(\overline{x}^k)\| \leq \widehat{L}\|x^k - \overline{x}^k\|$. Together with the last inclusion, using $\overline{\mu}_k < \widetilde{L}$ by Lemma 3.4 (iii) yields that

$$\|\nabla F_{S_k}(\overline{u}^k)\| = \text{dist}(0, \partial F(\overline{x}^k)) \leq (\widehat{L} + \widetilde{L})\|x^k - \overline{x}^k\|. \tag{3.35}$$

Let Ξ be a bounded open convex set containing $\mathcal{L}_F(x^0)$. By Lemma 3.4 (i) and the convexity of Ξ, $\overline{u}^k + \tau(u^k - \overline{u}^k) \in \Xi$ for all $\tau \in [0, 1]$. Recall that $k \in \mathcal{K}_2$. Hence, $x^k \neq 0$ and $\text{sign}(u^k) = \text{sign}(\overline{u}^k)$. Together with Lemma 3.6 (i), for all $\tau \in [0, 1]$, we have $\big|\overline{u}^k + \tau(u^k - \overline{u}^k)\big|_{\min} \geq \nu$ and $\text{sign}(\overline{u}^k + \tau(u^k - \overline{u}^k)) = \text{sign}(u^k)$. By Lemma 3.3 (ii), there exists a constant $\widehat{c}_2 > 0$ (independent of k) such that for all $\tau \in [0, 1]$,

$\|\nabla^2 F_{S_k}(\overline{u}^k+\tau(u^k-\overline{u}^k))\| \le \widehat{c}_2$. Note that $\mathrm{dist}(0,\partial F(x^k)) = \|r^k\|$ and

$$
\begin{aligned}
\|r^k\| &= \|r^k-\nabla F_{S_k}(\overline{u}^k) + \nabla F_{S_k}(\overline{u}^k)\| \le \|r^k-\nabla F_{S_k}(\overline{u}^k)\| + \|\nabla F_{S_k}(\overline{u}^k)\| \\
&\le \int_0^1 \|\nabla^2 F_{S_k}(\overline{u}^k+\tau(u^k-\overline{u}^k))(u^k-\overline{u}^k)\|_2 d\tau + \|\nabla F_{S_k}(\overline{u}^k)\| \\
&\le \widehat{c}_2\|u^k-\overline{u}^k\| + \|\nabla F_{S_k}(\overline{u}^k)\| \le \big[(\widehat{L}+\widetilde{L})+\widehat{c}_2\big]\|x^k-\overline{x}^k\|, \qquad (3.36)
\end{aligned}
$$

where the last inequality is due to (3.35). The first part of the conclusions follows. From (3.36), Lemma 3.9 (ii) and (3.25), we obtain the second part. □

To achieve the linear convergence rate of the objective sequence $\{F(x^k)\}_{k\in\mathbb{N}}$, we first argue that for all sufficiently large k, the support of the iterate x^k is stable, and $k\in\mathcal{K}_2$. The latter means that after a finite number of iterates, Algorithm 3 reduces to a regularized Newton method for minimizing the function F_{S_*}, where S_* is defined below in Proposition 3.2 (i).

Proposition 3.2. *Let $\{x^k\}_{k\in\mathbb{N}}$ and $\{\overline{x}^k\}_{k\in\mathbb{N}}$ be the sequences given by Algorithm 3. Then, under Assumption 3.1, the following assertions hold.*

(i) There exists an index set $S_*\subseteq[n]$ such that for all sufficiently large k,

$$\mathrm{supp}(x^k) = \mathrm{supp}(\overline{x}^k) = S_*;$$

furthermore, every cluster point x^* of $\{x^k\}_{k\in\mathbb{N}}$ satisfies $\mathrm{supp}(x^*) = S_*$.

(ii) There exists $\overline{k}\in\mathbb{N}$ such that for all $k\ge\overline{k}$, $k\in\mathcal{K}_2$.

Proof. **(i)** First we argue that $|x^k|_{\min} > \frac{\nu}{2}$ for all sufficiently large k. Indeed, by Lemma 3.6 (i), if $k-1\in\mathcal{K}_1$, i.e., $x^k=\overline{x}^{k-1}$, we have $|x^k|_{\min}\ge\nu$. If $k-1\in\mathcal{K}_2$, we have $|x^{k-1}|_{\min}\ge\nu$, while by Lemma 3.10, for all sufficiently large k, $\|d^{k-1}\|<\frac{\nu}{3}$, which along with $x^k = x^{k-1}+\alpha_k d^k, \alpha_k\in(0,1]$, $d^k_{S_k^c}=0$ and $|x^{k-1}|_{\min}\ge\nu$ implies that $|x^k|_{\min}>\frac{\nu}{2}$. Next we argue that for all sufficiently large k, $\mathrm{supp}(x^k)=\mathrm{supp}(\overline{x}^k)$.

Indeed, by Lemma 3.9 (ii), for all sufficiently large k, $\|x^k - \overline{x}^k\| < \frac{\nu}{3}$. Hence, for every $i \in \mathrm{supp}(x^k)$, we have $|\overline{x}_i^k| \geq |x_i^k| - |x_i^k - \overline{x}_i^k| > \frac{\nu}{2} - \frac{\nu}{3} > 0$, which implies that $\mathrm{supp}(x^k) \subseteq \mathrm{supp}(\overline{x}^k)$; and for every $i \in \mathrm{supp}(\overline{x}^k)$, we have $|x_i^k| > |\overline{x}_i^k| - \frac{\nu}{3} > 0$, which implies that $\mathrm{supp}(\overline{x}^k) \subseteq \mathrm{supp}(x^k)$. Thus, $\mathrm{supp}(\overline{x}^k) = \mathrm{supp}(x^k)$ holds for all k large enough. It remains to show that for all k large enough, $\mathrm{supp}(x^k) = \mathrm{supp}(x^{k+1})$. For all sufficiently large $k \in \mathcal{K}_1$, the conclusion holds since $x^{k+1} = \overline{x}^k$ and $\mathrm{supp}(\overline{x}^k) = \mathrm{supp}(x^k)$. For all sufficiently large $k \in \mathcal{K}_2$, by Lemma 3.10, we have $\|d^k\| < \frac{\nu}{3}$ and then $\|x^{k+1} - x^k\| < \frac{\nu}{3}$, and the conclusion follows by the above arguments. To sum up, $\mathrm{supp}(x^{k+1}) = \mathrm{supp}(x^k) = \mathrm{supp}(\overline{x}^k)$ holds for all sufficiently large k. Since $|x^k|_{\min} > \frac{\nu}{2}$ for all sufficiently large k, following a similar arguments as above we have every cluster point x^* of $\{x^k\}$ satisfies $\mathrm{supp}(x^*) = S_*$.

(ii) By the proof of part (i), we have $|x^k|_{\min} > \frac{\nu}{2}$ for all sufficiently large k. Together with Lemmas 3.9 (ii) and 3.6 (iii), the two conditions in (3.10) are satisfied for all k large enough, so there exists $\overline{k} \in \mathbb{N}$ such that for all $k \geq \overline{k}$, $k \in \mathcal{K}_2$. □

Now we are in a position to achieve the Q-linear convergence rate of the objective value sequence $\{F(x^k)\}_{k\in\mathbb{N}}$ under the KL property of the exponent 1/2 of F.

Proposition 3.3. *Suppose that Assumption 3.1 holds, and that F is a KL function of exponent 1/2. Then $\{F(x^k)\}_{k\in\mathbb{N}}$ converges to some value F^* in a Q-linear rate.*

Proof. If there exists some $k \in \mathbb{N}$ such that $F(x^k) = F(x^{k+1})$, by Lemma 3.9 (i), we have $x^k = \overline{x}^k$, and the stopping condition in (1b) of Algorithm 3 is satisfied, so $\{x^k\}_{k\in\mathbb{N}}$ converges to an L-type stationary point within a finite number of steps. Hence, it suffices to consider that $F(x^k) > F(x^{k+1})$ for all $k \in \mathbb{N}$. Since F is assumed to be a KL function of exponent 1/2, by (Bolte et al., 2014, Lemma 6) and Lemma 3.4 (ii), there exist $\varepsilon > 0$ and $\eta > 0$ such that for all $\overline{x} \in \omega(x^0)$ and all $z \in \{x \in \mathbb{R}^n \,|\, \mathrm{dist}(x, \omega(x^0)) < \varepsilon\} \cap [F(\overline{x}) < F < F(\overline{x}) + \eta]$,

$$\varphi'(F(z) - F(\overline{x}))\mathrm{dist}(0, \partial F(z)) \geq 1, \tag{3.37}$$

where $\varphi(t) = c\sqrt{t}$ for some $c > 0$. Let x^* be a cluster point of $\{x^k\}_{k\in\mathbb{N}}$. Clearly, $\lim_{k\to\infty} \mathrm{dist}(x^k, \omega(x^0)) = 0$. Along with $\lim_{k\to\infty} F(x^k) = F(x^*)$, for all sufficiently large k, $x^k \in \{x \in \mathbb{R}^n \,|\, \mathrm{dist}(x, \omega(x^0)) < \varepsilon\} \cap [F(x^*) < F < F(x^*) + \eta]$, and then

$$\frac{c}{2}(F(x^k) - F(x^*))^{-1/2}\mathrm{dist}(0, \partial F(x^k)) \geq 1.$$

Let $\Delta_k = F(x^k) - F(x^*)$ for each k. By Proposition 3.2 (ii), when $k > \overline{k}$, $k \in \mathcal{K}_2$. Combining the above inequality with Lemma 3.10 yields that for all $k > \overline{k}$ (if necessary by increasing $\overline{k}$),

$$\begin{aligned} 4c^{-2} &\leq \big[(\Delta_k)^{-1/2}\mathrm{dist}(0, \partial F(x^k))\big]^2 \leq \widetilde{c}_2^2(\Delta_k)^{-1}\|x^k - \overline{x}^k\|^2 \\ &\leq 2\widetilde{c}_2^2\widehat{\gamma}^{-1}(\Delta_k)^{-1}[F(x^k) - F(x^{k+1})] = 2\widetilde{c}_2^2\widehat{\gamma}^{-1}(\Delta_k)^{-1}(\Delta_k - \Delta_{k+1}), \end{aligned}$$

where the third inequality is due to Lemma 3.9 (i). The last inequality, along with $0 < \Delta_{k+1} < \Delta_k$ implies that $\rho = 1 - \frac{2\widehat{\gamma}}{(c\widetilde{c}_2)^2} \in (0, 1)$. Then, for all $k \geq \overline{k}$, we have $\Delta_{k+1} \leq \rho\Delta_k$, so that $\{F(x^k)\}_{k\in\mathbb{N}}$ converges to $F^* = F(x^*)$ in a Q-linear rate. □

3.3.2 Convergence Analysis of Iterate Sequence

In order to achieve the convergence of the sequence $\{x^k\}_{k\in\mathbb{N}}$, we also need the following assumption:

Assumption 3.2. *It holds that* $\liminf\limits_{\mathcal{K}_2 \ni k\to\infty} \frac{-\langle r^k, d^k_{S_k}\rangle}{\|r^k\|\|d^k_{S_k}\|} > 0.$

Assumption 3.2 is very common in the global convergence analysis of line search Newton-type methods (see, e.g., Nocedal and Wright (2006)), which essentially requires that the angle between r^k and $d^k_{S_k}$ is sufficiently away from $\pi/2$ and close to π. Note that the early global convergence analysis of Newton-type methods aims to achieve $\lim_{k\to\infty} \|r^k\| = 0$ under Assumption 3.2. Here, under this assumption, we establish the convergence of the whole iterate sequence for the KL function F.

Theorem 3.1. *Suppose Assumptions 3.1 and 3.2 hold. The following assertions hold.*

(i) If F 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-type stationary point of (3.1)*, say x^*.*

(ii) If F is a KL function of exponent $1/2$ at x^*, then $\{x^k\}_{k\in\mathbb{N}}$ converges R-linearly to x^*.

Proof. **(i)** By the proof of Proposition 3.3, it suffices to consider the case where $F(x^k) > F(x^{k+1})$ for all k. Let x^* be a cluster point of $\{x^k\}_{k\in\mathbb{N}}$. Following a similar argument to the proof of Proposition 3.3 and from Definition 2.4, there exists $\varphi \in \Upsilon_\eta$ such that for sufficiently large k,

$$\varphi'(F(x^k) - F(x^*))\text{dist}(0, \partial F(x^k)) \geq 1. \tag{3.38}$$

By Assumption 3.2, there exists $c_{\min} > 0$ such that for all sufficiently large $k \in \mathcal{K}_2$,

$$-\langle r^k, d^k_{S_k}\rangle > c_{\min}\|r^k\|\|d^k_{S_k}\|. \tag{3.39}$$

By Proposition 3.2, there exists $\overline{k} \in \mathbb{N}$ such that for all $k \geq \overline{k}$, $k \in \mathcal{K}_2$ and $S_k = S_{k+1}$. Together with (3.11) and (3.39), if necessary by increasing $\overline{k}$, for all $k \geq \overline{k}$, we have

$$\frac{F(x^k) - F(x^{k+1})}{\|r^k\|} \geq \frac{-\varrho\alpha_k\langle r^k, d^k_{S_k}\rangle}{\|r^k\|} \geq \varrho c_{\min}\|\alpha_k d^k_{S_k}\| = \varrho c_{\min}\|x^{k+1} - x^k\|. \tag{3.40}$$

In addition, from the concavity of φ on $[0, \eta)$, for all $k > \overline{k}$, it holds that

$$\varphi(F(x^k)-F(x^*)) - \varphi(F(x^{k+1})-F(x^*)) \geq \varphi'(F(x^k)-F(x^*))(F(x^k)-F(x^{k+1})). \tag{3.41}$$

For each k, let $\bar{\Delta}_k := \varphi(F(x^k) - F(x^*))$. From (3.38) and (3.40)-(3.41), if possibly enlarging $\overline{k}$, we have for all $k \geq \overline{k}$,

$$\begin{aligned}\bar{\Delta}_k - \bar{\Delta}_{k+1} &\geq \varphi'(F(x^k) - F(x^*))(F(x^k) - F(x^{k+1})) \\ &\geq \frac{F(x^k) - F(x^{k+1})}{\text{dist}(0, \partial F(x^k))} = \frac{F(x^k) - F(x^{k+1})}{\|r^k\|} \geq \varrho c_{\min}\|x^{k+1} - x^k\|.\end{aligned}$$

Summing this inequality from $\overline{k}$ to any $k > \overline{k}$ yields that

$$\sum_{j=\overline{k}}^{k} \|x^{j+1} - x^j\| \le \frac{1}{\varrho c_{\min}} \sum_{j=\overline{k}}^{k} (\bar{\Delta}_j - \bar{\Delta}_{j+1}) = \frac{1}{\varrho c_{\min}} (\bar{\Delta}_{\overline{k}} - \bar{\Delta}_{k+1}) \le \frac{1}{\varrho c_{\min}} \bar{\Delta}_{\overline{k}}.$$

Passing the limit $k \to \infty$ to this inequality yields that $\sum_{j=\overline{k}}^{\infty} \|x^{j+1} - x^j\| < \infty$. Thus the sequence $\{x^k\}$ converges. By Lemma 3.9 (iii), the desired result then follows.

(ii) For each $k \in \mathbb{N}$, write $\Delta_k := F(x^k) - F(x^*)$. From Proposition 3.2 and the proof of Proposition 3.3, there exists $\overline{k}$ such that for all $k > \overline{k}$, $k \in \mathcal{K}_2$ and $\Delta_{k+1} \le \rho \Delta_k$. From this recursion formula,

$$F(x^k) - F(x^*) \le \Delta_{\overline{k}} \rho^{k - \overline{k}}. \tag{3.42}$$

By (3.25) and Lemma 3.10, for all $k \ge \overline{k}$, $\|d^k\| \le b_2^{-1} \widetilde{c}_2^{1-\sigma} \|x^k - \overline{x}^k\|^{1-\sigma}$. Together with part (i), Lemma 3.9 (i) and (3.42), for all $k \ge \overline{k}$ it holds that

$$\begin{aligned}
\|x^k - x^*\| &\le \sum_{j=k}^{\infty} \|x^j - x^{j+1}\| = \sum_{j=k}^{\infty} \alpha_j \|d^j\| \le \sum_{j=k}^{\infty} \|d^j\| \le b_2^{-1} \widetilde{c}_2^{1-\sigma} \sum_{j=k}^{\infty} \|x^j - \overline{x}^j\|^{1-\sigma} \\
&\le b_2^{-1} \widetilde{c}_2^{1-\sigma} \sum_{j=k}^{\infty} \Big(\frac{2(F(x^j) - F(x^{j+1}))}{\widehat{\gamma}} \Big)^{\frac{1-\sigma}{2}} \\
&\le b_2^{-1} \widetilde{c}_2^{1-\sigma} \Big(\frac{2\Delta_{\overline{k}}}{\widehat{\gamma} \rho^{\overline{k}}} \Big)^{\frac{1-\sigma}{2}} \sum_{j=k}^{\infty} \rho^{\frac{(1-\sigma)j}{2}} \le \Big(\frac{2\Delta_{\overline{k}}}{\widehat{\gamma} \rho^{\overline{k}}} \Big)^{\frac{1-\sigma}{2}} \frac{\widetilde{c}_2^{1-\sigma}}{b_2(1 - \rho^{1/4})} \rho^{k/4}.
\end{aligned}$$

This means that the sequence $\{x^k\}_{k \in \mathbb{N}}$ converges to x^* in an R-linear rate. $\square$

By Proposition 3.1, to check the KL property with exponent 1/2 of F at x^*, it suffices to verify that of F_{S_*} at $x^*_{S_*}$, and due to the sufficient smoothness of F_{S_*} at $x^*_{S_*}$, the verification of the latter is easier than that of the former. In fact, by (Zeng et al., 2016, Lemma 3), the nonsingularity of $\nabla^2 F_{S_*}(x^*_{S_*})$ implies the KL property of exponent 1/2 for F_{S_*} at $x^*_{S_*}$.

By Theorem 3.1, if Assumptions 3.1-3.2 hold and F is a KL function, the sequence $\{x^k\}_{k\in\mathbb{N}}$ is convergent. In the sequel, we denote its limit by x^*. By Proposition 3.2 (i), $\text{supp}(x^*) = S_*$. Write

$$u^* := x^*_{S_*} \text{ and } \mathcal{U}^* := \big\{u \in \mathbb{R}^{|S_*|} \mid \nabla F_{S_*}(u) = 0, \nabla^2 F_{S_*}(u) \succeq 0\big\}.$$

To achieve the superlinear convergence rate of $\{x^k\}_{k\in\mathbb{N}}$, we need to bound Λ_k involved in the matrix G_k by $\text{dist}(u^k,\mathcal{U}^*)$ as in the following lemma.

Lemma 3.11. *Suppose that Assumptions 3.1 and 3.2 hold, and that F is a KL function. If $\nabla^2 F_{S_*}(u^*) \succeq 0$, then there exists $c_H > 0$ such that for all sufficiently large k, $\Lambda_k \le c_H \text{dist}(u^k,\mathcal{U}^*)$.*

Proof. By the proof of Proposition 3.2 (i), we have $|x^*|_{\min} \ge \frac{\nu}{2}$. Fix any $\varepsilon < \frac{\nu}{4}$. From Proposition 3.2 and Theorem 3.1 (i), if necessary enlarging $\overline{k}$, we have for all $k > \overline{k}$, $k \in \mathcal{K}_2$, $S_k = S_*$ and $u^k \in \mathbb{B}(u^*, \varepsilon/2)$. By following the proof of Lemma 3.8, there exists $c_H > 0$ such that for any $u', u'' \in \mathbb{B}(u^*, \varepsilon)$,

$$\|\nabla^2 F_{S_*}(u') - \nabla^2 F_{S_*}(u'')\|_2 \le c_H \|u' - u''\|. \tag{3.43}$$

Fix any $k > \overline{k}$. When $\lambda_{\min}(\nabla^2 F_{S_k}(u^k)) > 0$, the desired result is trivial, so it suffices to consider the case $\lambda_{\min}(\nabla^2 F_{S_k}(u^k)) \le 0$. Pick any $\widetilde{u}^k \in \text{proj}_{\mathcal{U}^*}(u^k)$. Since $u^* \in \mathcal{U}^*$, one can deduce that $\|\widetilde{u}^k - u^*\| \le \|\widetilde{u}^k - u^k\| + \|u^k - u^*\| \le 2\|u^k - u^*\| \le \varepsilon$. If $\lambda_{\min}(\nabla^2 F_{S_k}(\widetilde{u}^k)) = 0$, then by Weyl's inequality (Bhatia, 2013, Corollary III.2.6) we have $\Lambda_k = -\lambda_{\min}(\nabla^2 F_{S_k}(u^k)) \le \|\nabla^2 F_{S_k}(\widetilde{u}^k) - \nabla^2 F_{S_k}(u^k)\|_2$, which together with (3.43) implies that $\Lambda_k \le c_H \|u^k - \widetilde{u}^k\| = c_H \text{dist}(u^k,\mathcal{U}^*)$. Now suppose that $\lambda_{\min}(\nabla^2 F_{S_k}(\widetilde{u}^k)) > 0$. Let $\phi_k(t) := \lambda_{\min}[\nabla^2 F_{S_k}(u^k + t(\widetilde{u}^k - u^k))]$ for $t \ge 0$. Clearly, ϕ_k is continuous on an open interval containing $[0, 1]$. Note that $\phi_k(0) < 0$ and $\phi_k(1) > 0$. There necessarily exists $\overline{t}_k \in (0, 1)$ such that $\phi_k(\overline{t}_k) = 0$. Consequently, by Weyl's inequality,

$$\begin{aligned}\Lambda_k &= \big[\lambda_{\min}(\nabla^2 F_{S_k}(u^k + \overline{t}_k(\widetilde{u}^k - u^k))) - \lambda_{\min}(\nabla^2 F_{S_k}(u^k))\big] \\ &\le \|\nabla^2 F_{S_k}(u^k + \overline{t}_k(\widetilde{u}^k - u^k)) - \nabla^2 F_{S_k}(u^k)\|_2 \le c_H \|\widetilde{u}^k - u^k\|.\end{aligned}$$

This shows that the desired result holds. The proof is completed. □

Ueda and Yamashita ever obtained a similar result in (Ueda and Yamashita, 2010, Lemma 5.2) under the condition that $\mathcal{U}^*$ is the set of local minima of F_{S_*}. Here, we remove the local optimality of $\mathcal{U}^*$ and provide a simpler proof. Based on this result, we establish the superlinear convergence rate of $\{x^k\}_{k\in\mathbb{N}}$ under a local error bound condition.

Theorem 3.2. *Suppose that Assumptions 3.1 and 3.2 hold, and that F is a KL function. If $\nabla^2 F_{S_*}(u^*) \succeq 0$ and there exist $\delta > 0$ and $\kappa_1 > 0$ such that for all $u \in \mathbb{B}(u^*, \delta)$,*

$$\kappa_1 \mathrm{dist}(u, \mathcal{U}^*) \leq \|\nabla F_{S_*}(u)\|, \tag{3.44}$$

then the sequence $\{x^k\}_{k\in\mathbb{N}}$ converges to x^* in a Q-superlinear rate of order $1+\sigma$.

Proof. By Theorem 3.1 and Proposition 3.2, if necessary enlarging $\overline{k}$, wer have for all $k \geq \overline{k}$, $k \in \mathcal{K}_2$ and $S_k = S_*$. By comparing the iterate steps of Algorithm 3 for $k \geq \overline{k}$ with those of E-RNM proposed in (Ueda and Yamashita (2010)), we conclude that the sequence $\{u^k\}_{k\geq\overline{k}}$ is the same as the one generated by E-RNM of (Ueda and Yamashita (2010)). By Lemma 3.11, there exists a constant $c_H > 0$ such that for all $k \geq \overline{k}$ (if necessary by increasing $\overline{k}$), $\Lambda_k \leq c_H \mathrm{dist}(u^k, \mathcal{U}^*)$. Then, by (Ueda and Yamashita, 2010, Theorem 5.1) $\mathrm{dist}(u^k, \mathcal{U}^*)$ converges to 0 superlinearly with rate $1+\sigma$.

Write $\mathcal{X}^* := \{x \in \mathbb{R}^n \,|\, x_{S_*} \in \mathcal{U}^*, x_{S_*^c} = 0\}$. For all $k \geq \overline{k}$, from $S_k = S_*$, clearly, $\mathrm{dist}(x^k, \mathcal{X}^*) = \mathrm{dist}(u^k, \mathcal{U}^*)$. Consequently, $\mathrm{dist}(x^k, \mathcal{X}^*)$ converges to 0 superlinearly with rate $1+\sigma$, i.e., for all $k \geq \overline{k}$ (if necessary by enlarging $\overline{k}$),

$$\mathrm{dist}(x^{k+1}, \mathcal{X}^*) = O([\mathrm{dist}(x^k, \mathcal{X}^*)]^{1+\sigma}). \tag{3.45}$$

Also, by (Ueda and Yamashita, 2010, Lemma 5.3) there exists a constant $c_0 > 0$ such that for all $k \geq \overline{k}$ (if necessary by increasing $\overline{k}$),

$$\|d^k_{S_k}\| = \|d^k_{S_*}\| \leq c_0 \text{dist}(u^k, \mathcal{U}^*) = c_0 \text{dist}(x^k, \mathcal{X}^*). \tag{3.46}$$

For each $k \geq \overline{k}$, pick any $\widetilde{x}^k \in \text{proj}_{\mathcal{X}^*}(x^k)$. By the definition of $\mathcal{X}^*$, $\text{supp}(\widetilde{x}^k) \subseteq S_*$; while from $\lim_{k\to\infty} x^k = x^*$, we have $\text{supp}(\widetilde{x}^k) \supseteq S_*$ for all $k \geq \overline{k}$ (if necessary by increasing $\overline{k}$). Then, for all $k \geq \overline{k}$, $\text{supp}(\widetilde{x}^k) = S_*$. In addition, by (3.45) there exists $\rho \in (0, 1)$ such that $\text{dist}(x^{k+1}, \mathcal{X}^*) \leq \rho \text{dist}(x^k, \mathcal{X}^*)$ for all $k > \overline{k}$. Together with (3.46), for all $k \geq \overline{k}$ it holds that

$$\begin{aligned}\|x^k - x^*\| &\leq \sum_{j=k}^{\infty} \|x^j - x^{j+1}\| \leq \sum_{j=k}^{\infty} \|d^j_{S_j}\| \leq c_0 \sum_{j=k}^{\infty} \text{dist}(x^j, \mathcal{X}^*) \\ &< c_0 \Big(\sum_{j=k}^{\infty} \rho^{j-k} \Big) \text{dist}(x^k, \mathcal{X}^*) = \frac{c_0}{1-\rho} \text{dist}(x^k, \mathcal{X}^*).\end{aligned}$$

By combining this inequality and (3.45), it follows that for all $k > \overline{k}$,

$$\|x^k - x^*\| \leq \frac{c_0}{1-\rho} \text{dist}(x^k, \mathcal{X}^*) = O([\text{dist}(x^{k-1}, \mathcal{X}^*)]^{1+\sigma}) \leq O(\|x^{k-1} - x^*\|^{1+\sigma}).$$

The desired conclusion then follows. The proof is completed. □

Remark 3.2. **(a)** *Note that we do not require the isolatedness of u^* and its local optimality. The local error bound condition* (3.44) *is a little stronger than the metric subregularity of ∇F_{S_*} at u^* for the origin because $\mathcal{U}^*$ may be a strict subset of $\nabla F_{S_*}^{-1}(0)$. For example, let $h(t) := \frac{1}{2}(t - 3.5)^2 + 5\sqrt{|t|}$. Elementary calculation yields that $h'(1) = 0$ and $h''(1) = -0.25$. It is clear that h' is metrically subregular at $t = 1$, while the local error bound condition* (3.44) *does not hold at $t = 1$ since $t = 1$ is not a local minimum of h.*

(b) *The proof of the superlinear convergence of E-RNM in (Ueda and Yamashita (2010)) relies on Assumption 5.1 therein, which requires the local optimality of x^*. After checking its proof, we found that the local optimality of x^* was only used to* *achieve (Ueda and Yamashita, 2010, Lemma 5.2). Thus, by following the same arguments as those for Lemma 3.11, the local optimality of x^* in their Assumption 5.1 can be removed.*

To conclude this section, we take a closer look at Assumption 3.2. The following lemma shows that if the regularized Newton direction d^k from Step 2 satisfies condition (3.16) for all $k \in \mathcal{K}_2$, Assumption 3.2 necessarily holds. Together with Example 3.1 later, we conclude that Assumption 3.2 is weaker than condition (3.16) for our regularized Newton direction d^k.

Lemma 3.12. *Suppose that Assumption 3.1 holds. If d^k yielded by Step 2 of Algorithm 3 satisfies condition* (3.16) *for all $k \in \mathcal{K}_2$, then Assumption 3.2 holds.*

Proof. By Lemma 3.4, $\{x^k\}_{k\in\mathbb{N}}$ is bounded. Let x^* be an arbitrary accumulation point of $\{x^k\}_{k\in\mathbb{N}}$. Then, there exists a subsequence $\{x^{k_j}\}_{j\in\mathbb{N}}$ with $k_j \in \mathcal{K}_2$ such that $\lim_{j\to\infty} x^{k_j} = x^*$. By Proposition 3.2, for all sufficiently large $j \in \mathbb{N}$, $\mathrm{supp}(x^{k_j}) = \mathrm{supp}(x^*) = S_*$. Write $s = |S_*|$. From the continuity, the sequence $\{G^{k_j}\}_{j\in\mathbb{N}}$ is convergent and let $G^* = \lim_{j\to\infty} G^{k_j}$. Clearly, G^* is an $s \times s$ positive semidefinite matrix. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_s \geq 0$ be the eigenvalues of G^*. For each $j \in \mathbb{N}$, let $\lambda_1^{k_j} \geq \lambda_2^{k_j} \geq \cdots \geq \lambda_s^{k_j} > 0$ be the eigenvalues of the $s \times s$ positive definite matrix G^{k_j}. Then, for each $i \in [s]$, $\lim_{j\to\infty} \lambda_i^{k_j} = \lambda_i$.

Case 1: $\lambda_s > 0$. Now the matrix G^* is positive definite. Also, for all sufficiently large $j \in \mathbb{N}$, $\lambda_s^{k_j} > \frac{\lambda_s}{2}$ and $0 < \lambda_1^{k_j} \leq \frac{3\lambda_1}{2}$. Consequently, for all sufficiently large $j \in \mathbb{N}$,

$$\frac{-\langle r^{k_j}, d^{k_j}_{S_{k_j}}\rangle}{\|r^{k_j}\|\|d^{k_j}_{S_{k_j}}\|} = \frac{\langle G^{k_j} d^{k_j}_{S_{k_j}}, d^{k_j}_{S_{k_j}}\rangle}{\|G^{k_j} d^{k_j}_{S_{k_j}}\|\|d^{k_j}_{S_{k_j}}\|} \geq \frac{\lambda_s^{k_j}\|d^{k_j}_{S_{k_j}}\|^2}{\lambda_1^{k_j}\|d^{k_j}_{S_{k_j}}\|^2} \geq \frac{\lambda_s}{3\lambda_1} > 0. \tag{3.47}$$

Case 2: $\lambda_s = 0$. Now there exists $t \in [s]$ such that $\lambda_i = 0$ for $i \in [t\!:\!s]$ and $\lambda_i > 0$ for $i \in [t-1]$. Fix any $0 < \varepsilon < \min\left\{\frac{\varpi}{8\hat{c}}, \frac{\varpi}{4\hat{c}\sqrt{s-t+1}}\right\}$. From $\lim_{j\to\infty} \lambda_i^{k_j} = \lambda_i$ for each

$i \in [s]$ and $G^{k_j} \succ 0$ for each $j \in \mathbb{N}$, for all sufficiently large $j \in \mathbb{N}$,

$$0 < \lambda_i^{k_j} < \varepsilon \ \text{ for } i \in [t\!:\!s] \ \text{ and } \ \frac{1}{2}\lambda_i < \lambda_i^{k_j} < \frac{3}{2}\lambda_i \ \text{ for } i \in [t-1]. \tag{3.48}$$

We claim that $t > 1$. If not, $t = 1$, by Lemma 3.7 (ii), $\|d^{k_j}\| = \|(G^{k_j})^{-1}r^{k_j}\| \geq \frac{\|r^{k_j}\|}{\lambda_1^{k_j}} \geq \frac{\varpi}{4\lambda_1^{k_j}}\|x^{k_j} - \overline{x}^{k_j}\| \geq \frac{\varpi}{4\varepsilon}\|x^{k_j} - \overline{x}^{k_j}\|$, which along with $\varepsilon \leq \frac{\varpi}{8\widehat{c}}$ implies that $\|d^{k_j}\| > 2\widehat{c}\|x^{k_j} - \overline{x}^{k_j}\|$, a contradiction to condition (3.16). Now let G^{k_j} have the eigenvalue decomposition given by $G^{k_j} = (V^{k_j})^\top \mathrm{diag}(\lambda_1^{k_j}, \ldots, \lambda_s^{k_j})V^{k_j}$, where V^{k_j} is an $s \times s$ orthogonal matrix. For each $j \in \mathbb{N}$, since the column vectors $v_1^{k_j}, \ldots, v_s^{k_j}$ of the matrix V^{k_j} are linearly independent, there exist $\gamma_1^{k_j}, \ldots, \gamma_s^{k_j} \in \mathbb{R}$ such that

$$\frac{r^{k_j}}{\|r^{k_j}\|} = \sum_{i=1}^{s} \gamma_i^{k_j} v_i^{k_j} \ \text{ with } \ \sum_{i=1}^{s} (\gamma_i^{k_j})^2 = 1. \tag{3.49}$$

Together with the definition of $d_{S_{k_j}}^{k_j}$, it follows that

$$\frac{d_{S_{k_j}}^{k_j}}{\|r^{k_j}\|} = \frac{-(G^{k_j})^{-1}r^{k_j}}{\|r^{k_j}\|} = -\sum_{i=1}^{s} \frac{\gamma_i^{k_j}}{\lambda_i^{k_j}} v_i^{k_j}. \tag{3.50}$$

By combining condition (3.16) and Lemma 3.7 (ii), for each $j \in \mathbb{N}$, we have

$$\|d_{S_{k_j}}^{k_j}\| = \|d^{k_j}\| \leq (4\widehat{c}/\varpi)\|r^{k_j}\|, \tag{3.51}$$

which by (3.50) means that $\sum_{i=1}^{s} \left(\gamma_i^{k_j}/\lambda_i^{k_j}\right)^2 \leq \frac{16\widehat{c}^2}{\varpi^2}$. This by (3.48) implies that for all sufficiently large $j \in \mathbb{N}$, $\gamma_i^{k_j} \leq \frac{4\varepsilon\widehat{c}}{\varpi}$ with $i \in [t\!:\!s]$. Together with $\sum_{i=1}^{s}(\gamma_i^{k_j})^2 = 1$, we obtain that $\sum_{i=1}^{t-1}(\gamma_i^{k_j})^2 \geq 1 - \frac{16(s-t+1)\varepsilon^2\widehat{c}^2}{\varpi^2}$ and then for all sufficiently large $j \in \mathbb{N}$, there exists $l_j \in [t-1]$ such that $(\gamma_{l_j}^{k_j})^2 \geq \frac{\varpi^2 - 16(s-t+1)\varepsilon^2\widehat{c}^2}{\varpi^2(t-1)}$. Thus, for all sufficiently large $j \in \mathbb{N}$, it follows from (3.49)-(3.51) that

$$\frac{-\langle r^{k_j}, d_{S_{k_j}}^{k_j}\rangle}{\|r^{k_j}\|\|d_{S_{k_j}}^{k_j}\|} \geq \frac{\varpi}{4\widehat{c}} \sum_{i=1}^{s} \frac{(\gamma_i^{k_j})^2}{\lambda_i^{k_j}} \geq \frac{\varpi(\gamma_{l_j}^{k_j})^2}{4\widehat{c}\lambda_{l_j}^{k_j}} \geq \frac{\varpi^2 - 16(s-t+1)\varepsilon^2\widehat{c}^2}{6\widehat{c}\lambda_{l_j}\varpi(t-1)} > 0,$$

where the third inequality is also using $\lambda_l^{k_j} \leq \frac{3}{2}\lambda_l$ by (3.48), and the last one is by $0 < \varepsilon < \frac{\varpi}{4\widehat{c}\sqrt{s-t+1}}$. From the last inequality and (3.47), we obtain the conclusion. □

The example below shows that the inverse of Lemma 3.12 does not hold.

Example 3.1. *Consider the problem* $\min_{t\in\mathbb{R}} f(t) + |t|^{\frac{1}{2}}$ *with* f *defined as follows:*

$$f(t) := \begin{cases} \frac{49}{8}t^2 - \frac{67}{4}t + \frac{85}{8} & \text{if } t \in (-\infty, 1), \\ (t-2)^4 - t^{\frac{1}{2}} & \text{if } t \in [1, 4), \\ \frac{1537}{64}t^2 - \frac{5132}{32}t + \frac{1085}{4} & \text{if } t \in [4, \infty). \end{cases}$$

We use Algorithm 3 with $\tau = 2$, $\alpha = 1, \mu_{\min} = 40, \widetilde{L} = 49$ *and* $\sigma = \frac{1}{3}, b_2 = 1, \varrho = 10^{-4}, \beta = \frac{1}{2}, t^0 = 2.1$ *to seek a critical point of this problem. From the iterates of Algorithm 3, the generated sequence* $\{t^k\}$ *satisfies* $\lim_{k\to\infty} t^k = 2$. *When* t^k *is sufficiently close to 2, all the iterates are from regularized Newton step and* $|d^k| = \left|\frac{4(t^k-2)^3}{12(t^k-2)^2+4^{\frac{1}{3}}(t^k-2)}\right| = O(|t^k-2|^2)$, *while by Lemmas 3.7 (ii) and 3.10 we have* $|t_k - \bar{t}_k| = O(|f'(t_k)|) = O(|t_k-2|^3)$. *Then,* $|t^k - \bar{t}^k| = o(|d^k|)$ *and the condition in* (3.16) *does not hold for all sufficiently large* k. *However, Assumption 3.2 always holds because* $-\frac{d^k f'(t^k)}{|d^k||f'(t^k)|} = 1$ *for all* k.

3.4 Numerical Experiments

In this section we apply HpgSRN to solving the ℓ_q quasi-norm regularized linear and logistic regression problems on real data, which respectively take the form of (3.1) with $f = f_1$ or f_2, where $f_1(x) := \frac{1}{2}\|Ax-b\|^2$ and $f_2(x) := \sum_{i=1}^m \log\big(1+\exp(-b_i(Ax)_i)\big)$ for $x \in \mathbb{R}^n$. Here, $A \in \mathbb{R}^{m\times n}$ is a given matrix and $b \in \mathbb{R}^m$ is a given vector. Clearly, f satisfies Assumption 3.1, and from Proposition 2.3 we have that equipped with either f_1 or f_2, F is a KL function. All numerical tests are conducted on a desktop running in MATLAB R2020b and 64-bit Windows System with an Intel(R) Core(TM) i7-

10700 CPU 2.90GHz and 32.0 GB RAM. The MATLAB code is available at `https://github.com/yuqiawu/HpgSRN`.

3.4.1 Implementation of HpgSRN

In Algorithm 3, we set $\mu_0 = 1$ and when $k \geq 1$, μ_k is chosen by the Barzilai-Borwein (BB) rule (Barzilai and Borwein (1988)), that is,

$$\mu_k = \max\left\{\mu_{\min}, \min\left\{\mu_{\max}, \frac{\langle x^k - x^{k-1}, \nabla f(x^k) - \nabla f(x^{k-1})\rangle}{\|x^k - x^{k-1}\|^2}\right\}\right\}$$

with $\mu_{\min} = 10^{-20}$ and $\mu_{\max} = 10^{20}$. For each $k \in \mathcal{K}_2$, we call the MATLAB function `eigs` to compute the approximate smallest eigenvalue of $\nabla^2 F_{S_k}(u^k)$, which requires about $O(|S_k|^2)$ flops by Stewart (2002). Since $|S_k|$ is usually much smaller than n, this computation cost is not expensive. In addition, we choose

$$\tau = 10,\ \alpha = 10^{-8},\ \sigma = 0.5,\ b_1 = 1 + 10^{-8},\ b_2 = 10^{-3},\ \varrho = 10^{-4}, \beta = 2.$$

During the testing, we solve the linear system in (2a) via a direct method if $|S_k| < 500$, otherwise a conjugate gradient method. The direct method for computing the inverse of the G^k needs about $O(|S_k|^3)$ flops, so that HpgSRN is well adapted to high dimensional problems if $|S_k|$ is small. Our preliminary tests indicate that (3.1) with $q = 1/2$ usually has better performance than (3.1) with other $q \in (0, 1)$ in terms of the CPU time and the sparsity. This coincides with the conclusion in (Hu et al. (2017); Xu et al. (2010)). Inspired by this, we choose $q = 1/2$ for the subsequent numerical testing. The parameter λ in (3.1) is specified in the corresponding experiments.

We compare the performance of HpgSRN with that of ZeroFPR (Themelis et al. (2018)). The code package of ZeroFPR is downloaded from `http://github.com/kul-forbes/ForBES`. Consider that the iterate steps of PG method with a monotone line search (PGls), a monotone version of SpaRSA (Wright et al. (2009)), are the same as those of step (1a) of Algorithm 3 with the above BB rule for updating μ_k.

We also compare the performance of HpgSRN with that of PGls to check the effect of the additional subspace regularized Newton step on HpgSRN. The parameters of PGls are chosen to be the same as those involved in Step 1 of HpgSRN except $\tau = 2$. For the three algorithms, we adopt the same stopping criterion

$$\gamma\|x^k - \text{prox}_{\gamma^{-1}g}(x^k - \gamma^{-1}\nabla f(x^k))\|_\infty < 10^{-3} \quad \text{or} \quad k \geq 50000,$$

where $\gamma = L/0.95$ and L is an estimation of the Lipschitz constant of $\nabla f(\cdot)$. It is well known that the Lipschitz constants of ∇f_1 and ∇f_2 are $\|A\|_2^2$ and $0.25\|A\|_2^2$, respectively. We use the following MATLAB code to estimate the spectral norm of A:

```
Amap = @(x) A*x; ATmap = @(x) A'*x; AATmap = @(x) Amap(ATmap(x));
eigsopt.issym = 1; L = eigs(AATmap, m, 1, 'LA', eigsopt).
```

As in ZeroFPR, we choose $x^0 = 0$ as the starting point. Although $x^0 = 0$ is a local minimizer of F and hence an L-type stationary point by (Ahookhosh et al., 2021, Theorem 4.4), it is not a good one in terms of objective value; see the difference between $F(0)$ and Fval, the objective value of the output, for each example in Tables 3.1 and 3.2. It is worth noting that equipped with such an initial point, Algorithm 3 may stop in the first iteration and in this case, x^0 is regarded as an acceptable solution.

In the next two subsections, we will conduct the experiments on real data and report the numerical results including the number of iterations (Iter#), the CPU times in seconds (Time), the objective function values (Fval) and the cardinality of the outputs (Nnz). For the reason that ℓ_q quasi-norm is not a soft thresholding, we simply calculate Nnz of x by MATLAB sentence `Nnz = sum(abs(x)>0)`. In particular, to check the effect of the regularized Newton steps in HpgSRN, we record its number of iterations in the form $M(N)$, where M means the total number of iterates and N means the number of regularized Newton steps.

3.4.2 ℓ_q Regularized Linear Regression

We conduct the experiments for the ℓ_q quasi-norm regularized linear regressions with (A, b) from LIBSVM datasets (see `https://www.csie.ntu.edu.tw`). As suggested in (Huang et al. (2010)), for **housing** and **space_ga**, we expand their original features with polynomial basis functions. The second column of Table 3.1 lists the values of $\|A\|_2^2$ and $F(0)$. Among others, large $\|A\|_2^2$ leads to a difficult implementation of PG method. In fact, the step length of PG is related to the inverse of Lipschitz constant of f, generally propotional to $\|A\|_2^2$. Therefore, the larger the $\|A\|_2^2$, the smaller the step length, making the problem more difficult to solve. On the other hand, the term $F(0)$ reflects the quality of the starting point x^0. For each dataset, we solve (3.1) associated to f_1 and $\lambda = \lambda_c\|A^\top b\|_\infty$ for two different λ_c's with the three solvers.

From Table 3.1, we see that for all test examples HpgSRN spends much less time than ZeroFPR and PGls. For example, for **log1p.E2006.train** with $\lambda_c = 10^{-5}$, ZeroFPR and PGls require more than one hour to yield an output, but HpgSRN returns an output within only $314s$. In terms of the objective function value and sparsity, the outputs of HpgSRN are comparable with those of ZeroFPR and PGls, and even in some examples, these outputs of HpgSRN are better. For example, for **housing7** with both λ_c's the objective function values of HpgSRN are better than those of ZeroFPR and PGls as well as the sparsity of HpgSRN is much less.

3.4.3 ℓ_q Regularized Logistic Regression

We conduct the experiments for the ℓ_q quasi-norm regularized logistic regressions with (A, b) from LIBSVM datasets. For each data, we solve (3.1) associated to f_2 and $\lambda = \lambda_c \max_{1\le j\le n} \|A_{\cdot j}\|_1$ for two different λ_c's with the three solvers. Table 3.2 records their numerical results. We see that in terms of CPU time, HpgSRN is still the best one among the three solvers; in terms of the quality of the other outputs,

Table 3.1: Numerical comparisons on ℓ_q regularized linear regressions with LIBSVM datasets

Data (m, n)	$\|A\|_2^2$ $F(0)$	λ_c	Index	HpgSRN	ZeroFPR	PGls
space_ga9 (3107, 5505)	4.01e3 5.77e3	10^{-3}	Iter#	17(5)	43	180
			Time	0.45	0.98	0.93
			Fval	36.47	37.24	37.15
			Nnz	7	7	6
		10^{-4}	Iter#	230(64)	476	3058
			Time	2.26	9.03	16.48
			Fval	20.93	20.31	21.57
			Nnz	15	19	15
housing7 (506, 77520)	3.28e5 1.50e5	10^{-3}	Iter#	639(157)	4164	25133
			Time	14.45	2.13e2	4.08e2
			Fval	2.25e3	2.57e3	2.56e3
			Nnz	27	49	57
		10^{-4}	Iter#	1765(485)	18807	50000
			Time	49.26	9.81e2	8.59e2
			Fval	8.89e2	9.27e2	9.17e2
			Nnz	82	123	135
E2006.test (3308, 72812)	4.79e4 2.46e4	10^{-4}	Iter#	3(0)	3	3
			Time	0.03	0.25	0.03
			Fval	2.45e2	2.45e2	2.45e2
			Nnz	1	1	1
		10^{-5}	Iter#	3(0)	4	4
			Time	0.05	0.25	0.04
			Fval	2.40e2	2.40e2	2.40e2
			Nnz	1	1	1
E2006.train (16087, 150348)	1.91e5 1.03e5	10^{-4}	Iter#	3(0)	3	3
			Time	0.09	1.06	0.09
			Fval	1.22e3	1.22e3	1.22e3
			Nnz	1	1	1
		10^{-5}	Iter#	4(0)	4	4
			Time	0.11	1.05	0.11
			Fval	1.20e3	1.20e3	1.20e3
			Nnz	1	1	1
log1p.E2006.test (3308, 1771946)	1.46e7 2.46e4	10^{-4}	Iter#	372(88)	827	1416
			Time	33.54	2.87e2	1.16e2
			Fval	2.35e2	2.43e2	2.37e2
			Nnz	5	4	6
		10^{-5}	Iter#	755(166)	6708	22305
			Time	1.01e2	2.28e3	2.30e3
			Fval	1.54e2	1.53e2	1.49e2
			Nnz	385	460	389
log1p.E2006.train (16087, 4265669)	5.86e7 1.03e5	10^{-4}	Iter#	286(58)	855	1621
			Time	77.95	8.57e2	3.85e2
			Fval	1.16e3	1.16e3	1.16e3
			Nnz	7	5	4
		10^{-5}	Iter#	944(195)	5610	33112
			Time	3.14e2	5.26e3	8.83e3
			Fval	1.02e3	1.02e3	1.01e3
			Nnz	141	184	155

HpgSRN has a comparable performance with ZeroFPR and PGls.

To sum up, HpgSRN requires the least CPU time for all the test examples com-

Table 3.2: Numerical comparisons on ℓ_q regularized logistic regressions with LIBSVM datasets

Data (m,n)	$\|A\|_2^2$ $F(0)$	λ_c	Index	HpgSRN	ZeroFPR	PGls
colon-cancer $(62, 2000)$	1.94e4 42.98	10^{-2}	Iter#	48(6)	730	94
			Time	0.04	0.74	0.06
			Fval	7.97	10.58	7.77
			Nnz	10	9	9
		10^{-3}	Iter#	94(9)	1853	175
			Time	0.07	2.07	0.11
			Fval	1.03	1.07	1.07
			Nnz	11	12	12
rcv1 $(20242, 47236)$	4.48e2 1.40e4	10^{-2}	Iter#	65(10)	448	1193
			Time	1.00	6.35	11.24
			Fval	4.23e3	4.35e3	4.24e3
			Nnz	165	167	164
		10^{-3}	Iter#	365(96)	2081	5536
			Time	7.78	29.27	88.65
			Fval	1.28e3	1.53e3	1.27e3
			Nnz	704	741	717
news20 $(19996, 1355191)$	1.73e3 1.39e4	10^{-2}	Iter#	44(6)	170	981
			Time	2.65	36.61	53.14
			Fval	9.73e3	1.04e4	9.53e3
			Nnz	51	42	50
		10^{-3}	Iter#	410(99)	1528	18538
			Time	41.45	3.44e2	1.43e3
			Fval	4.31e3	4.71e3	4.25e3
			Nnz	385	371	401

pared to ZeroFPR and PGls, and for those large scale examples, HpgSRN is at least ten times faster than ZeroFPR and PGls. The outputs of the objective function value and the sparsity yielded by HpgSRN have a comparable even better quality. This indicates that the introduction of Newton steps improves greatly the performance of the PG method. We also observe that for most of examples, the iterates generated by the regularized Newton step account for about 10%–35% of the total iterates.

Chapter 4

An Inexact Regularized Projected Newton Method for Fused ℓ_0-norms Regularized Problems

Given a matrix $B \in \mathbb{R}^{p\times n}$, $\lambda_1 > 0$, $\lambda_2 > 0$, $l_b \in \mathbb{R}^n_-$ and $u_b \in \mathbb{R}^n_+$, in this chapter we consider the following structured ℓ_0-norms regularization problem with a box constraint:

$$\min_{x\in\mathbb{R}^n} F(x) := f(x) + \lambda_1\|Bx\|_0 + \lambda_2\|x\|_0 \ \text{ s.t. } \ l_b \le x \le u_b, \tag{4.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function, $\|\cdot\|_0$ denotes the ℓ_0-norm (or cardinality) function. This model encourages sparsity of both variable x and its linear transformation Bx. Throughout this chapter, we define

$$g(\cdot) := \lambda_1\|B\cdot\|_0 + \lambda_2\|\cdot\|_0 + \delta_\Omega(\cdot) \ \text{ and } \ \Omega := \{x \in \mathbb{R}^n \mid l_b \le x \le u_b\},$$

where $\delta_\Omega(\cdot)$ denotes the indicator function of Ω.

In this chapter, we aim to design a hybrid of PG and inexact projected regularized Newton methods (PGiPN) to solve the structured ℓ_0-norms regularization problem (4.1), whose main idea is similar to that of HpgSRN in Chapter 3. In particular, let $x^k \in \Omega$ be the current iterate. Our method first runs a PG step with line search at

x^k to produce $\overline{x}^k$ with

$$\overline{x}^k \in \operatorname{prox}_{\overline{\mu}_k^{-1} g}(x^k - \overline{\mu}_k^{-1} \nabla f(x^k)), \tag{4.2}$$

where $\overline{\mu}_k > 0$ is a constant such that F gains a sufficient decrease from x^k to $\overline{x}^k$, and then judges whether the iterate enters Newton step or not in terms of some switch condition, which takes the following forms of structured stable supports:

$$\operatorname{supp}(x^k) = \operatorname{supp}(\overline{x}^k) \;\text{ and }\; \operatorname{supp}(Bx^k) = \operatorname{supp}(B\overline{x}^k). \tag{4.3}$$

If this switch condition does not hold, we set $x^{k+1} = \overline{x}^k$ and return to the PG step. Otherwise, due to the nature of ℓ_0-norm, the restriction of $\lambda_1\|Bx\|_0 + \lambda_2\|x\|_0$ on the supports $\operatorname{supp}(Bx^k)$ and $\operatorname{supp}(x^k)$, i.e., $\lambda_1\|(Bx)_{\operatorname{supp}(Bx^k)}\|_0 + \lambda_2\|x_{\operatorname{supp}(x^k)}\|_0$, is a constant near x^k and does not provide any useful information at all and thus, unlike dealing with the ℓ_q regularization problem in Chapter 3, we introduce the following multifunction $\Pi : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$:

$$\begin{aligned}\Pi(z) &:= \{x \in \Omega \mid \operatorname{supp}(x) \subseteq \operatorname{supp}(z),\ \operatorname{supp}(Bx) \subseteq \operatorname{supp}(Bz)\} \\ &= \left\{x \in \Omega \mid x_{[\operatorname{supp}(z)]^c} = 0,\ (Bx)_{[\operatorname{supp}(Bz)]^c} = 0\right\},\end{aligned} \tag{4.4}$$

and consider the associated subproblem

$$\min_{x\in\mathbb{R}^n} f(x) + \delta_{\Pi_k}(x) \;\text{ with }\; \Pi_k := \Pi(x^k). \tag{4.5}$$

It is noted that the set $\Pi(x^k)$ contains all the points whose supports are a subset of the support of x^k as well as the supports of their linear transformation are a subset of the support of the linear transformation of x^k. It is clear that Π is closed-valued. For z_1, z_2 satisfying $\operatorname{supp}(z_1) = \operatorname{supp}(z_2)$ and $\operatorname{supp}(Bz_1) = \operatorname{supp}(Bz_2)$, we have $\Pi(z_1) = \Pi(z_2)$. Since for $z_k \to \overline{z}$, $\operatorname{supp}(\overline{z}) \subset \operatorname{supp}(z^k)$ and $\operatorname{supp}(B\overline{z}) \subset \operatorname{supp}(Bz^k)$, the multifunction Π is not closed.

We will show that a critical point of (4.5) is one for problem (4.1). Thus, instead of a subspace regularized Newton step in Chapter 3, 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); Liu et al. (2024)), our projected regularized Newton step minimizes the following second-order approximation of (4.5) on Π_k:

$$\mathop{\arg\min}_{x\in\mathbb{R}^n}\Big\{\Theta_k(x):=f(x^k)+\langle\nabla f(x^k),x-x^k\rangle+\frac{1}{2}\langle x-x^k,G_k(x-x^k)\rangle+\delta_{\Pi_k}(x)\Big\},\tag{4.6}$$

where G_k is an approximation to the Hessian $\nabla^2 f(x^k)$, satisfying the following positive definiteness condition:

$$G_k\succeq b_2\|\overline{\mu}_k(x^k-\overline{x}^k)\|^{\sigma}I,\tag{4.7}$$

where $b_2>0$, $\sigma\in(0,\frac{1}{2})$ and $\overline{\mu}_k$ is the one in (4.2). The detailed construction of G_k is presented in (4.30)-(4.32). To cater for the practical computation, our Newton step seeks an inexact solution y^k of (4.6) satisfying

$$\begin{cases}\Theta_k(y)-\Theta_k(x^k)\le 0, & (4.8)\\[2mm] \mathrm{dist}(0,\partial\Theta_k(y))\le\dfrac{\min\{\overline{\mu}_k^{-1},1\}}{2}\min\big\{\|\overline{\mu}_k(x^k-\overline{x}^k)\|,\|\overline{\mu}_k(x^k-\overline{x}^k)\|^{1+\varsigma}\big\} & (4.9)\end{cases}$$

with $\varsigma\in(\sigma,1]$. Setting the direction $d^k:=y^k-x^k$, a step size $\alpha_k\in(0,1]$ is found in the direction d^k via backtrackings, and set $x^{k+1}:=x^k+\alpha_k d^k$. To ensure the global convergence, the next iterate still returns to the PG step. The details of the algorithm are given in Section 4.2.

The main contributions of this chapter are as follows:

(i) Based on dynamic programming principle, we develop a polynomial-time algorithm with complexity ($O(n^{3+o(1)})$) for seeking a point $\overline{x}_k$ in the proximal mapping (4.2) of g when $B=\widehat{B}$, with $\widehat{B}$ being the one in (1.4). This generalizes the corresponding result in (Jewell et al. (2020)) for finding $\overline{x}^k$ in (4.2) from

$g(\cdot) = \lambda_1\|\widehat{B}\cdot\|_0$ to $g(\cdot) = \lambda_1\|\widehat{B}\cdot\|_0 + \lambda_2\|\cdot\|_0 + \delta_\Omega(\cdot)$. This also provides a PG algorithm for solving (4.1). We establish a uniform lower bound on $\text{prox}_{\mu^{-1}g}(x)$ for x on a compact set and μ on a closed interval. This generalizes the corresponding results in Lu (2014a) for ℓ_0-norm and in Lemma 3.1 for ℓ_q-norm with $0<q<1$, respectively.

(ii) We design a hybrid algorithm (PGiPN) of PG and inexact projected regularized Newton method to solve the structured ℓ_0-norms regularization problem (4.1), 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 (4.3) 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 isolatedness and the local optimality of the limit point.

(iii) The numerical experiments show that our PGiPN is more effective than some existing algorithms in the literature in terms of solution quality and efficiency.

The rest of the paper is organized as follows. In Section 4.1 we give some preliminaries on stationary conditions of model (4.1) and some results related to g, including a lower bound of proximal mapping of g, and an algorithm for finding a point in the proximal mapping of $\lambda_1\|\widehat{B}x\|_0 + \lambda_2\|x\|_0 + \delta_\Omega(x)$. In Section 4.2, we introduce our algorithm and show that it is well defined. In Section 4.3 we present the convergence analysis of our algorithm. The implementation scheme of our algorithm and the numerical experiments are presented in Section 4.4.

4.1 Preliminaries on Structured ℓ_0-norms Regularied Problem

4.1.1 Generalized Subdifferential

Since problem (4.1) involves a box constraint and the structured ℓ_0-norms function is lower semicontinuous, the set of global optimal solutions of model (4.1) is nonempty and compact. Moreover, by the continuity of $\nabla^2 f$ and the compactness of Ω, we have ∇f is Lipschitz continuous on Ω, i.e., there exists $L_1 > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \le L_1\|x - y\| \ \text{ for all } x, y \in \Omega. \tag{4.10}$$

Recall that multifunction function $\Pi : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is defined in (4.4), by which we characterize the generalized subdifferential of F.

Before presenting the theoretical results, we make an overview on the regular subdifferential of $\|B\cdot\|_0$. Fix any $z \in \mathbb{R}^n$. It is not hard to check that for $z^k \to z$, it holds that $\|Bz^k\|_0 \ge \|Bz\|_0$. Let

$$\Gamma_1 := \{x \mid \|Bx\|_0 = \|Bz\|_0\} \ \text{ and } \ \Gamma_2 := \{x \mid \|Bx\|_0 > \|Bz\|_0\}.$$

For $\Gamma_2 \ni z^k \to z$, we deduce that for any $v \in \mathbb{R}^n$,

$$\liminf_{\Gamma_2 \ni z^k \to z} \frac{\|Bz^k\|_0 - \|Bz\|_0 - \langle v, z^k - z\rangle}{\|z^k - z\|} \ge 0.$$

Therefore, from the definition of regular subdifferential, we have

$$\begin{aligned}\widehat{\partial}\|B\cdot\|_0(z) &= \left\{v \in \mathbb{R}^n \mid \liminf_{\Gamma_1 \ni z^k \to z} \frac{\|Bz^k\|_0 - \|Bz\|_0 - \langle v, z^k - z\rangle}{\|z^k - z\|} \ge 0\right\} \\ &= \left\{v \in \mathbb{R}^n \mid \liminf_{\Gamma_1 \ni z^k \to z} \frac{-\langle v, z^k - z\rangle}{\|z^k - z\|} \ge 0\right\}.\end{aligned}$$

From this deduction, the regular subdifferential of $\|B\cdot\|_0$ at z is likely to be a normal cone of a subspace at z. We make a formal statement in the following lemma.

Lemma 4.1. *Fix any $z \in \Omega$. The following statements are true.*

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

(ii) $0 \in \nabla f(x) + \mathcal{N}_{\Pi(z)}(x)$ *implies that* $0 \in \partial F(x)$.

Proof. The first equality of part (i) follows by (Rockafellar and Wets, 2009, Exercise 8.8), and the second one uses (Pan et al., 2023, Lemma 2.2 (i)). Next we consider part (ii). Let $x \in \Pi(z)$. From the definition of $\Pi(\cdot)$, we have $\Pi(x) \subseteq \Pi(z)$, which along with $x \in \Pi(x)$ implies that $\mathcal{N}_{\Pi(z)}(x) \subseteq \mathcal{N}_{\Pi(x)}(x)$. Combining part (i), we obtain the desired result. □

Remark 4.1. *Lemma 4.1 (ii) provides a way to seek a critical point of F. Indeed, for any given $z \in \mathbb{R}^n$, if x is a critical point of problem* $\arg\min\{f(y) \mid y \in \Pi(z)\}$, *i.e.,* $0 \in \nabla f(x) + \mathcal{N}_{\Pi(z)}(x)$, *then by Lemma 4.1 (ii) it necessarily satisfies* $0 \in \partial F(x)$. *This technique will be utilized in the design of our algorithm. In particular, when obtaining a good estimate of the critical point, say x^k, we use a Newton step to minimize f over the polyhedral set $\Pi(x^k)$, so as to enhance the speed of the algorithm.*

4.1.2 Prox-regularity of g

In this subsection, we aim at proving the prox-regularity of g, which together with Proposition 2.1 and the prox-boundedness of g indicates that the set of critical points of problem (4.1) coincides with that of its L-type stationary points.

We remark here that the prox-regularity of g cannot be obtained from the existing calculus of prox-regularity. In fact, 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 $v_i \in \partial f_i(\overline{x})$ and let $v := v_1 + v_2$, and $f_0 := f_1 + f_2$, a sufficient condition such that f_0 is prox-regular at $\overline{x}$ for v is

$$w_1 + w_2 = 0 \text{ with } w_i \in \partial^\infty f_i(\overline{x}) \Longrightarrow w_i = 0,\ i = 1, 2, \tag{4.11}$$

where ∂^∞ denotes the horizon subdifferential (Definition 2.1). 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,

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

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 (4.11) does not hold. Next, we give our proof toward the prox-regularity of g.

Lemma 4.2. *The function g is prox-regular on its domain Ω. Consequently, the set of critical points of model* (4.1) *coincides with its set of L-type stationary points.*

Proof. 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 $\varepsilon \in (0, \min\{\lambda, \|\overline{v}\|, \frac{\lambda}{5\|\overline{v}\|}\})$ such that for all $x \in \mathbb{B}(\overline{x}, \varepsilon)$, $\text{supp}(Cx) \supseteq \text{supp}(C\overline{x})$. Next we prove that

$$g(x') \geq g(x) + v^\top(x' - x), \text{ for all } \|x' - \overline{x}\| \leq \varepsilon, \ v \in \partial g(x), \ \|v - \overline{v}\| < \varepsilon \text{ and } x \in \Xi, \tag{4.12}$$

where $\Xi := \{x \mid \|x - \overline{x}\| < \varepsilon, \ g(x) < g(\overline{x}) + \varepsilon\}$, which implies that g is prox-regular at $\overline{x}$ for $\overline{v}$.

We first claim that for each $x \in \Xi$, it holds that $\text{supp}(Cx) = \text{supp}(C\overline{x})$ and $x \in \Omega$. If fact, by the definition of ε, $\text{supp}(Cx) \supseteq \text{supp}(C\overline{x})$. If $\text{supp}(Cx) \neq \text{supp}(C\overline{x})$, we have $g(x) \geq g(\overline{x}) + \lambda > g(\overline{x}) + \varepsilon$, which yields that $x \notin \Xi$. Therefore, $\text{supp}(Cx) = \text{supp}(C\overline{x})$. 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 (4.12) holds, so it suffices to consider $x' \in \mathbb{B}(\overline{x}, \varepsilon) \cap \Omega$. Note that $\text{supp}(Cx') \supseteq \text{supp}(C\overline{x}) = \text{supp}(Cx)$. If $\text{supp}(Cx') \neq \text{supp}(Cx)$, then

$g(x') \geq g(x) + \lambda$. For any $v \in \partial g(x)$ with $v \in \mathbb{B}(\overline{v}, \varepsilon)$, $\|v\| \leq \|\overline{v}\| + \varepsilon \leq 2\|\overline{v}\|$, which along with $\|x' - x\| \leq \|x' - \overline{x}\| + \|x - \overline{x}\| \leq 2\varepsilon$ implies that

$$g(x') - g(x) - v^\top (x' - x) \geq \lambda - \|v\| \|x' - x\| \geq \lambda - 4\|\overline{v}\|\varepsilon > 0.$$

Equation (4.12) holds. Next we consider the case $\mathrm{supp}(Cx') = \mathrm{supp}(Cx)$. Define

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

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 (4.12) 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. □

4.1.3 Lower Bound of the Proximal Mapping of g

Given $\lambda > 0$ and $x \in \mathbb{R}^n$, for any $z \in \mathrm{prox}_{\lambda\|\cdot\|_0}(x)$, it holds that if $|z_i| > 0$, then $|z_i| \geq \sqrt{2\lambda}$ (Lu, 2014a, Lemma 3.3). This indicates that $|z|_{\min}$ has a uniform lower bound. Such a uniform lower bound is shown to hold for ℓ_q-norm with $0 < q < 1$ (Lemma 3.1). Next, we show that such a uniform lower bound exists for g.

Lemma 4.3. *For any given compact set $\Xi \subseteq \mathbb{R}^n$ and constants $0 < \underline{\mu} < \overline{\mu}$, define*

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

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

Proof. Write $C := [B; I]$. By invoking (Bauschke et al., 1999, Corollary 3) and the compactness of Ω, there exists $\kappa_0 > 0$ such that for all index set $J \subseteq [n+p]$,

$$\mathrm{dist}(x, \mathrm{Null}(C_{J\cdot}) \cap \Omega) \leq \kappa_0 \mathrm{dist}(x, \mathrm{Null}(C_{J\cdot})) \quad \text{for any } x \in \Omega. \tag{4.13}$$

In addition, there exists $\sigma_0 > 0$ such that for any index set $J \subseteq [n+p]$ with $\{C_{j\cdot}\}_{j\in J}$ being linearly independent,

$$\lambda_{\min}(C_{J\cdot} C_{J\cdot}^{\top}) \geq \sigma_0. \tag{4.14}$$

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,

$$\begin{aligned} |h_{z,\mu}(x) - h_{z,\mu}(y)| &= \frac{\mu}{2}|\langle x - y, x + y - 2z\rangle| \\ &\leq \frac{\overline{\mu}}{2}(\|x\| + \|y\| + 2\|z\|)\|x - y\| < \frac{\lambda}{2}. \end{aligned} \tag{4.15}$$

Now suppose that the conclusion does not hold. Then there is a sequence $\{\overline{z}^k\}_{k\in\mathbb{N}} \subseteq \mathcal{Z}\backslash\{0\}$ such that $|C\overline{z}^k|_{\min} \leq \frac{1}{k}$ 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 exist $\mathcal{K} \subseteq \mathbb{N}$ and an index $i \in [n+p]$ such that

$$0 < |(C\overline{z}^k)_i| = |C\overline{z}^k|_{\min} < \frac{\delta_0\sigma_0}{\kappa_0\|C\|_2} \quad \text{for each } k \in \mathcal{K}, \tag{4.16}$$

where κ_0 and σ_0 are the ones appearing in (4.13) and (4.14), respectively. Fix any $k \in \mathcal{K}$. Write $Q_k := [n+p]\backslash\mathrm{supp}(C\overline{z}^k)$ and choose $J_k \subseteq Q_k$ such that the rows of $C_{J_k\cdot}$ form a basis of those of $C_{Q_k\cdot}$. Let $\widehat{J}_k := J_k \cup \{i\}$. If $J_k = \emptyset$, then $C_{\widehat{J}_k\cdot}$ has a full row rank. If $J_k \neq \emptyset$, 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 $|(C\overline{z}^k)_i| = |C\overline{z}^k|_{\min} > 0$). 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 $(\overline{z}^k - \widetilde{z}^k) \in \mathrm{Range}\,(C_{\widehat{J}_k\cdot}^\top)$. The latter means that 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

$$|(C\overline{z}^k)_i| = \|C_{\widehat{J}_k\cdot}\overline{z}^k - C_{\widehat{J}_k\cdot}\widetilde{z}^k\| = \|C_{\widehat{J}_k\cdot}C_{\widehat{J}_k\cdot}^\top\xi^k\| \geq \sigma_0\|\xi^k\|, \tag{4.17}$$

where the last inequality is due to (4.14). Combining (4.17) with (4.16) yields $\|\xi^k\| < \kappa_0^{-1}\|C\|_2^{-1}\delta_0$. Therefore,

$$\|\overline{z}^k - \widetilde{z}^k\| = \|C_{\widehat{J}_k\cdot}^\top\xi^k\| \leq \|C_{\widehat{J}_k\cdot}\|_2\|\xi^k\| \leq \|C\|_2\|\xi^k\| < \kappa_0^{-1}\delta_0. \tag{4.18}$$

Let $\widehat{z}^k := \mathrm{proj}_{\mathrm{Null}(C_{\widehat{J}_k\cdot})\cap\Omega}(\overline{z}^k)$. From (4.13) and (4.18), it follows that

$$\|\overline{z}^k - \widehat{z}^k\| = \mathrm{dist}(\overline{z}^k, \mathrm{Null}(C_{\widehat{J}_k\cdot}) \cap \Omega) \leq \kappa_0 \mathrm{dist}(\overline{z}^k, \mathrm{Null}(C_{\widehat{J}_k\cdot})) = \kappa_0\|\overline{z}^k - \widetilde{z}^k\| < \delta_0. \tag{4.19}$$

Note that $\widehat{z}^k, \overline{z}^k \in \Omega$. From (4.19) and (4.15), it follows that

$$|h_{z^k,\mu_k}(\widehat{z}^k) - h_{z^k,\mu_k}(\overline{z}^k)| < \frac{\lambda}{2}. \tag{4.20}$$

Next we claim that $\mathrm{supp}(C\widehat{z}^k) \cup \{i\} \subseteq \mathrm{supp}(C\overline{z}^k)$. 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 $\mathrm{supp}(C_{[Q_k\cup\{i\}]^c\cdot}\widehat{z}^k) \subseteq \mathrm{supp}(C_{[Q_k\cup\{i\}]^c\cdot}\overline{z}^k)$, which implies that $\mathrm{supp}(C\widehat{z}^k) \cup \{i\} \subseteq \mathrm{supp}(C\overline{z}^k)$. Thus, the claimed inclusion follows, which implies that $g(\overline{z}^k) - g(\widehat{z}^k) \geq \lambda$. This together with (4.20) yields

$$h_{z^k,\mu_k}(\overline{z}^k) + g(\overline{z}^k) - (h_{z^k,\mu_k}(\widehat{z}^k) + g(\widehat{z}^k)) \geq \lambda - \frac{\lambda}{2} = \frac{\lambda}{2},$$

contradicting to $\overline{z}^k \in \mathrm{prox}_{\mu_k^{-1}g}(z^k)$. The proof is completed. $\square$

The result of Lemma 4.3 will be utilized in Proposition 4.2 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 (4.2) (or (4.33) 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.

4.1.4 Proximal Mapping of a Fused ℓ_0-norms Function with a Box Constraint

Using the idea of (Killick et al. (2012)), Jewell et al. (2020) presented a polynomial-time algorithm for computing the proximal mapping of the fused ℓ_0-norm $\lambda_1\|\widehat{B}\cdot\|_0$, where $\widehat{B}x = (x_1 - x_2; ...; x_{n-1} - x_n)$ for any $x \in \mathbb{R}^n$. We extend the result of (Jewell et al. (2020)) for computing 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 $z \in \mathbb{R}^n$, seeking a global optimal solution of the problem

$$\min_{x\in\mathbb{R}^n} h(x;z) := \frac{1}{2}\|x-z\|^2 + \lambda_1\|\widehat{B}x\|_0 + \lambda_2\|x\|_0 + \delta_\Omega(x). \tag{4.21}$$

To simplify the deduction, for each $i \in [n]$, we define $w_i : \mathbb{R} \to \overline{\mathbb{R}}$ by $w_i(\alpha) := \lambda_2|\alpha|_0 + \delta_{[(l_b)_i,(u_b)_i]}(\alpha)$. It is clear that for all $x \in \mathbb{R}^n$, $\lambda_2\|x\|_0 + \delta_\Omega(x) = \sum_{i=1}^n w_i(x_i)$. Let $H(0) := -\lambda_1$, and for each $s \in [n]$, define

$$H(s) := \min_{y\in\mathbb{R}^s} h_s(y;z_{1:s}), \ \text{where } h_s(y;z_{1:s}) := \frac{1}{2}\|y - z_{1:s}\|^2 + \lambda_1\|\widehat{B}_{\cdot[s]}y\|_0 + \sum_{j=1}^s w_j(y_j). \tag{4.22}$$

It is immediate to see that $H(n)$ is the optimal value to (4.21). For each $s \in [n]$, define function $P_s\colon [0{:}s-1]\times\mathbb{R} \to \overline{\mathbb{R}}$ by

$$P_s(i,\alpha) := H(i) + \frac{1}{2}\|\alpha\mathbf{1} - z_{i+1:s}\|^2 + \sum_{j=i+1}^s w_j(\alpha) + \lambda_1. \tag{4.23}$$

For any given $y \in \mathbb{R}^s$, if i is the largest integer in $[0\!:\!s-1]$ such that $y_i \neq y_{i+1}$, and $y_{i+1} = y_{i+2} = \ldots = y_s = \alpha$, then $y = (y_{1:i}; \alpha\mathbf{1})$ and

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

If $y_{1:i}$ is optimal to $\min_{y' \in \mathbb{R}^i} h_i(y'; z_{1:i})$, then $H(i) = h_i(y_{1:i}; z_{1:i})$, which by the definitions of P_s and h_s yields that $P_s(i, \alpha) = h_s(y; z_{1:s})$. 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 $H(s)$, by which we give characterization to an optimal solution of $h_s(\cdot; z_{1:s})$.

Lemma 4.4. *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) *Assume that $(i_s^*, \alpha_s^*) \in \arg\min_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i, \alpha)$. Then $y^* = (y^*_{1:i_s^*}; \alpha_s^*\mathbf{1})$ is a global solution of $\min_{y\in\mathbb{R}^s} h_s(y; z_{1:s})$ with $y^*_{1:i_s^*} \in \arg\min_{v\in\mathbb{R}^{i_s^*}} h_{i_s^*}(v; z_{1:i_s^*})$.*

Proof. **(i)** Let y^* be an optimal solution to problem (4.22). If $y_i^* = y_j^*$ for all $i, j \in [s]$, let $i_s^* = 0$; otherwise, let i_s^* 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

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

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

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

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

$$H(s) \leq h_s(y^*; z_{1:s}) \leq h_{i^*_s}(y^*_{1:i^*_s}; z_{1:i^*_s}) + \frac{1}{2}\|y^*_{i^*_s+1:s} - z_{i^*_s+1:s}\|^2 + \sum_{j=i^*_s+1}^{s} w_j(y^*_j) + \lambda_1$$

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

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

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

Therefore, $H(s) \leq \min_{i\in[0:s-1],\alpha\in\mathbb{R}} P_s(i,\alpha)$. These two inequalities imply the result.

(ii) If $i^*_s \neq 0$, by part (i) and the definitions of α^*_s and i^*_s,

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

$$= h_{i^*_s}(y^*_{1:i^*_s}; z_{1:i^*_s}) + \frac{1}{2}\|y^*_{i^*_s+1:s} - z_{i^*_s+1:s}\|^2 + \sum_{j=i^*_s+1}^{s} w_j(y^*_j) + \lambda_1 \geq h_s(y^*, z_{1:s}),$$

where the last inequality follows by the definition of $h_s(\cdot, z_{1:s})$. If $i^*_s = 0$,

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

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

□

From Lemma 4.4 (i), the nonconvex nonsmooth problem (4.21) can be recast as a mixed-integer programming with objective function given in (4.23). Lemma 4.4 (ii) suggests a recursive method to obtain an optimal solution to (4.21) via solving (4.23). In fact, by setting $s = n$, we can obtain that there exists an optimal solution to (4.21), 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 $x^*_{i^*_{i^*_n}+1:i^*_n}$. Repeating this loop backward until $s = 0$, we can obtain the full expression of an optimal solution to (4.21). The outline of computing $\mathrm{prox}_{\lambda_1\|\widehat{B}\cdot\|_0+w(\cdot)}(z)$ is shown as follows.

$$\begin{cases} \text{Set the current changepoint } s = n. \\ \text{While } s > 0 \text{ do} \\ \qquad \text{Find } (i^*_s, \alpha^*_s) \in \underset{i \in [0:s-1], \alpha \in \mathbb{R}}{\arg\min} \; P_s(i, \alpha). \\ \qquad \text{Let } x^*_{i^*_s+1:s} = \alpha^*_s \mathbf{1} \text{ and } s \leftarrow i^*_s. \\ \text{End} \end{cases} \tag{4.24}$$

To obtain an optimal solution to (4.21), the remaining issue is how to execute the first line in **while** loop of (4.24), 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 4.4 (ii). The following proposition provides some preparations.

Proposition 4.1. *For each $s \in [n]$, let $P^*_s(\alpha) := \min_{i \in [0:s-1]} P_s(i, \alpha)$.*

(i) *For all $\alpha \in \mathbb{R}$,*

$$P^*_s(\alpha) = \begin{cases} \frac{1}{2}(\alpha - z_1)^2 + w_1(\alpha) & \text{if } s = 1, \\ \min\left\{P^*_{s-1}(\alpha), \min_{\alpha' \in \mathbb{R}} P^*_{s-1}(\alpha') + \lambda_1\right\} + \frac{1}{2}(\alpha - z_s)^2 + w_s(\alpha) & \text{if } s \in [2:n]. \end{cases}$$

(ii) *Let $\mathcal{R}^0_1 := \mathbb{R}$, and $\mathcal{R}^i_s := \mathcal{R}^i_{s-1} \cap (\mathcal{R}^{s-1}_s)^c$ for all $s \in [2:n]$ and $i \in [0:s-2]$, where*

$$\mathcal{R}^{s-1}_s := \left\{\alpha \in \mathbb{R} \mid P^*_{s-1}(\alpha) \geq \min_{\alpha' \in \mathbb{R}} P^*_{s-1}(\alpha') + \lambda_1\right\}. \tag{4.25}$$

(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)** Note that $P_1^*(\alpha) = P_1(0,\alpha) = H(0) + \frac{1}{2}(\alpha - z_1)^2 + w_1(\alpha) + \lambda_1 = \frac{1}{2}(\alpha - z_1)^2 + w_1(\alpha)$. Now fix any $s \in [2:n]$. By the definition of P_s^*, for any $\alpha \in \mathbb{R}$,

$$P_s^*(\alpha) = \min_{i\in[0:s-1]} P_s(i,\alpha) = \min\Big\{ \min_{i\in[0:s-2]} P_s(i,\alpha),\ P_s(s\!-\!1,\alpha)\Big\}. \tag{4.26}$$

From the definition of P_s in (4.23), for each $i \in [0\!:\!s\!-\!2]$ and $\alpha \in \mathbb{R}$, it holds that

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

while for any $\alpha \in \mathbb{R}$, $P_s(s\!-\!1,\alpha) = H(s\!-\!1) + \frac{1}{2}(\alpha - z_s)^2 + w_s(\alpha) + \lambda_1$. By combining the last two equalities with (4.26), we immediately obtain that

$$\begin{aligned} P_s^*(\alpha) &= \min\Big\{ \min_{i\in[0:s-2]} P_{s-1}(i,\alpha),\ H(s\!-\!1) + \lambda_1\Big\} + \frac{1}{2}(\alpha - z_s)^2 + w_s(\alpha) \\ &= \min\Big\{P_{s-1}^*(\alpha), \min_{\alpha'\in\mathbb{R}} P_{s-1}^*(\alpha') + \lambda_1\Big\} + \frac{1}{2}(\alpha - z_s)^2 + w_s(\alpha), \end{aligned} \tag{4.27}$$

where the last equality follows by Lemma 4.4 (i). 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 when $s = j$ for some $j \in [2\!:\!n\!-\!1]$. We consider the case $s = j\!+\!1$. Since $\mathcal{R}_s^i := \mathcal{R}_{s-1}^i \cap (\mathcal{R}_s^{s-1})^c$ for all $i \in [0\!:\!s\!-\!2]$ and $\bigcup_{i\in[0:s-2]} \mathcal{R}_{s-1}^i = \mathbb{R}$, it holds that

$$\textstyle\bigcup_{i\in[0:s-1]}\mathcal{R}_s^i = \big[\bigcup_{i\in[0:s-2]}(\mathcal{R}_{s-1}^i \cap (\mathcal{R}_s^{s-1})^c)\big] \cup \mathcal{R}_s^{s-1} = \big(\mathbb{R} \cap (\mathcal{R}_s^{s-1})^c\big) \cup \mathcal{R}_s^{s-1} = \mathbb{R}.$$

Thus we obtain the first part of (a) by deduction. For any $i \in [0\!:\!s\!-\!2]$, by definition, $\mathcal{R}_s^i \cap \mathcal{R}_s^{s-1} = \emptyset$. It suffices to show that $\mathcal{R}_s^i \cap \mathcal{R}_s^j = \emptyset$ for any $i \neq j \in [0\!:\!s\!-\!2]$. By definition,

$$\mathcal{R}_s^i \cap \mathcal{R}_s^j = \left[\mathcal{R}_{s-1}^i \cap (\mathcal{R}_s^{s-1})^c\right] \cap \left[\mathcal{R}_{s-1}^j \cap (\mathcal{R}_s^{s-1})^c\right] = \emptyset,$$

where the last equality is using $\mathcal{R}_{s-1}^i \cap \mathcal{R}_{s-1}^j = \emptyset$. Thus, the second part of (a) is obtained.

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

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

where the second equality is using Lemma 4.4 (i) and the definition of P_s. 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 (4.25) we have $P_1^*(\alpha) < \min_{\alpha' \in \mathbb{R}} P_1^*(\alpha') + \lambda_1$, which by part (i) implies that $P_2^*(\alpha) = P_1^*(\alpha) + \frac{1}{2}(\alpha - z_2)^2 + w_2(\alpha) = P_1(0, \alpha) + \frac{1}{2}(\alpha - z_2)^2 + w_2(\alpha) = P_2(0, \alpha)$. 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 $i \in [0\!:\!s\!-\!2]$, by definition, $\mathcal{R}_s^i = \mathcal{R}_{s-1}^i \cap (\mathcal{R}_s^{s-1})^c$. Then, from (4.27) for any $\alpha \in \mathcal{R}_s^i$,

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

where the second equality is using $P_{s-1}^*(\alpha) = P_{s-1}(i, \alpha)$ 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 4.1. 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 w_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 $P_s^*(\cdot)=\min\{P_s(0,\cdot),P_s(1,\cdot),\ldots,P_s(s-1,\cdot)\}$. Note that 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 $0, \frac{\sum_{j=i+1}^{s} z_j}{s-i}, \max_{j\in[i+1:s]}\{(l_b)_j\}, \min_{j\in[i+1:s]}\{(u_b)_j\}$), 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 4 for solving $\mathrm{prox}_{\lambda_1\|\widehat{B}\cdot\|_0+w(\cdot)}(z)$, whose iterate steps are described as follows.

Algorithm 4 (Computing $\mathrm{prox}_{\lambda_1\|\widehat{B}\cdot\|_0+w(\cdot)}(z)$)

---

1. **Initialize:** Compute $P_1^*(\alpha)=\frac{1}{2}(z_1-\alpha)^2+w_1(\alpha)$ and set $\mathcal{R}_1^0=\mathbb{R}$.
2. **For** $s=2,\ldots,n$ **do**
3. $\quad P_s^*(\alpha):=\min\{P_{s-1}^*(\alpha),\min_{\alpha'\in\mathbb{R}} P_{s-1}^*(\alpha')+\lambda_1\}+\frac{1}{2}(\alpha-z_s)^2+w_s(\alpha)$.
4. $\quad$ Compute $\mathcal{R}_s^{s-1}$ by (4.25).
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 the current changepoint $s=n$.
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**

The main computation cost of Algorithm 4 comes from lines 3 and 6, in which the number of pieces of the linear-quadratic functions involved in P_s^* plays a crucial role. The following lemma gives a worst-case estimation for the number of pieces of P_s^* in the s-th iterate.

Lemma 4.5. *Fix any $s\in[2:n]$. The function P_s^* in line 3 of Algorithm 4 has at most $O(s^{1+o(1)})$ linear-quadratic pieces.*

Proof. Let $h_i(\alpha) := H(i) + \frac{1}{2}\|\alpha\mathbf{1} - z_{i+1:s}\|^2 + \lambda_1 + (s-i)\lambda_2|\alpha|_0 + \sum_{j=i+1}^{s} \delta_{[(l_b)_j,(u_b)_j]}(\alpha)$ for $\alpha \in \mathbb{R}$ with $i \in [0\!:\!s-1]$. From the definition of P_s^*, it holds that

$$P_s^*(\alpha) = \min_{i\in[0:s-1]} \{h_i(\alpha)\}, \text{ for } \alpha \in \mathbb{R}. \tag{4.28}$$

For each $i \in [0\!:\!s–1]$, 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 $i \in [0\!:\!s-1]$,

$$h_i = \min\left\{h_{i,1}, h_{i,2}, h_{i,3}\right\}, \text{ with} \tag{4.29}$$

$$h_{i,1}(\alpha) := h_i(\alpha) - (s-i)\lambda_2|\alpha|_0 + (s-i)\lambda_2 + \delta_{(-\infty,0]}(\alpha),\ h_{i,2}(\alpha) := h_i(\alpha) + \delta_{\{0\}}(\alpha),$$

$$h_{i,3}(\alpha) := h_i(\alpha) - (s-i)\lambda_2|\alpha|_0 + (s-i)\lambda_2 + \delta_{[0,\infty)}(\alpha).$$

Obviously, $h_{i1}, h_{i,2}$ and $h_{i,3}$ are piecewise linear-quadratic functions with domain being a closed interval. Combining (4.29) with (4.28), for any $\alpha \in \mathbb{R}$,

$$P_s^*(\alpha) = \left\{h_0(\alpha), h_1(\alpha), \ldots, h_{s-2}(\alpha), h_{s-1}(\alpha)\right\} = \min_{i\in[0:s-1], j\in[3]} \left\{h_{i,j}(\alpha)\right\}.$$

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, 1988, 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 $O(s^{1+o(1)})$. The proof is finished. □

By invoking Lemma 4.5, we are able to provide a worst-case estimation for the complexity of Algorithm 4. Indeed, the main cost of Algorithm 4 consists in lines 3 and 5-7. Since line 3 involves the computation cost proportional to the pieces of P_{s-1}^*, from Lemma 4.5, it requires $O(s^{1+o(1)})$ operation. For each $i \in [0 : s-1]$, from part (b) of Proposition 4.1 (ii), we know that $\mathcal{R}_s^i$ consists of at most $O(s^{1+o(1)})$ intervals, which means that line 6 requires at most $O(s^{1+o(1)})$ operations and then the complexity of lines 5-7 is $O(s^{2+o(1)})$. Thus, the worst-case complexity of Algorithm 4 is $\sum_{s=2}^{n} O(s^{2+o(1)}) = O(n^{3+o(1)})$.

4.2 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. We now state the details of our algorithm, a hybrid of PG and inexact projected regularized Newton methods (PGiPN), for solving problem (4.1), where the introduction of the switch condition (4.3) is due to the consideration that the PG step is more cost-effective than the Newton step when the iterates are far from a critical point. Let $x^k \in \Omega$ be the current iterate. It is noted that the PG step is always executed and if condition (4.3) is met, we need to solve (4.6), which involves constructing G_k to satisfy (4.7). Such G^k can be easily achieved in the following situations.

For some generalized linear models, 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 twice continuously differentiable, separable h. For this case, $\nabla^2 h$ is a diagonal matrix, and $\nabla^2 f(x) = A^\top \nabla^2 h(Ax - b)A$. Since $\nabla^2 f(x^k)$ is not necessarily positive definite, following the method in (Liu et al. (2024)), we construct $G_k := G_k^1$, where

$$G_k^1 := \nabla^2 f(x^k) + b_1[-\lambda_{\min}(\nabla^2 h(Ax^k - b))]_+ A^\top A + b_2\|\overline{\mu}_k(x^k - \overline{x}^k)\|^\sigma I \qquad (4.30)$$

with $b_1 \geq 1$. 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 and simultaneously make G_k positive definite, Zhang et al. (2023) considered $G_k := G_k^2$, where

$$G_k^2 := A^\top [\nabla^2 h(Ax^k - b)]_+ A + b_2 \|\overline{\mu}_k (x^k - \overline{x}^k)\|^\sigma I. \tag{4.31}$$

For the case where $\nabla^2 f(x^k) \succeq 0$, $G_k^1 = G_k^2$. If $\nabla^2 f(x^k) \not\succeq 0$, it is immediate to see that $\|G_k^1 - \nabla^2 f(x^k)\|_2 \geq \|G_k^2 - \nabla^2 f(x^k)\|_2$, which means that G_k^2 is a better approximation to $\nabla^2 f(x^k)$ than G_k^1. On the other hand, for those f's not owning a separable structure, we form $G_k := G_k^3$ as in (Ueda and Yamashita (2010)) and HpgSRN in Chapter 3, where

$$G_k^3 := \nabla^2 f(x^k) + \big(b_1[-\lambda_{\min}(\nabla^2 f(x^k))]_+ + b_2 \|\overline{\mu}_k (x^k - \overline{x}^k)\|^\sigma\big) I. \tag{4.32}$$

It is not hard to check that for $i = 1, 2, 3$, G_k^i meets the requirement in (4.7). We remark here that the sequel convergence analysis holds for all three G_k^i, and we write them by G_k for simplicity.

Now we are in position to present the detailed iterates of our algorithm.

Algorithm 5 (a hybrid of PG and inexact projected regularized Newton methods)

Initialization: Choose $\epsilon \geq 0$ and parameters $\mu_{\max} > \mu_{\min} > 0$, $\tau > 1$, $\alpha > 0, b_2 > 0$, $\varrho \in (0, \frac{1}{2})$, $\sigma \in (0, \frac{1}{2})$, $\varsigma \in (\sigma, 1]$ 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

$$F(\overline{x}^k) \leq F(x^k) - \frac{\alpha}{2} \|x^k - \overline{x}^k\|^2 \ \text{ with } \ \overline{x}^k \in \text{prox}_{(\mu_k \tau^m)^{-1} g}(x^k - (\mu_k \tau^m)^{-1} \nabla f(x^k)). \tag{4.33}$$

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

(1c) If condition (4.3) 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 (4.6) satisfying (4.8)-(4.9).

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

$$f(x^k + \beta^t d^k) \leq f(x^k) + \varrho \beta^t \langle \nabla f(x^k), d^k \rangle. \tag{4.34}$$

(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 4.2. **(a)** *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 4.3, switch condition* (4.3) *always holds and the supports of* $\{Bx^k\}_{k\in\mathbb{N}}$ *and* $\{x^k\}_{k\in\mathbb{N}}$ *keep unchanged when* k *is sufficiently large, so that Algorithm 5 will reduce to an inexact projected regularized Newton method for solving* (4.5) *with* $\Pi_k \equiv \Pi_*$. *In this sense, the PG step plays a crucial role in transforming the original challenging problem* (4.1) *into a problem that can be efficiently solved by the inexact projected regularized Newton method.*

(b) *When* x^k *enters the Newton step, from the inexact criterion* (4.8) *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*

$$\langle \nabla f(x^k), d^k\rangle \leq -\frac{1}{2}\langle d^k, G_k d^k\rangle \leq -\frac{b_2}{2}\|\overline{\mu}_k(x^k - \overline{x}^k)\|^\sigma \|d^k\|^2 < 0, \tag{4.35}$$

where the second inequality is due to (4.7). *In addition, the inexact criterion* (4.8) *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, $\mathrm{supp}(B(x^k+\alpha d^k)) \subseteq \mathrm{supp}(Bx^k)$ *and* $\mathrm{supp}(x^k+\alpha d^k) \subseteq \mathrm{supp}(x^k)$, *so* $g(x^k+\alpha d^k) \leq g(x^k)$ *for any* $\alpha \in (0,1]$. *This together with* (4.35) *shows that the iterate along the direction* d^k *will reduce the value of* F *at* x^k.

(c) *When* $\epsilon = 0$, *by Definition 2.2 the output* x^k *of Algorithm 5 is an L-type stationary point of* (4.1), *which is also a critical point of problem* (4.5) *from Proposition 2.1 and Lemma 4.1 (i). Let* $r_k : \mathbb{R}^n \to \mathbb{R}^n$ *be the KKT residual mapping of* (4.5) *defined by*

$$r_k(x) := \overline{\mu}_k[x - \mathrm{proj}_{\Pi_k}(x - \overline{\mu}_k^{-1}\nabla f(x))]. \tag{4.36}$$

It is not difficult to verify that when x^k *satisfies condition* (4.3), *the following relation* *holds*

$$r_k(x^k) = \overline{\mu}_k(x^k - \overline{x}^k), \tag{4.37}$$

for which it suffices to argue that $\overline{x}^k = \mathrm{proj}_{\Pi_k}(x^k - \overline{\mu}_k^{-1}\nabla f(x^k))$*. Indeed, if not, 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* $\mathrm{supp}(B\overline{z}^k) \subseteq \mathrm{supp}(B\overline{x}^k)$ *and* $\mathrm{supp}(\overline{z}^k) \subseteq \mathrm{supp}(\overline{x}^k)$*, which implies that* $g(\overline{z}^k) \le g(\overline{x}^k)$ *and then* $\widetilde{h}_k(\overline{z}^k)+g(\overline{z}^k) < \widetilde{h}_k(\overline{x}^k)+g(\overline{x}^k)$*, a contradiction to* $\overline{x}^k \in \mathrm{prox}_{\overline{\mu}_k^{-1}g}(x^k - \overline{\mu}_k^{-1}\nabla f(x^k))$*.*

(d) *The line search in step (1a) must stop after a finite number of backtrackings. In fact, by using equation* (4.10) *and Remark 2.2, we deduce that when* $\mu_k\tau^m \ge L_1+\alpha$*,* (4.33) *must hold, which implies that* $\overline{\mu}_k < \widetilde{\mu} := \tau(L_1+\alpha)$ *for each* $k \in \mathbb{N}$*.*

By Remark 3.1 (d), to show that Algorithm 5 is well defined, we only need to argue that the Newton steps in Algorithm 5 are well defined, which is implied by the following lemma.

Lemma 4.6. *For each* $k \in \mathbb{N}$*, define the KKT residual mapping* $R_k : \mathbb{R}^n \to \mathbb{R}^n$ *of* (4.6) *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* (4.3)*, the following statements are true.*

(i) *For any* y *close enough to the optimal solution of* (4.6)*,* $y - \overline{\mu}_k^{-1}R_k(y)$ *satisfies inexact conditions* (4.8)-(4.9)*.*

(ii) *The line search step in* (4.34) *terminates after a finite number of backtrackings, and* $\alpha_k \ge \min\{1, \frac{(1-\varrho)b_2\beta}{L_1}\|\overline{\mu}_k(x^k - \overline{x}^k)\|^\sigma\}$*.*

(iii) *The inexact criterion* (4.9) *implies that* $\|R_k(y^k)\| \le \frac{1}{2}\min\{\|r_k(x^k)\|, \|r_k(x^k)\|^{1+\varsigma}\}$*.*

Proof. Pick any x^k satisfying (4.3). We proceed the proof of parts (i)-(iii) as follows.

(i) Let $\widehat{y}^k$ be the unique optimal solution to (4.6). Then $\widehat{y}^k \neq x^k$ (if not, x^k is the optimal solution of (4.6) and $0 = R_k(x^k) = r_k(x^k)$, which by (4.37) means that $x^k = \overline{x}^k$ and Algorithm 5 stops at x^k). By the optimality condition of (4.6), $-\nabla f(x^k) - G_k(\widehat{y}^k - x^k) \in \mathcal{N}_{\Pi_k}(\widehat{y}^k)$, which by the convexity of Π_k and $x^k \in \Pi_k$ implies that $\langle \nabla f(x^k) + G_k(\widehat{y}^k - x^k), \widehat{y}^k - x^k \rangle \le 0$. Along with the expression of Θ_k, we have $\Theta_k(\widehat{y}^k) - \Theta_k(x^k) \le -\frac{1}{2}\langle \widehat{y}^k - x^k, G_k(\widehat{y}^k - x^k)\rangle < 0$. Since Θ_k is continuous relative to Π_k, for any $z \in \Pi_k$ sufficiently close to $\widehat{y}^k$, $\Theta_k(z) - \Theta_k(x^k) \le 0$. From $R_k(\widehat{y}) = 0$ and the continuity of R_k, $y - \overline{\mu}_k^{-1} R_k(y)$ is close to $\widehat{y}$ when y sufficiently 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 (4.8) when y is sufficiently close to $\widehat{y}$. 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, $\mathrm{dist}(0, \partial\Theta_k(y - \overline{\mu}_k^{-1} R_k(y))) \le \|\overline{\mu}_k^{-1} G_k R_k(y) + R_k(y)\|$. 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 (4.9).

(ii) By (4.10) and the descent lemma (Bertsekas, 1997, Proposition A.24), for any $\alpha \in (0, 1]$,

$$\begin{aligned} f(x^k + \alpha d^k) - f(x^k) - \varrho\alpha\langle \nabla f(x^k), d^k\rangle &\le (1-\varrho)\alpha\langle \nabla f(x^k), d^k\rangle + \frac{L_1\alpha^2}{2}\|d^k\|^2 \\ &\le -\frac{(1-\varrho)\alpha b_2}{2}\|\overline{\mu}_k(x^k - \overline{x}^k)\|^\sigma \|d^k\|^2 + \frac{L_1\alpha^2}{2}\|d^k\|^2 \\ &= \Big(-\frac{(1-\varrho)b_2}{2}\|\overline{\mu}_k(x^k - \overline{x}^k)\|^\sigma + \frac{L_1\alpha}{2}\Big)\alpha\|d^k\|^2, \end{aligned}$$

where the second inequality uses (4.35). Therefore, when

$$\alpha \le \min\Big\{1, \frac{(1-\varrho)b_2}{L_1}\|\overline{\mu}_k(x^k - \overline{x}^k)\|^\sigma\Big\},$$

the line search in (4.34) holds, which implies that

$$\alpha_k \geq \min\big\{1, \frac{(1-\varrho)b_2\beta}{L_1}\|\overline{\mu}_k(x^k - \overline{x}^k)\|^\sigma\big\}.$$

(iii) Let $\xi^k \in \partial\Theta_k(y^k)$ be such that $\|\xi^k\| = \mathrm{dist}(0, \partial\Theta_k(y^k))$. From $\xi^k \in \partial\Theta_k(y^k)$ and the expression of Θ_k, we have $y^k = \mathrm{proj}_{\Pi_k}(y^k + \xi^k - (G_k(y^k - x^k) + \nabla f(x^k)))$. Along with the nonexpansiveness of proj_{Π_k}, $\|y^k - \mathrm{proj}_{\Pi_k}(y^k - (G_k(y^k - x^k) + \nabla f(x^k)))\| \leq \|\xi^k\|$. 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 by (Sra, 2012, Lemma 4) and the expression of R_k. Combining the last inequality with (4.9) and (4.37) leads to the desired inequality. $\square$

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 first inexact condition in (Yue et al., 2019, Equa (6a)) or the inexact condition in (Mordukhovich et al., 2023, Equa (14)), which by Lemma 4.6 (iii) shows that criterion (4.9) with $\overline{\mu}_k = 1$ is stronger than those ones.

To analyze the convergence of Algorithm 5 with $\epsilon = 0$, henceforth we assume $x^k \neq \overline{x}^k$ for all k (if not, Algorithm 5 will produce an L-type stationary point within finite number of steps, and its convergence holds automatically). From the iterate steps of Algorithm 5, 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 4.2. *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 $|B\overline{x}^k|_{\min} \geq \nu$ and $|\overline{x}^k|_{\min} \geq \nu$ for all $k \in \mathbb{N}$.*

(iii) *There exist $c_1, c_2 > 0$ such that $c_1\|r_k(x^k)\| \leq \|d^k\| \leq c_2\|r_k(x^k)\|^{1-\sigma}$ 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 (4.34) and (4.35), it follows that $f(x^{k+1}) < f(x^k)$, which along with $g(x^{k+1}) \leq g(x^k)$ by Remark 4.2 (b) 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 4.2 (d), $\overline{\mu}_k \in [\mu_{\min}, \widetilde{\mu})$ for all $k \in \mathbb{N}$. Note that $\{x^k\}_{k\in\mathbb{N}} \subseteq \Omega$, 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 4.3 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 $|B\overline{x}^k|_{\min} \geq |[B; I]\overline{x}^k|_{\min}$ and $|\overline{x}^k|_{\min} \geq |[B; I]\overline{x}^k|_{\min}$.

(iii) From the definition of G_k, the continuity of $\nabla^2 f$, $\{x^k, \overline{x}^k\}_{k\in\mathbb{N}} \subseteq \Omega$ and Remark 4.2 (d), there exists $\overline{c} > 0$ such that

$$\|G_k\|_2 \leq \overline{c} \text{ for all } k \in \mathcal{K}_2. \tag{4.38}$$

Fix any $k \in \mathcal{K}_2$. By Lemma 4.6 (iii), $\|R_k(y^k)\| \leq \frac{1}{2}\|r_k(x^k)\|$. Then, it holds that

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

where the third inequality is using the nonexpansiveness of proj_{Π_k}, and the last one is due to (4.38) 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 (4.7), (4.37) and Lemma 4.6 (iii) leads to

$$\begin{aligned}b_2\|r_k(x^k)\|^{\sigma}\|d^k\|^2 &\le (1+\overline{\mu}_k^{-1}\|G_k\|_2)(\|R_k(y^k)\|+\|r_k(x^k)\|)\|d^k\| \qquad (4.39)\\ &\le (3/2)(1+\overline{\mu}_k^{-1}\|G_k\|_2)\|r_k(x^k)\|\|d^k\|,\end{aligned}$$

which along with (4.38) and $\mu_k \ge \mu_{\min}$ implies that $\|d^k\| \le \frac{3}{2}(1+\mu_{\min}^{-1}\overline{c})b_2^{-1}\|r_k(x^k)\|^{1-\sigma}$. Then, $\|d^k\| \le c_2\|r_k(x^k)\|^{1-\sigma}$ holds with $c_2 := \frac{3}{2}(1+\mu_{\min}^{-1}\overline{c})b_2^{-1}$. The proof is completed. □

4.3 Convergence Analysis

Before analyzing the convergence of Algorithm 5, we show that Algorithm 5 finally reduces to an inexact projected regularized Newton method for seeking a critical point of a problem to minimize a smooth function over a polyhedral set. This requires the following lemma, which presents the descent proeprty of $\{F(x^k)\}_{k\in\mathbb{N}}$, and proves that all the elements of $\omega(x^0)$ is an L-type stationary point (recall that $\omega(x^0)$ denotes the set of accumulation points of $\{x^k\}_{k\in\mathbb{N}}$).

Lemma 4.7. *For the sequences* $\{x^k\}_{k\in\mathbb{N}}$ *and* $\{\overline{x}^k\}_{k\in\mathbb{N}}$ *generated by Algorithm 5, the following assertions are true.*

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

$$F(x^{k+1}) - F(x^k) \le \begin{cases} -\gamma\|x^k - \overline{x}^k\|^2 & \text{if } k \in \mathcal{K}_1, \\ -\gamma\|x^k - \overline{x}^k\|^{2+\sigma} & \text{if } k \in \mathcal{K}_2,\ \alpha_k = 1, \\ -\gamma\|x^k - \overline{x}^k\|^{2+2\sigma} & \text{if } k \in \mathcal{K}_2,\ \alpha_k \ne 1. \end{cases} \tag{4.40}$$

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

(iii) $\omega(x^0)$ *is nonempty and compact, and every element of* $\omega(x^0)$ *is an L-type stationary point of problem* (4.1).

Proof. **(i)** Fix any $k \in \mathcal{K}_2$. From inequalities (4.34) and (4.35),

$$\begin{aligned} f(x^{k+1}) - f(x^k) &\le -\frac{\varrho b_2\alpha_k}{2}\|\overline{\mu}_k(x^k - \overline{x}^k)\|^\sigma\|d^k\|^2 \le -\frac{\varrho c_1^2 b_2\alpha_k}{2}\|\overline{\mu}_k(x^k - \overline{x}^k)\|^{2+\sigma} \\ &\le -\frac{\varrho c_1^2 b_2\alpha_k\mu_{\min}^{2+\sigma}}{2}\|x^k - \overline{x}^k\|^{2+\sigma}, \end{aligned} \tag{4.41}$$

where the second inequality is using Proposition 4.2 (iii) and equality (4.37). By Remark 4.2 (b), we have $g(x^{k+1}) \le g(x^k)$, so that $F(x^{k+1}) - F(x^k) \le f(x^{k+1}) - f(x^k)$, which along with the last equation yields that

$$F(x^{k+1}) - F(x^k) \le -\frac{\varrho c_1^2 b_2\alpha_k\mu_{\min}^{2+\sigma}}{2}\|x^k - \overline{x}^k\|^{2+\sigma}.$$

Take $\gamma := \min\left\{\frac{\alpha}{2}, \frac{\varrho c_1^2 b_2\mu_{\min}^{2+\sigma}}{2}, \frac{\beta(1-\varrho)\varrho c_1^2 b_2^2\mu_{\min}^{2+2\sigma}}{2L_1}\right\}$. The desired result then follows by using Lemma 4.6 (ii) and recalling that $F(x^{k+1}) - F(x^k) \le \frac{\alpha}{2}\|x^k - \overline{x}^k\|^2$ for $k \in \mathcal{K}_1$.

(ii) Let $\widetilde{\mathcal{K}}_2 := \{k \in \mathcal{K}_2 \mid \alpha_k = 1\}$. Doing summation for inequality (4.40) from $k = 0$

to any $j \in \mathbb{N}$ yields that

$$\sum_{k\in\mathcal{K}_1\cap[j]} \gamma\|x^k - \overline{x}^k\|^2 + \sum_{k\in\widetilde{\mathcal{K}}_2\cap[j]} \gamma\|x^k - \overline{x}^k\|^{2+\sigma} + \sum_{k\in(\mathcal{K}_2\setminus\widetilde{\mathcal{K}}_2)\cap[j]} \gamma\|x^k - \overline{x}^k\|^{2+2\sigma}$$

$$\leq \sum_{k=0}^{j} \left[F(x^k) - F(x^{k+1})\right] = F(x^0) - F(x^{j+1}),$$

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

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

Thus, we obtain $\lim_{k\to\infty} \|x^k - \overline{x}^k\| = 0$. Together with Proposition 4.2 (iii), (4.37) and Remark 4.2 (d), it follows that $\lim_{\mathcal{K}_2\ni k\to\infty} \|d^k\| = 0$.

(iii) Recall that $\{x^k\}_{k\in\mathbb{N}} \subseteq \Omega$, so its accumulation point set $\omega(x^0)$ is nonempty. The compactness of $\omega(x^0)$ can be obtained by following the proof of (Bolte et al., 2014, Lemma 5(iii)). Pick any $x^* \in \omega(x^0)$. Then, there exists an index set $\mathcal{K} \subseteq \mathbb{N}$ such that $\lim_{\mathcal{K}\ni k\to\infty} x^k = x^*$. From part (ii), $\lim_{\mathcal{K}\ni k\to\infty} \overline{x}^k = x^*$. For each $k \in \mathcal{K}$, $\overline{x}^k \in \mathrm{prox}_{\overline{\mu}_k^{-1} g}\big(x^k - \overline{\mu}_k^{-1}\nabla f(x^k)\big)$ with $\overline{\mu}_k \in [\mu_{\min}, \widetilde{\mu})$ by step (1a) of Algorithm 5 and Remark 4.2 (d). We assume that $\lim_{\mathcal{K}\ni k\to\infty} \overline{\mu}_k = \overline{\mu}_* \in [\mu_{\min}, \widetilde{\mu}]$ (if necessary by taking a subsequence). Define the function

$$h(z,x,\mu) := \begin{cases} \frac{\mu}{2}\left\|z - (x - \mu^{-1}\nabla f(x))\right\|^2 + g(z) & \text{if } (z,x,\mu) \in \mathbb{R}^n \times \mathbb{R}^n \times [\mu_{\min}, \widetilde{\mu}], \\ \infty & \text{otherwise,} \end{cases}$$

and write $\mathcal{P}(x,\mu) := \arg\min_{z\in\mathbb{R}^n} h(z,x,\mu)$ for $(x,\mu) \in \mathbb{R}^n \times \mathbb{R}$. Note that $h : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \to \overline{\mathbb{R}}$ is a proper and lower semicontinuous function and is level-bounded in z locally uniformly in (x,μ). In addition, by (Rockafellar and Wets, 2009, Theorem 1.25), the function $\widehat{h}(x,\mu) := \inf_{z\in\mathbb{R}^n} h(z,x,\mu)$ is finite and continuous on $\mathbb{R} \times [\mu_{\min}, \widetilde{\mu}]$. From (Rockafellar and Wets, 2009, Example 5.22), the multifunction $\mathcal{P} : \mathbb{R}^n \times \mathbb{R} \rightrightarrows \mathbb{R}^n$ is outer semicontinuous relative to $\mathbb{R}^n \times [\mu_{\min}, \widetilde{\mu}]$. Note that

$\overline{x}^k \in \mathcal{P}(x^k, \overline{\mu}_k)$ for each $k \in \mathcal{K}$. Then, $x^* \in \mathrm{prox}_{\overline{\mu}_*^{-1} g}\big(x^* - \overline{\mu}_*^{-1} \nabla f(x^*)\big)$, so x^* is an L-type stationary point of (4.1). □

Next we use Lemma 4.7 (ii) to show that, after a finite number of iterations, the switch condition in (4.3) always holds and the Newton step is executed. To this end, define

$$T_k := \mathrm{supp}(Bx^k),\ \overline{T}_k := \mathrm{supp}(B\overline{x}^k),\ S_k := \mathrm{supp}(x^k) \ \text{ and } \ \overline{S}_k := \mathrm{supp}(\overline{x}^k). \quad (4.42)$$

Proposition 4.3. *For the index sets defined in* (4.42)*, there exist index sets* $T \subseteq [p], S \subseteq [n]$ *and an index* $\overline{k} \in \mathbb{N}$ *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* $x^* \in \omega(x^0)$*,* $\mathrm{supp}(Bx^*) = T$ *and* $\mathrm{supp}(x^*) = S$*.*

Proof. We complete the proof via the following three claims:

Claim 1: There exists $\overline{k} \in \mathbb{N}$ such that for $k > \overline{k}$, $|Bx^k|_{\min} \geq \frac{\nu}{2}$, where ν is the same as the one in Proposition 4.2 (ii).

Indeed, for each $k-1 \in \mathcal{K}_1$, $x^k = \overline{x}^{k-1}$, and $|Bx^k|_{\min} = |B\overline{x}^{k-1}|_{\min} \geq \nu > \frac{\nu}{2}$ follows by Proposition 4.2 (ii). Hence, it suffices to consider that $k-1 \in \mathcal{K}_2$. By Lemma 4.7 (ii), there exists $\overline{k} \in \mathbb{N}$ such that for all $k > \overline{k}$, $\|x^{k-1} - \overline{x}^{k-1}\| < \frac{\nu}{4\|B\|_2}$ and $\|d^{k-1}\| < \frac{\nu}{4\|B\|_2}$, 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 (4.3) 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 4.2 (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 3.1 (ii), $\mathrm{supp}(Bx^k) \subseteq \mathrm{supp}(Bx^{k-1})$ 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 $\mathcal{K}_2 \ni k-1 > \overline{k}-1$,

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

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 $T_k \subseteq \overline{T}_k$. 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 $\overline{T}_k \subseteq T_k$. 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**. 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 $T_k \subseteq T_{k+1}$. 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 $T_{k+1} \subseteq T_k$.

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

Pick any $x^* \in \omega(x^0)$. Let $\{x^k\}_{k\in\mathcal{K}}$ be a subsequence such that $\lim_{\mathcal{K}\ni k\to\infty} x^k = x^*$. By the above proof, for all sufficiently large $k \in \mathcal{K}$, $|Bx^k|_{\min} \geq \frac{\nu}{2}$ 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**. The proof is completed. □

By Proposition 4.3, $k \in \mathcal{K}_2$ for all $k > \overline{k}$, 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 $\Pi_k = \Pi_* := \Pi_{\overline{k}+1}$ for all $k > \overline{k}$, Algorithm 5 finally reduces to the inexact projected regularized Newton method for solving

$$\min_{x\in\mathbb{R}^n} \ \Psi(x) := f(x) + \delta_{\Pi_*}(x), \tag{4.43}$$

which is a minimization problem of function f over polyhedron Π_*, much simpler than the original problem (4.1). Consequently, the global convergence and local convergence rate analysis of PGiPN for model (4.1) boils down to analyzing those of the inexact projected regularized Newton method for (4.43). The rest of this section is devoted to this. Unless otherwise stated, the notation $\overline{k}$ in the sequel is always the same as that of Proposition 4.3. In addition, we require the assumption that $\nabla^2 f$ is locally Lipschitz continuous on $\omega(x^0)$.

Assumption 4.1. *$\nabla^2 f$ is locally Lipschitz continuous on an open set $\mathcal{O} \supseteq \omega(x^0)$.*

Assumption 4.1 is very standard when analyzing the convergence behavior of Newton-type method. In fact, if f is assumed to be third time continuously differentiable on $\mathbb{R}^n$, this assumption directly holds. 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 (Liu et al., 2022, Lemma B.1), the details are omitted here.

Lemma 4.8. *Suppose that Assumption 4.1 holds. Then $\alpha_k = 1$ for sufficiently large k.*

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

$$\Pi_* = \big\{x \in \mathbb{R}^n \mid B_{T^c_{\overline{k}+1}\cdot}x = 0,\ x_{S^c_{\overline{k}+1}} = 0,\ x \geq l_b,\ -x \geq -u_b\big\}. \tag{4.44}$$

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

$$\mathcal{A}(x) := \{i \mid x_i = (l_b)_i\} \cup \{i + n \mid x_i = (u_b)_i\}.$$

Clearly, for $x \in \Pi_*$, $\mathcal{A}(x)$ is the active set of constraint Π_* at x. To prove the global convergence for PGiPN, we first show that $\{\mathcal{A}(x^k)\}_{k\in\mathbb{N}}$ remains stable when k is sufficiently large, under the following non-degeneracy assumption.

Assumption 4.2. *For all $x^* \in \omega(x^0)$, $0 \in \nabla f(x^*) + \mathrm{ri}(\mathcal{N}_{\Pi_*}(x^*))$.*

It follows from Proposition 2.1 and Lemma 4.7 (iii) that for each $x^* \in \omega(x^0)$, x^* is a critical point of F, which together with Proposition 4.3 and Lemma 4.1 (i) yields that $0 \in \nabla f(x^*) + \mathcal{N}_{\Pi_*}(x^*)$, so that Assumption 4.2 substantially requires that $-\nabla f(x^*)$ does not belong to the relative boundary[1] of $\mathcal{N}_{\Pi_*}(x^*)$. In the next lemma, we prove that under Assumptions 4.1-4.2, $\mathcal{A}(x^k) = \mathcal{A}(x^{k+1})$ for sufficiently large k.

Lemma 4.9. *Let $\{x^k\}_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 5. Suppose that Assumptions 4.1-4.2 hold. Then, there exist $\mathcal{A}^* \subseteq [2n]$ and a closed and convex cone $\mathcal{N}^* \subseteq \mathbb{R}^n$ 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,

$$\|\mathrm{proj}_{\mathcal{T}_{\Pi_*}(x^k)}(-\nabla f(x^k))\| = \| -\nabla f(x^k) - \mathrm{proj}_{\mathcal{N}_{\Pi_*}(x^k)}(-\nabla f(x^k))\| = \mathrm{dist}(0, \partial\Psi(x^k)).$$

Therefore, it suffices to prove that $\lim_{k\to\infty} \mathrm{dist}(0, \partial\Psi(x^k)) = 0$. By Proposition 4.3, equation (4.9), Remark 3.1 (d) and Lemma 4.7 (ii), there exists $\{\zeta_k\}_{k>\overline{k}}$ with

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

$\lim_{k\to\infty}\|\zeta_k\| = 0$ such that $0 \in \nabla f(x^k) + G_k d^k + \zeta_k + \mathcal{N}_{\Pi_k}(x^k + d^k)$ for each $k \geq \overline{k}$, which implies that $\nabla f(x^k + d^k) - \nabla f(x^k) - G_k d^k - \zeta_k \in \partial\Psi(x^k + d^k)$ for each $k \geq \overline{k}$. This together with Lemma 4.7 (ii) and the continuity of ∇f implies that $\lim_{k\to\infty} \mathrm{dist}(0, \partial\Psi(x^k + d^k)) = 0$. Thus, by Lemma 4.8 we obtain the desired result.

Claim 2: $\mathcal{A}(x^k) \subseteq \mathcal{A}(x^{k+1})$ for all sufficiently large k.

We prove by contradiction. If this claim does not hold, there exists $\mathcal{K} \subseteq \mathbb{N}$ such that $\mathcal{A}(x^k) \nsubseteq \mathcal{A}(x^{k+1})$ for all $k \in \mathcal{K}$. If necessary taking a subsequence, we assume that $\{x^k\}_{k\in\mathcal{K}}$ converges to x^*. By Lemma 4.7 (ii), $\{x^{k+1}\}_{k\in\mathcal{K}}$ converges to x^*. In addition, from **Claim 1** it follows that $\lim_{k\to\infty}\|\mathrm{proj}_{\mathcal{T}_{\Pi_*}(x^{k+1})}(-\nabla f(x^{k+1}))\| = 0$. The two sides along with Assumption 4.2 and (Burke and Moré, 1988, Corollary 3.6) yields that $\mathcal{A}(x^{k+1}) = \mathcal{A}(x^*)$ for all sufficiently large $k \in \mathcal{K}$. Since $\mathcal{A}(x^k) \subseteq \mathcal{A}(x^*)$ for sufficiently large $k \in \mathcal{K}$, we have $\mathcal{A}(x^k) \subseteq \mathcal{A}(x^{k+1})$ for sufficiently large $k \in \mathcal{K}$, contradicting to $\mathcal{A}(x^k) \nsubseteq \mathcal{A}(x^{k+1})$ for $k \in \mathcal{K}$. The claimed fact that $\mathcal{A}(x^k) \subseteq \mathcal{A}(x^{k+1})$ then follows.

From $\mathcal{A}(x^k) \subseteq \mathcal{A}(x^{k+1})$ for all sufficiently large k, $\{\mathcal{A}(x^k)\}_{k\in\mathcal{K}_2}$ converges to for some $A^* \subseteq [2n]$ in the sense of Painlevé-Kuratowski. From the finiteness of $\mathcal{A}^*$, we conclude that $\mathcal{A}(x^k) = \mathcal{A}^*$ for all sufficiently large k. From the expression of Π_* in (4.44) and $\mathcal{A}(x^k) = \mathcal{A}^*$ for all sufficiently large k, we have $\mathcal{N}_{\Pi_*}(x^k) = \mathcal{N}^*$ for all sufficiently large k. □

Our proof for the global convergence of PGiPN additionally requires the following assumption.

Assumption 4.3. *For every sufficiently large k, there exists $\xi_k \in \mathcal{N}_{\Pi_*}(x^k)$ such that*

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

This assumption essentially requires that for every sufficiently large k there exists 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 every 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 $\langle \xi^k, d^k \rangle \le 0$. Together with (4.35), for every 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 4.3 to hold. Obviously, when $n = 1$, it automatically holds.

Next, we show that if Ψ is a KL function and Assumptions 4.1-4.3 hold, the sequence generated by PGiPN is Cauchy and converges to an L-type stationary point.

Theorem 4.1. *Let $\{x^k\}_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 5. Suppose that Assumptions 4.1-4.3 hold, and that Ψ 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-type stationary point of* (4.1).

Proof. If there exists $k > \overline{k}$ such that $\Psi(x^k) = \Psi(x^{k+1})$, then $F(x^k) = F(x^{k+1})$ by Proposition 4.3, which together with Lemma 4.7 (i) yields that $x^k = \overline{x}^k$. Consequently, x^k meets the termination condition of Algorithm 5, so that $\{x^k\}_{k\in\mathbb{N}}$ converges to an L-type stationary point of (4.1) within a finite number of steps. Thus, we only need to consider the case that $\Psi(x^k) > \Psi(x^{k+1})$ for all $k > \overline{k}$. By (Bolte et al., 2014, Lemma 6), there exist $\varepsilon > 0, \eta > 0$ and a continuous concave function $\varphi \in \Upsilon_\eta$ (see Definition 2.4) such that for all $\overline{x} \in \omega(x^0)$ and $x \in \{z \in \mathbb{R}^n \mid \mathrm{dist}(z, \omega(x^0)) < \varepsilon\} \cap [\Psi(\overline{x}) < \Psi < \Psi(\overline{x}) + \eta]$,

$$\varphi'(\Psi(x) - \Psi(\overline{x}))\mathrm{dist}(0, \partial\Psi(x)) \ge 1,$$

where $\omega(x^0)$ is defined in Lemma 4.7 (iii). Clearly, $\lim_{k\to\infty} \mathrm{dist}(x^k, \omega(x^0)) = 0$. Pick any $x^* \in \omega(x^0)$. By the definition of Ψ, Propositions 4.2 (i) and 4.3, we have $\lim_{k\to\infty} \Psi(x^k) = \Psi(x^*)$. Then, for $k > \overline{k}$ (if necessary by increasing $\overline{k}$), $x^k \in \{z \in \mathbb{R}^n \mid \mathrm{dist}(z, \omega(x^0)) < \varepsilon\} \cap [\Psi(x^*) < \Psi < \Psi(x^*) + \eta]$. Consequently, for all $k > \overline{k}$,

$$\varphi'(\Psi(x^k) - \Psi(x^*))\mathrm{dist}(0, \partial\Psi(x^k)) \ge 1. \tag{4.45}$$

By Assumption 4.3, there exist $c_{\min} > 0$ and $\xi_k \in \mathcal{N}_{\Pi_*}(x^k)$ such that for all suffciently large k,

$$-\langle \nabla f(x^k) + \xi_k, d^k \rangle > c_{\min} \|\nabla f(x^k) + \xi_k\| \|d^k\|. \tag{4.46}$$

From Lemma 4.9 we have $\mathcal{N}_{\Pi_*}(x^k) = \mathcal{N}_{\Pi_*}(x^{k+1})$ for all $k > \overline{k}$ (by possibly enlarging $\overline{k}$), which implies that $\xi_k \in \mathcal{N}_{\Pi_*}(x^{k+1})$. Together with (4.34), (4.46) and Lemma 4.8, we have that for all $k > \overline{k}$ (if necessary enlarging $\overline{k}$),

$$\begin{aligned} \frac{\Psi(x^k) - \Psi(x^{k+1})}{\mathrm{dist}(0, \partial\Psi(x^k))} &\geq \frac{-\varrho\langle \nabla f(x^k) + \xi_k, d^k \rangle}{\mathrm{dist}(0, \partial\Psi(x^k))} \\ &\geq \frac{\varrho c_{\min} \|\nabla f(x^k) + \xi_k\| \|d^k\|}{\|\nabla f(x^k) + \xi_k\|} = \varrho c_{\min} \|x^{k+1} - x^k\|, \end{aligned} \tag{4.47}$$

where the second inequality follows by $\nabla f(x^k) + \xi_k \in \partial\Psi(x^k)$ and (4.46). For each k, let $\Delta_k := \varphi(\Psi(x^k) - \Psi(x^*))$. From (4.45), (4.47) and the concavity of φ on $[0, \eta)$, for all $k > \overline{k}$,

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

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

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

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 4.7 (iii) that x^* is an L-type stationary point of problem (4.1). The proof is completed. □

If Ψ has the KL property of exponent $1/2$ and Assumptions 4.1, 4.2 are satisfied, both $\{F(x^k)\}_{k\in\mathbb{N}}$ and $\{x^k\}_{k\in\mathbb{N}}$ converge at a linear rate.

Theorem 4.2. *Suppose that $\{x^k\}_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 5, and Assumptions 4.1, 4.2 hold. Pick any $x^* \in \omega(x^0)$. If Ψ has the KL property with exponent $\frac{1}{2}$ at x^*, $\{F(x^k)\}_{k\in\mathbb{N}}$ converges to $F(x^*)$ at a Q-linear rate, and $\{x^k\}_{k\in\mathbb{N}}$ converges to x^* at an R-linear convergence rate.*

Proof. Let $\overline{k}$ be such that the results in Proposition 4.3 and Lemmas 4.8, 4.9 hold for $k > \overline{k}$. For any $k > \overline{k}$, by equations (4.9) and (4.37) there exist $\xi_k \in \mathcal{N}_{\Pi_*}(x^k + d^k)$ and $\zeta_k \in \mathbb{R}^n$ with $\|\zeta_k\| \leq \frac{1}{2}\|r_k(x^k)\|^{1+\varsigma}$ such that $\nabla f(x^k) + G_k d^k + \xi_k + \zeta_k = 0$, which implies that $d^k = -G_k^{-1}(\nabla f(x^k) + \xi_k + \zeta_k)$. From Lemma 4.9, (Li and Pong, 2018, Lemma 4.1) and (Sra, 2012, Lemma 4), we have

$$\|\nabla f(x^k) + \xi_k\| \geq \mathrm{dist}(0, \partial\Psi(x^k)) \geq \min\{\overline{\mu}_k^{-1}, 1\}\|r_k(x^k)\|. \tag{4.48}$$

Since $\varsigma > \sigma$, the above inequality along with Remark 4.2 (d), (4.37), Lemma 4.7(ii) yields

$$\lim_{k\to\infty} \frac{\|r_k(x^k)\|^{1+\varsigma-\sigma}}{\|\nabla f(x^k) + \xi_k\|} \leq (\min\{\overline{\mu}_k^{-1}, 1\})^{-1} \lim_{k\to\infty} \|r_k(x^k)\|^{\varsigma-\sigma} = 0. \tag{4.49}$$

Then, we have for $k > \overline{k}$ (if necessary enlarging $\overline{k}$),

$$\begin{aligned}
&-\langle \nabla f(x^k) + \xi_k, d^k\rangle = \langle \nabla f(x^k) + \xi_k, G_k^{-1}(\nabla f(x^k) + \xi_k + \zeta_k)\rangle \\
&\geq \overline{c}^{-1}\|\nabla f(x^k) + \xi_k\|^2 - (b_2\|r_k(x^k)\|^\sigma)^{-1}\|\nabla f(x^k) + \xi_k\|\|\zeta_k\| \\
&\geq \overline{c}^{-1}\|\nabla f(x^k) + \xi_k\|^2 - \frac{\|\nabla f(x^k) + \xi_k\|\|r_k(x^k)\|^{1+\varsigma}}{2b_2\|r_k(x^k)\|^\sigma} \\
&\geq \overline{c}^{-1}\|\nabla f(x^k) + \xi_k\|^2\Big(1 - \frac{\overline{c}\|r_k(x^k)\|^{1+\varsigma-\sigma}}{2b_2\|\nabla f(x^k) + \xi_k\|}\Big) \geq (2\overline{c})^{-1}\|\nabla f(x^k) + \xi_k\|^2,
\end{aligned} \tag{4.50}$$

where the first inequality uses (4.38), the second inequality follows by the definition of ζ_k, and the last inequalities use (4.49). From (4.34), Proposition 4.3 and Lemmas 4.8, 4.9, it holds that

$$\Psi(x^k) - \Psi(x^{k+1}) \geq \frac{\varrho}{2\overline{c}}\mathrm{dist}(0, \partial\Psi(x^k))^2, \tag{4.51}$$

for $k > \overline{k}$. Since Ψ has the KL property of exponent $1/2$, following a discussion similar to the proof of Theorem 3.1, we have that there exist $c > 0$ and $x^* \in \omega(x^0)$ such that for all $k > \overline{k}$ (if necessary enlarging $\overline{k}$),

$$\frac{c}{2}(\Psi(x^k) - \Psi(x^*))^{-1/2}\mathrm{dist}(0, \partial\Psi(x^k)) \geq 1.$$

Let $\bar{\Delta}_k = \Psi(x^k) - \Psi(x^*)$ for each $k > \overline{k}$. Then, it follows from (4.51) that

$$4c^{-2} \leq [\bar{\Delta}_k^{-1/2}\mathrm{dist}(0, \partial\Psi(x^k))]^2 \leq \frac{2\overline{c}}{\varrho}\bar{\Delta}_k^{-1}(\Psi(x^k) - \Psi(x^{k+1})) = \frac{2\overline{c}}{\varrho}\bar{\Delta}_k^{-1}(\bar{\Delta}_k - \bar{\Delta}_{k+1}),$$

which implies that $\Psi(x^{k+1}) - \Psi(x^*) \leq \widehat{c}(\Psi(x^k) - \Psi(x^*))$ with $\widehat{c} := \frac{\overline{c}c^2 - 2\varrho}{\overline{c}c^2}$. That is, $\{\Psi(x^k) - \Psi(x^*)\}_{k > \overline{k}}$ converges to 0 at a Q-linear convergence rate. If necessary enlarging $\overline{k}$, we have for $k > \overline{k}$,

$$\sum_{j=k}^{\infty}\|x^j - x^{j+1}\| = \sum_{j=k}^{\infty}\|d^j\| \leq \sum_{j=k}^{\infty} c_2\widetilde{\mu}^{1-\sigma}\|x^j - \overline{x}^j\|^{1-\sigma}$$

$$\leq \sum_{j=k}^{\infty} c_2\widetilde{\mu}^{1-\sigma}(\gamma^{-1}(\Psi(x^j) - \Psi(x^{j+1})))^{\frac{1-\sigma}{2+2\sigma}} \leq \sum_{j=k}^{\infty} c_2\widetilde{\mu}^{1-\sigma}(\gamma^{-1}(\Psi(x^j) - \Psi(x^*)))^{\frac{1-\sigma}{2+2\sigma}}$$

$$\leq \sum_{j=k}^{\infty} c_2\widetilde{\mu}^{1-\sigma}\gamma^{\frac{\sigma-1}{2+2\sigma}}\widehat{c}^{\frac{(1-\sigma)(j-k)}{2+2\sigma}}(\Psi(x^k) - \Psi(x^*))^{\frac{1-\sigma}{2+2\sigma}} \leq \frac{c_2\widetilde{\mu}^{1-\sigma}\gamma^{\frac{\sigma-1}{2+2\sigma}}}{1 - \widehat{c}^{\frac{(1-\sigma)}{2+2\sigma}}}(F(x^0) - \Psi(x^*))^{\frac{1-\sigma}{2+2\sigma}},$$

where the first equality holds by Lemma 4.8, the first inequality follows by equation (4.37), Proposition 4.2(iii) and Remark 4.2 (d), the second inequality uses Lemma 4.7 (i), and the last inequality holds by $F(x^0) \geq F(x^k) \geq \Psi(x^k)$. Therefore, we conclude that $\{x^k\}_{k\in\mathbb{N}}$ is a convergent sequence. By noting that $x^* \in \omega(x^0)$, $\{x^k\}_{k\in\mathbb{N}}$ converges to x^*. Since $\|x^k - x^*\| \leq \sum_{j=k}^{\infty}\|x^j - x^{j+1}\|$, by the above group of inequalities we conclude that $\{x^k\}_{k\in\mathbb{N}}$ converges to x^* at an R-linear convergence rate. □

We now focus on the superlinear rate analysis of PGiPN. Denote

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

which we call by the set of second-order stationary points of (4.43). Based on this notation, we assume that a local Hölderian error bound condition holds with $\mathcal{X}^*$ in Assumption 4.4. 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 4.4. *The q-subregularity of function* $r(x) := x - \mathrm{proj}_{\Pi_*}(x - \nabla f(x))$ *holds at* x^* *for the origin with* $\mathcal{X}^*$, *i.e., there exist* $\varepsilon > 0,\ \kappa > 0$ *and* $q \in (0,1]$ *such that for all* $x \in \mathbb{B}(x^*,\varepsilon) \cap \Pi_*$, $\mathrm{dist}(x, r^{-1}(0)) = \mathrm{dist}(x, \mathcal{X}^*) \le \kappa \|r(x)\|^q$.

Recently, Liu et al. (2024) proposed an inexact regularized proximal Newton method (IRPNM) for solving the problems, consisting of a smooth function and an extended real-valued convex function, which includes (4.43) as a special case. They studied the superlinear convergence rate of IRPNM under Assumptions 4.1 and 4.4. 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 4.4 implies that there exists $\widehat{\kappa} > 0$ such that for sufficiently large k with $x^k \in \mathbb{B}(x^*, \varepsilon)$,

$$\mathrm{dist}(x^k, \mathcal{X}^*) \le \widehat{\kappa} \|r_k(x^k)\|^q. \tag{4.52}$$

Recall that PGiPN finally reduces to an inexact projected regularized Newton method for solving (4.43). From Lemma 4.6 (iii) and Lemma 4.8, we have

$$\Theta_k(x^{k+1}) - \Theta_k(x^k) \le 0 \quad \text{and} \quad \|R_k(x^{k+1})\| \le \frac{1}{2} \min\{\|r_k(x^k)\|, \|r_k(x^k)\|^{1+\varsigma}\}, \tag{4.53}$$

for sufficiently large k. Let $\Lambda_k^i := G_k^i - \nabla^2 f(x^k) - b_2 \|\overline{\mu}_k(x^k - \overline{x}^k)\|^\sigma I$, where G_k^i are those in (4.30)-(4.32). Under Assumption 4.4, from Lemma 3.11 and (Liu et al., 2024, Lemma 7) and the fact that $G_k^1 - G_k^2 \succeq 0$, we have for sufficiently large k with $x^k \in \mathbb{B}(x^*, \varepsilon)$,

$$\max\left\{\lambda_{\min}(\Lambda_k^1), \lambda_{\min}(\Lambda_k^2), \lambda_{\min}(\Lambda_k^3)\right\} = O(\mathrm{dist}(x^k, \mathcal{X}^*)), \tag{4.54}$$

By using (4.52), (4.53) and (4.54), and following a proof similar to (Liu et al., 2024, Theorem 6), we can obtain the following result.

Theorem 4.3. *Let $\{x^k\}_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 5. Suppose that Assumption 4.1 holds, and that $\{x^k\}_{k\in\mathbb{N}}$ converges to $x^* \in \omega(x^0)$. If Assumption 4.4 holds with $q \in (\frac{1}{1+\sigma}, 1]$ at x^*, then the sequence $\{x^k\}_{k\in\mathbb{N}}$ converges to x^* with the Q-superlinear convergence rate at order $q(1+\sigma)$.*

4.4 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 5 in Section 4.4.1. In Section 4.4.2, we make comparison between model (4.1) with the least-squares loss function f and the fused Lasso model (1.6) 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`. We note that problem (4.43) with f considered in this subsection satisfies KL property, see Proposition 2.3. Finally, in Section 4.4.3, we present some numerical results toward the comparison among several variants of PGiPN and ZeroFPR and PG method for (4.1) in terms of efficiency and the quality of the output. The MATLAB code of PGiPN is available at `https://github.com/yuqiawu/PGiPN`.

4.4.1 Implementation of Algorithm 5

Dimension reduction of (4.6)

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 (4.6) by solving a problem in a lower dimension. Specifically, for each $k \in \mathcal{K}_2$, write

$$H^k := (G_k)_{S_k S_k},\ u^k := x^k_{S_k},\ \nabla f_{S_k}(u^k) = [\nabla f(x^k)]_{S_k},$$

$$\widehat{\Pi}_k := \{u \in \mathbb{R}^{|S_k|} \mid \widetilde{B}_k u = 0,\ (l_b)_{S_k} \le u \le (u_b)_{S_k}\},$$

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

$$\widehat{u}^k \approx \underset{u\in\mathbb{R}^{|S_k|}}{\arg\min} \left\{\theta_k(u) := f(I_{\cdot S_k} u^k) + \langle \nabla f_{S_k}(u^k), u - u^k\rangle + \frac{1}{2}(u - u^k)^\top H^k (u - u^k) + \delta_{\widehat{\Pi}_k}(u)\right\}. \tag{4.55}$$

The following lemma gives a way to find y^k satisfying (4.8)-(4.9) by inexactly solving problem (4.55), whose dimension is much smaller than that of (4.6) if $|S_k| \ll n$.

Lemma 4.10. *Let* $y^k_{S_k} = \widehat{u}^k$ *and* $y^k_{S_k^c} = 0$. *Then,* $\Theta_k(y^k) = \theta_k(\widehat{u}^k)$ *and* $\mathrm{dist}(0, \partial\Theta_k(y^k)) = \mathrm{dist}(0, \partial\theta_k(\widehat{u}^k))$. *Consequently, the vector* $\widehat{u}^k$ *satisfies*

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

if and only if the vector y^k *satisfies the inexact conditions in* (4.8)-(4.9).

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, $\mathcal{N}_{\Pi_k}(y^k) = \mathrm{Range}(B^\top_{T_k^c \cdot}) + \mathrm{Range}(I^\top_{S_k^c \cdot}) + \mathcal{N}_\Omega(y^k)$, and

$$\begin{aligned}
&\mathrm{dist}(0, \partial\Theta_k(y^k)) \\
&= \mathrm{dist}\big(0, \nabla f(x^k) + G_k(y^k - x^k) + \mathrm{Range}(B^\top_{T_k^c \cdot}) + \mathrm{Range}(I^\top_{S_k^c \cdot}) + \mathcal{N}_\Omega(y^k)\big) \\
&= \mathrm{dist}\big(0, \nabla f_{S_k}(u^k) + H^k(\widehat{u}^k - u^k) + \mathrm{Range}(B^\top_{T_k^c S_k}) + \mathcal{N}_{[(l_b)_{S_k}, (u_b)_{S_k}]}(\widehat{u}^k)\big) \\
&= \mathrm{dist}(0, \nabla f_{S_k}(u^k) + H^k(\widehat{u}^k - u^k) + \mathcal{N}_{\widehat{\Pi}_k}(\widehat{u}^k)) = \mathrm{dist}(0, \theta_k(\widehat{u}^k)),
\end{aligned}$$

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

Acceleration of Algorithm 1

The switch condition in (4.3) is in general difficult to be satisfied when $\|Bx^k\|_0$ or $\|x^k\|_0$ is large. Consequently, PGiPN is continuously executing PG steps. This phenomenon is evident in the numerical experiment of the restoration of blurred image, see Section 4.4.3. To further accelerate the iterates of Algorithm 5 into the Newton step, we introduce the following relaxed switch condition:

$$\||\mathrm{sign}(Bx^k)| - |\mathrm{sign}(B\overline{x}^k)|\|_1 \leq \frac{\eta_1 n}{k} \quad \text{and} \quad \||\mathrm{sign}(x^k)| - |\mathrm{sign}(\overline{x}^k)|\|_1 \leq \frac{\eta_2 n}{k}, \tag{4.56}$$

where η_1, η_2 are two nonnegative constants. Following the arguments similar to those in Lemma 4.6, we have that Algorithm 5 equipped with (4.56) is also well defined. Obviously, when $\frac{\eta_i n}{k} \geq 1$, condition (4.56) 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 (4.56) is identical to (4.3). This means that as k grows, Algorithm 5 with relaxed switch condition (4.56) will finally reduce to the one with (4.3). Since our convergence analysis does not specify the initial point, the asymptotic convergence results also hold for Algorithm 5 with condition (4.56).

Choice of parameters in Algorithm 5

We will test the performance of PGiPN with G_k given by G_k^2 in (4.31), and PGiPN(r), which is PGiPN with relaxed switch condition (4.56). We use Gurobi to solve subproblem (4.6) with such G_k, with inexact conditions (4.8), (4.9) controlled by options `params.Cutoff` and `params.OptimalityTol`, respectively. Also, we test PGilbfgs, which is the same as PGiPN, except using limited-memory BFGS (lbfgs) to construct G_k. In particular, we form $G_k = B_k + b_2\|\overline{\mu}_k(x^k - \overline{x}^k)\|^\sigma$, with B_k given by lbfgs. For solving (4.6) with such G_k, we use the method introduced in (Kanzow and Lechner (2022)). We set the parameters of all the variants of PGiPN by

$\alpha = 10^{-8},\ \tau = 2,\ \sigma = \frac{1}{2},\ \varrho = 10^{-4},\ \beta = \frac{1}{2},\ \varsigma = \frac{2}{3}$. We set $b_2 = 10^{-3}$ for PGiPN and PGiPN(r), and $b_2 = 10^{-8}$ for PGilbfgs.

We compare the numerical performance of our algorithms with those of ZeroFPR (Themelis et al. (2018)) and the PG method (Wright et al. (2009)). In particular, ZeroFPR uses the quasi-Newton method to minimize the forward-backward envelope of the objective. The code package of ZeroFPR is downloaded from `http://github.com/kul-forbes/ForBES`. We set "lbfgs" as the solver of ZeroFPR. On the other hand, the iterate 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 5. To be fair, as the default setting in ZeroFPR, in all variants of PGiPN and PG, we set $\mu_k = 0.95^{-1}L_1$ for all $k \in \mathbb{N}$, where L_1 is an estimation of the Lipschitz constant of ∇f obtained by computing $\|A\|_2$ from the following 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.

4.4.2 Model Comparison with the Fused Lasso

This subsection is devoted to the numerical comparison between the fused ℓ_0-norms regularization problem with a box constraint (FZNS), i.e., model (4.1) with $f = \frac{1}{2}\|A\cdot - b\|^2$ and $B = \widehat{B}$ and the fused Lasso (1.6). We apply PGiPN to solve FZNS and SSNAL to solve (1.6). Since the solved models are different, we only compare the quality of solutions returned by these two solvers, and will not compare their running time.

Our first empirical study focuses on the ability of regression. For this purpose, we use a commonly used dataset, prostate data, which can be downloaded from `https://hastie.su.domains/ElemStatLearn/`. 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, which composes $A \in \mathbb{R}^{50\times 8}$. The corresponding responses are represented by $b \in \mathbb{R}^{50}$. The reminders are left as 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 (1.6), with (A, b) given above, and $l_b = -1000 \times \mathbf{1}$, $u_b = 1000 \times \mathbf{1}$. 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 later one is defined as $\|\bar{A}x^* - \bar{b}\|$ with x^* being the output. The above procedure is repeated for 100 randomly constructed (A, b), 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, 3 and 5 in Table 4.1. For each sparsity pair, the mean testing errors of $\|\bar{A}x^* - \bar{b}\|$ for PGiPN and SSNAL corresponding to the given pair is recorded in lines 2, 4 and 6 in Table 4.1. Among others, since the fused Lasso may produce solutions with components being very small but not equal to 0, we define $\|y\|_0 := \min\{k \mid \sum_{i=1}^{k} |\hat{y}| \geq 0.999\|y\|_1\}$ as in (Li et al. (2018)), where $\hat{y}$ is obtained by sorting y in a nonincreasing order, for the outputs of the fused Lasso. As shown in Table 4.1, it is evident that when $(\|\widehat{B}x^*\|_0, \|x^*\|_0) = (6, 6)$, the mean testing error for FZNS is the smallest among all the testing examples. Furthermore, in the presented 21 comparative experiments, the fused Lasso outperforms FZNS for only 8 cases. Among these 8 experiments, in 7 cases, $\|\widehat{B}x^*\|_0 \geq 4$ and $\|x^*\|_0 \geq 6$.

This indicates that our model performs better when the solution is relatively sparse.

Table 4.1: Mean 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)
Mean testing error	8.35|8.54	7.34|7.36	5.45|5.15	5.15|5.74	5.21|6.32	5.08|5.27	5.11|5.70
$(\|\widehat{B}x^*\|_0, \|x^*\|_0)$	(4,4)	(4,5)	(4,6)	(4,7)	(4,8)	(5,5)	(5,6)
Mean testing error	5.07|5.52	5.31|5.86	5.49|4.99	5.25|4.97	5.33|4.78	5.10|5.48	5.74|5.38
$(\|\widehat{B}x^*\|_0, \|x^*\|_0)$	(5,7)	(5,8)	(6,6)	(6,7)	(6,8)	(7,7)	(7,8)
Mean testing error	5.46|5.58	5.35|5.19	4.41| 5.26	5.34|4.95	5.25|5.34	5.03|5.22	5.24|5.22

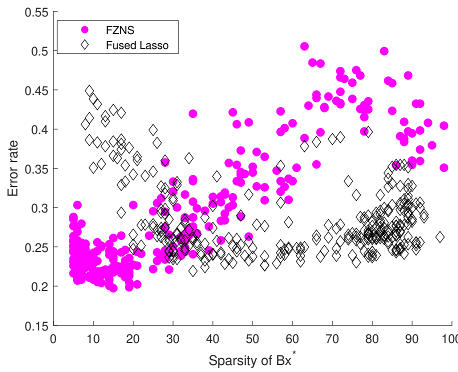
Our second numerical study is to evaluate the classification ability of the two models using the TIMIT database. As introduced in Section 1.2, the TIMIT database is a widely used resource for research in speech recognition. Following the approach described in (Land and Friedman (1997)), we compute a 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 $\lambda_2 = 0,\ l_b = -\mathbf{1}$ and $u_b = \mathbf{1}$, or in other words, the zero order variable fusion (1.4) plus a box constraint and the first order variable fusion (1.5).

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.

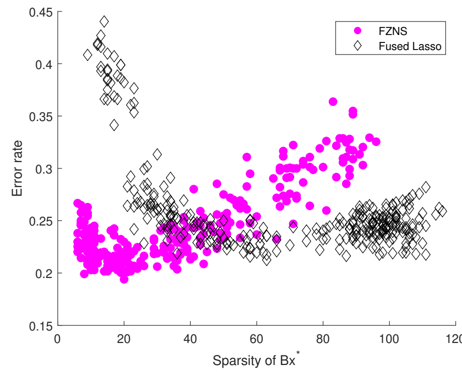
For (A, b), given 10 λ_1's randomly selected within $[2\times 10^{-5}, 300]$ such that the sparsity of the outputs $\|\widehat{B}x^*\|_0$ spans a wide range. If $\bar{A}_{i\cdot}x^* \le 1.5$, this phoneme is classified as "aa" and hence we set $\hat{b}_i = 1$; otherwise, $\hat{b}_i = 2$. If $\hat{b}_i \ne \bar{b}_i$, $A_{i\cdot}$ 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 figures in Figure 4.1 present $\|\widehat{B}x^*\|_0$ and the error rate for each output, with 4 different choices of (m_1, m_2). We can see that, for each figure the output with the smallest error rate is always achieved by the fused ℓ_0-norms regularization model. It is apparent that in general, FZNS performs better than the fused Lasso when $\|\widehat{B}x^*\|_0 \le 30$, while the mean error rate of the fused Lasso is lower than that of FZNS when $\|\widehat{B}x^*\|_0 \ge 60$. This phenomenon is especially evident when m_1 and m_2 are small.

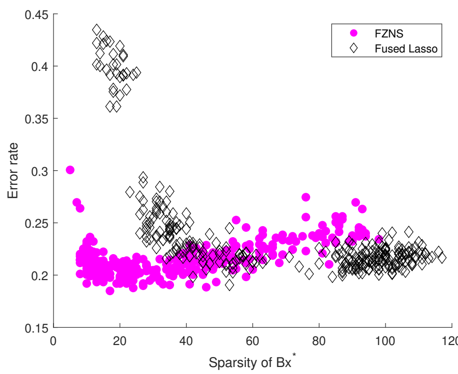
Based on the results of these two empirical studies, we deduce that the fused ℓ_0-norms regularization tends to outperform the fused Lasso regularization model when the output is sufficiently sparse. However, it is important to note that the numerical performance of the fused ℓ_0-norms regularization is not stable if the output is not sparse, especially when the number of observations is small, which suggests that when employing the fused ℓ_0-norms regularization, careful consideration should be given to selecting an appropriate penalty parameter. Moreover, due to the fact that 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.



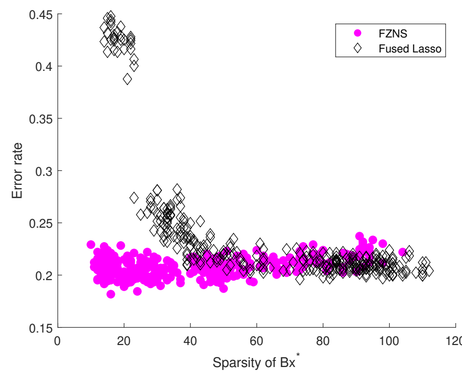
(a) $m_1 = 25, m_2 = 50$



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



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



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

Figure 4.1: $\|\widehat{B}x^*\|_0$ and the classification error rate for the outputs from FZNS and the fused Lasso under different m_1, m_2.

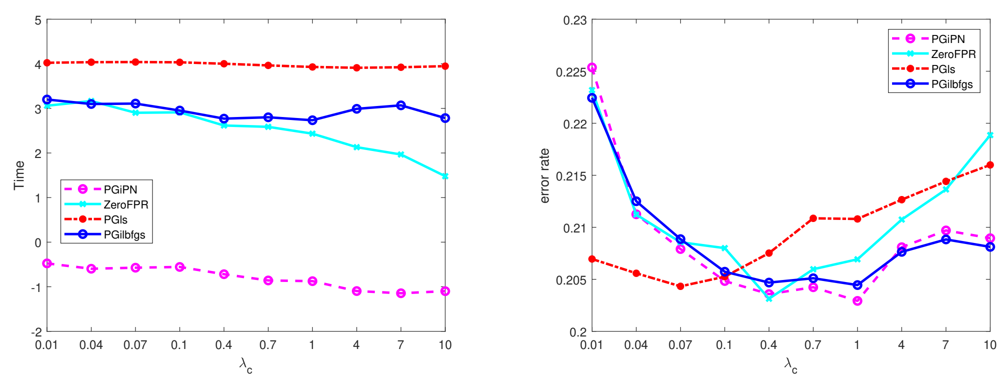
4.4.3 Comparison with ZeroFPR and PG

This subsection focuses on the comparison among several variants of PGiPN, ZeroFPR and PG, in terms of efficiency and the quality of the outputs.

Classification of TIMIT

The experimental data used in this part is the TIMIT dataset, the one in Section 4.4.2. To test the performance of the algorithms on (4.1) with nonconvex f, we consider solving model (4.1) with $f = \sum_{i=1}^{m} \log\left(1 + \frac{(A\cdot - b)_i}{\nu}\right)$, $B = \widehat{B}$, $l_b = -\mathbf{1}$ and $u_b = \mathbf{1}$, where $A \in \mathbb{R}^{m\times n}$ represents the training data and $b \in \mathbb{R}^m$ is the vector of the corresponding labels. It is worth noting that the loss function is nonconvex, and it was claimed in (Aravkin et al. (2012)) that this loss function is effective to process data denoised by heavy-tailed Student's t-noise.

Following the approach in Section 4.4.2, we use the first 150 frequencies of the log-periodograms. For the training set, we arbitrarily selecte 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 series of nonnegative λ_c, we set $\lambda_1 = \lambda_c \times 10^{-7}\|A^\top b\|_\infty$ and $\lambda_2 = 0.1\lambda_1$. We employ four solvers: PGiPN, PGilbfgs, PG, and ZeroFPR. Subsequently, we 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), and the mean CPU time and error rate are recorded for each λ_c, presented in Figure 4.2. Motivated by the experiment in Section 4.4.2, we also draw Figure 4.3, recording $\|\widehat{B}x^*\|_0$ and the error rate for all the tested cases for four solvers.



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

Figure 4.2: Mean of the cpu time and the error rate on 30 examples for four solvers

We see from Figure 4.2(a) that in terms of efficiency, PGiPN is always the best one, more than ten times faster than the other three solvers. The reason is that the

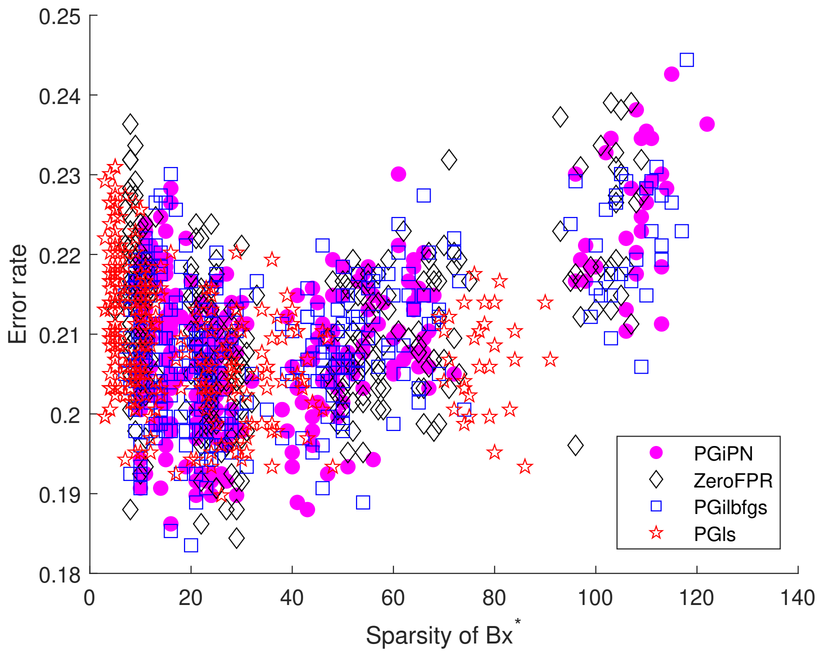


Figure 4.3: Scatter figure for all tested examples, recording the relationship of sparsity ($\|\widehat{B}x^*\|_0$) and the error rate of classification.

other three solvers depend heavily on the proximal mapping of g, and its computation is a little time-consuming, which reflects the advantage of the projected regularized Newton steps in PGiPN. From Figure 4.2(b), when $\lambda_c = 1$, PGiPN reaches the smallest mean error rate among four solvers for 10 λ_c's. When λ_c is large (> 0.4), PGiPN and PGilbfgs tend to outperform ZeroFPR and PG. Moreover, when λ_c is small (< 0.1), the solutions returned by PG have the best error rate among four solvers. This is because $\widehat{B}x^*$ produced by PG is sparser than those of the other three solvers under the same λ_c, which can be observed from Figure 4.3. For small λ_c, the solutions by the other three solvers are not sparse, leading to high error rate of classification.

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 be scaled such that all the entries belong to $[0, 1]$. 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 the vector $b \in \mathbb{R}^m$ represents a blurred image obtained by adding Gauss noise $e \sim \mathcal{N}(0, \varepsilon)$ with $\varepsilon > 0$ to $A\overline{x}$, i.e., $b = A\overline{x} + e$. We apply model (4.1) with $f = \frac{1}{2}\|A \cdot -b\|^2$, $B = \widehat{B}$, $l_b = 0$ and $u_b = \mathbf{1}$, to restore the blurred images. We test five solvers, which are PGiPN, PGiPN(r), PGilbfgs, ZeroFPR and PG. For PGiPN(r), we set $\eta_1 = 0.01, \eta_2 = 0.01$ in (4.56). For all these five solvers, we employ $\lambda_1 = \lambda_2 = 0.0005 \times \|A^\top b\|_\infty$. Under different ε's, we compare the performance of these five solvers in terms of required 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, we record the iterations (or time) in the form $M(N)$, where M means the total iterations (or time) and N means the iterations (or time) in Newton step. PSNR measures the quality of the restored images. The higher PSNR, the better the quality of restoration. Table 4.2 presents the numerical results.

From Table 4.2, PGiPN(r) always performs the best in terms of efficiency, which verifies the effectiveness of the acceleration scheme proposed in Section 4.4.1. PGiPN is faster than PGilbfgs, and PGilbfgs is faster than PG, supporting the effective acceleration of the Newton steps. However, 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 a computation of the proximal mapping of g, which is expensive (2-5 seconds).

Despite the superiority of efficiency, the solutions yielded by PGiPN(r) is not

Table 4.2: Numerical comparison of six solvers on recovery of blurred image with $\lambda_1 = \lambda_2 = 0.0005\|A^\top b\|_\infty$

Noise		PGiPN	PGiPN(r)	PGilbfgs	PG	ZeroFPR
$\varepsilon = 0.01$	Iter	379(3)	123(6)	529(31)	796	361
	Time	1.70e3(10.2)	5.61e2(21.4)	2.39e3(6.0)	3.43e3	2.35e4
	Fval	37.88	37.95	37.88	37.88	37.77
	xNnz	63858	63805	63858	63858	63717
	BxNnz	5767	5995	5776	5779	5834
	psnr	25.90	25.77	25.90	25.90	25.91
$\varepsilon = 0.02$	Iter	281(4)	109(6)	457(38)	853	286
	Time	1.20e3(8.8)	4.74e2(13.2)	1.95e3(7.4)	3.62e3	1.82e4
	Fval	45.98	46.05	45.98	45.98	45.83
	xNnz	63495	63440	63495	63495	63350
	BxNnz	6098	6320	6098	6099	6143
	psnr	25.41	25.23	25.42	25.42	25.33
$\varepsilon = 0.03$	Iter	234(3)	94(3)	325(15)	717	332
	Time	9.8e2(6.5)	3.98e2(6.5)	1.36e3(6.7)	2.97e3	1.88e4
	Fval	60.26	60.34	60.26	60.26	60.02
	xNnz	63006	62944	63006	63006	62800
	BxNnz	6594	6844	6597	6592	6710
	psnr	24.90	24.67	24.90	24.90	24.76
$\varepsilon = 0.04$	Iter	255(3)	78(5)	360(19)	526	230
	Time	1.04e3(6.2)	3.37e2(20.0)	1.45e3(3.9)	2.18e3	1.11e4
	Fval	77.82	77.87	77.82	77.82	77.44
	xNnz	62103	62002	62104	62104	61853
	BxNnz	7267	7553	7268	7271	7427
	psnr	24.20	23.85	24.20	24.20	24.00
$\varepsilon = 0.05$	Iter	263(3)	76(11)	389(29)	688	168
	Time	1.05e3(10.3)	3.30e2(28.5)	1.55e3(6.3)	2.71e3	5.91e3
	Fval	99.65	99.72	99.65	99.65	98.93
	xNnz	61376	61286	61381	61381	60963
	BxNnz	7955	8283	7955	7956	8240
	psnr	23.36	23.00	23.37	23.37	22.87

good. We also observe that $\|\widehat{B}x^*\|_0$ of PGiPN(r) is a little higher than those of PGiPN, PGilbfgs and PG, because PGiPN(r) runs few PG steps, so that its structured sparsity is not well reduced. Moreover, the PSNR is closely related to $\|\widehat{B}x^*\|_0$, and this leads to the weakest performance of PGiPN(r) in terms of PSNR. On the other hand, although ZeroFPR always outputs solutions with the smallest objective value, its PSNR is not as good as the objective value. The performance of PGiPN, PGilbfgs and PG in terms of the objective value and PSNR are quite similar. Taking the efficiency and the quality of the output into consideration, we conclude that PGiPN is the best solver for this test.

Chapter 5

Conclusion

In this thesis, we considered a hybrid framework of proximal gradient method and Newton-type method for two classes of nonconvex sparse optimization problems, which achieve global and superlinear convergence under mild conditions.

For the ℓ_q-norm regularized composite problem (3.1), we proposed a hybrid of PG and regularized Newton method by exploiting the special structures of the ℓ_q-norm. We not only established the convergence of the whole iterate sequence under a mild curve-ratio condition and the KL property of the objective function, but also achieved a superlinear convergence rate under an additional local error bound condition. In particular, the local superlinear convergence result neither requires the isolatedness of the limit point nor its local minimum property.

Moreover, we developed a polynomial-time algorithm for computing a point in the proximal mapping of $\lambda_1\|\widehat{B}x\|_0 + \lambda_2\|x\|_0 + \delta_\Omega(x)$, which makes PG available to solve (4.1) with $B = \widehat{B}$. To accelerate the PG method, we employed our hybrid framework to solve problem (4.1). We proved the convergence of the whole iterate sequence under a mild nondegeneracy condition, a curve-ratio condition and the KL property of the objective function, 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 minimality of the limit point.

In the future, we will consider one direction as extension of this thesis. In this study, we developed globalized algorithms for regularized Newton methods and projected regularized Newton methods. It would be intriguing to extend this framework to globalize semismooth Newton methods. In fact, assume that x is an L-stationary point of the composite problem (1.7). Then, there exists $\mu > 0$ such that that

$$x \in \mathrm{prox}_{\mu\phi}(x - \mu^{-1}\nabla\psi(x)). \tag{5.1}$$

For the reason that the proximal mapping may own a better smooth property than the original objective function, finding some stationary point of Ψ by solving (5.1) is a promising method. We plan to solve this system by using semismooth Newton method. To do it, the first thing we need to consider is that, in which case $\mathrm{prox}_{\mu\phi}(x - \mu^{-1}\nabla\psi(x))$ is single-valued and locally Lipschitz continuous, ensuring the existence of the Clarke generalized Jacobian. Existing result in (Themelis et al., 2018, Theorem 4.7) indicates that for given critical point $\overline{x}$ of Ψ, if ϕ is prox-regular at $\overline{x}$ for $-\nabla\psi(\overline{x})$ and prox-bounded, then for sufficiently small $\mu > 0$, $\mathrm{prox}_{\mu\phi}(x - \mu^{-1}\nabla\psi(x))$ is single-valued and Lipschitz continuous at a neighborhood of $\overline{x}$. This result provides a sufficient condition for the single-valuedness of $\mathrm{prox}_{\mu\phi}(x - \mu^{-1}\nabla\psi(x))$ at a neighborhood of some critical point. However, this is not enough, because to design a globalize semismooth Newton method, we also need to consider the single-valuedness and the locally Lipschitz continuity of the proximal mapping of the points far away from the critical points. For an arbitrarily given point x, how to ensure that $\mathrm{prox}_{\mu\phi}(x - \mu^{-1}\nabla\psi(x))$ is single-valued and locally Lipschitz continuous? If this question could be well resolved, it is possible to design a globally convergent semismooth Newton method for (5.1). We will leave it in our future study.

Bibliography

Absil, P.-A., Mahony, R., and Andrews, B. (2005), "Convergence of the Iterates of Descent Methods for Analytic Cost Functions," *SIAM Journal on Optimization*, 16, 531–547.

Ahookhosh, M., Themelis, A., and Patrinos, P. (2021), "A Bregman Forward-backward Linesearch Algorithm for Nonconvex Composite Optimization: Superlinear Convergence to Nonisolated Local Minima," *SIAM Journal on Optimization*, 31, 653–685.

Aravkin, A., Friedlander, M. P., Herrmann, F. J., and Van Leeuwen, T. (2012), "Robust Inversion, Dimensionality Reduction, and Randomized Sampling," *Mathematical Programming*, 134, 101–125.

Attouch, H. and Bolte, J. (2009), "On the Convergence of the Proximal Algorithm for Nonsmooth Functions Involving Analytic Features," *Mathematical Programming*, 116, 5–16.

Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. (2010), "Proximal Alternating Minimization and Projection Methods for Nonconvex Problems: an Approach Based on the Kurdyka-Łojasiewicz Inequality," *Mathematics of Operations Research*, 35, 438–457.

Attouch, H., Bolte, J., and Svaiter, B. F. (2013), "Convergence of Descent Methods for Semi-algebraic and Tame Problems: Proximal Algorithms, Forward–backward Splitting, and Regularized Gauss-Seidel Methods," *Mathematical Programming*, 137, 91–129.

Bao, C., Dong, B., Hou, L., Shen, Z., Zhang, X., and Zhang, X. (2016), "Image Restoration by Minimizing Zero Norm of Wavelet Frame Coefficients," *Inverse Problems*, 32, 115004.

Bareilles, G., Iutzeler, F., and Malick, J. (2023), "Newton Acceleration on Manifolds Identified by Proximal Gradient Methods," *Mathematical Programming*, 200, 37–70.

Barzilai, J. and Borwein, J. M. (1988), "Two-point Step Size Gradient Methods," *IMA Journal of Numerical Analysis*, 8, 141–148.

Bauschke, H. H., Borwein, J. M., and Li, W. (1999), "Strong Conical Hull Intersection Property, Bounded Linear Regularity, Jameson's Property (G), and Error Bounds in Convex Optimization," *Mathematical Programming*, 86, 135–160.

Bauschke, H. H., Bolte, J., and Teboulle, M. (2017), "A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications," *Mathematics of Operations Research*, 42, 330–348.

Beck, A. and Hallak, N. (2018), "Proximal Mapping for Symmetric Penalty and Sparsity," *SIAM Journal on Optimization*, 28, 496–527.

Beck, A. and Hallak, N. (2019), "Optimization Problems Involving Group Sparsity Terms," *Mathematical Programming*, 178, 39–67.

Beck, A. and Teboulle, M. (2009), "A Fast Iterative Shrinkage-thresholding Algorithm for Linear Inverse Problems," *SIAM Journal on Imaging Sciences*, 2, 183–202.

Bello Cruz, J. Y. and Nghia, T. T. (2016), "On the Convergence of the Forward–backward Splitting Method with Linesearches," *Optimization Methods and Software*, 31, 1209–1238.

Bello-Cruz, Y., Li, G., and Nghia, T. T. (2021), "On the linear convergence of forward–backward splitting method: Part I—Convergence analysis," *Journal of Optimization Theory and Applications*, 188, 378–401.

Bertsekas, D. P. (1982), "Projected Newton Methods for Optimization Problems with Simple Constraints," *SIAM Journal on Control and Optimization*, 20, 221–246.

Bertsekas, D. P. (1997), "Nonlinear Programming," *Journal of the Operational Research Society*, 48, 334–334.

Bertsimas, D., King, A., and Mazumder, R. (2016), "Best Subset Selection via a Modern Optimization Lens," *The Annals of Statistics*, 44, 813–852.

Bhatia, R. (2013), *Matrix Analysis*, vol. 169, Springer Science & Business Media.

Bian, W. and Chen, X. (2012a), "Smoothing Neural Network for Constrained Non-Lipschitz Optimization with Applications," *IEEE Transactions on Neural Networks and Learning Systems*, 23, 399–411.

Bian, W. and Chen, X. (2012b), "Smoothing SQP Algorithm for Non-Lipschitz Optimization with Complexity Analysis," *Preprint, February.*

Bian, W. and Chen, X. (2020), "A Smoothing Proximal Gradient Algorithm for Nonsmooth Convex Regression with Cardinality Penalty," *SIAM Journal on Numerical Analysis*, 58, 858–883.

Blumensath, T. and Davies, M. E. (2008), "Iterative Thresholding for Sparse Approximations," *Journal of Fourier Analysis and Applications*, 14, 629–654.

Blumensath, T. and Davies, M. E. (2010), "Normalized Iterative Hard Thresholding: Guaranteed Stability and Performance," *IEEE Journal of Selected Topics in Signal Processing*, 4, 298–309.

Bolte, J. and Pauwels, E. (2016), "Majorization-minimization Procedures and Convergence of SQP Methods for Semi-algebraic and Tame Programs," *Mathematics of Operations Research*, 41, 442–465.

Bolte, J., Daniilidis, A., and Lewis, A. (2006), "A Nonsmooth Morse–Sard Theorem for Subanalytic Functions," *Journal of Mathematical Analysis and Applications*, 321, 729–740.

Bolte, J., Daniilidis, A., Lewis, A., and Shiota, M. (2007a), "Clarke Subgradients of Stratifiable Functions," *SIAM Journal on Optimization*, 18, 556–572.

Bolte, J., Daniilidis, A., and Lewis, A. (2007b), "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM Journal on Optimization*, 17, 1205–1223.

Bolte, J., Daniilidis, A., Ley, O., and Mazet, L. (2008), "Characterizations of Lojasiewicz inequalities and applications," *arXiv preprint arXiv:0802.0826*.

Bolte, J., Sabach, S., and Teboulle, M. (2014), "Proximal Alternating Linearized Minimization for Nonconvex and Nonsmooth Problems," *Mathematical Programming*, 146, 459–494.

Bolte, J., Nguyen, T. P., Peypouquet, J., and Suter, B. W. (2017), "From Error Bounds to the Complexity of First-order Descent Methods for Convex Functions," *Mathematical Programming*, 165, 471–507.

Bolte, J., Sabach, S., Teboulle, M., and Vaisbourd, Y. (2018), "First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems," *SIAM Journal on Optimization*, 28, 2131–2151.

Bonnans, J. F. and Shapiro, A. (2013), *Perturbation Analysis of Optimization Problems*, Springer Science & Business Media.

Burke, J. V. and Moré, J. J. (1988), "On the Identification of Active Constraints," *SIAM Journal on Numerical Analysis*, 25, 1197–1211.

Candes, E. J., Wakin, M. B., and Boyd, S. P. (2008), "Enhancing Sparsity by Reweighted ℓ_q Minimization," *Journal of Fourier Analysis and Applications*, 14, 877–905.

Cao, W., Sun, J., and Xu, Z. (2013), "Fast Image Deconvolution Using Closed-form Thresholding Formulas of $L_q(q = \frac{1}{2}, \frac{2}{3})$ Regularization," *Journal of Visual Communication and Image Representation*, 24, 31–41.

Chartrand, R. (2007), "Exact Reconstruction of Sparse Signals via Nonconvex Minimization," *IEEE Signal Processing Letters*, 14, 707–710.

Chen, X. (2012), "Smoothing Methods for Nonsmooth, Nonconvex Minimization," *Mathematical Programming*, 134, 71–99.

Chen, X. and Zhou, W. (2014), "Convergence of the Reweighted ℓ_1 Minimization Algorithm for $\ell_2 - \ell_q$ Minimization," *Computational Optimization and Applications*, 59, 47–61.

Chen, X., Xu, F., and Ye, Y. (2010), "Lower Bound Theory of Nonzero Entries in Solutions of $\ell_2 - \ell_p$ Minimization," *SIAM Journal on Scientific Computing*, 32, 2832–2852.

Chen, X., Ng, M. K., and Zhang, C. (2012), "Non-Lipschitz ℓ_p-regularization and Box Constrained Model for Image Restoration," *IEEE Transactions on Image Processing*, 21, 4709–4721.

Chen, X., Niu, L., and Yuan, Y. (2013), "Optimality Conditions and a Smoothing Trust Region Newton Method for Nonlipschitz Optimization," *SIAM Journal on Optimization*, 23, 1528–1552.

Cheng, W., Chen, Z., and Hu, Q. (2020), "An Active Set Barzilar–Borwein Algorithm for ℓ_0 Regularized Optimization," *Journal of Global Optimization*, 76, 769–791.

Combettes, P. L. and Wajs, V. R. (2005), "Signal Recovery by Proximal Forward-backward Splitting," *Multiscale Modeling & Simulation*, 4, 1168–1200.

Davenport, H. and Schinzel, A. (1965), "A Combinatorial Problem Connected with Differential Equations," *American Journal of Mathematics*, 87, 684–694.

De Marchi, A. and Themelis, A. (2022), "Proximal Gradient Algorithms under Local Lipschitz Gradient Continuity," *Journal of Optimization Theory and Applications*, 194, 771–794.

Facchinei, F. and Pang, J.-S. (2003), *Finite-dimensional Variational Inequalities and Complementarity Problems*, Springer.

Figueiredo, M. A., Bioucas-Dias, J. M., and Nowak, R. D. (2007), "Majorization–minimization Algorithms for Wavelet-based Image Restoration," *IEEE Transactions on Image Processing*, 16, 2980–2991.

Frank, L. E. and Friedman, J. H. (1993), "A Statistical Biew of some Chemometrics Regression Tools," *Technometrics*, 35, 109–135.

Frankel, P., Garrigos, G., and Peypouquet, J. (2015), "Splitting Methods with Variable Metric for Kurdyka–Łojasiewicz Functions and General Convergence Rates," *Journal of Optimization Theory and Applications*, 165, 874–900.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *The Annals of Applied Statistics*, 1, 302–332.

Ge, D., Jiang, X., and Ye, Y. (2011), "A Note on the Complexity of L_p Minimization," *Mathematical Programming*, 129, 285–299.

Gong, P., Zhang, C., Lu, Z., Huang, J., and Ye, J. (2013), "A General Iterative Shrinkage and Thresholding Algorithm for Non-convex Regularized Optimization Problems," in *International Conference on Machine Learning*, pp. 37–45, PMLR.

Guo, K., Han, D., and Wu, T.-T. (2017), "Convergence of Alternating Direction Method for Minimizing Sum of two Nonconvex Functions with Linear Constraints," *International Journal of Computer Mathematics*, 94, 1653–1669.

Herrity, K. K., Gilbert, A. C., and Tropp, J. A. (2006), "Sparse Approximation via Iterative Thresholding," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 3, pp. III–III, IEEE.

Hu, Y., Li, C., Meng, K., Qin, J., and Yang, X. (2017), "Group Sparse Optimization via $\ell_{p,q}$ Regularization," *The Journal of Machine Learning Research*, 18, 960–1011.

Hu, Y., Li, C., Meng, K., and Yang, X. (2021), "Linear Convergence of Inexact Descent Method and Inexact Proximal Gradient Algorithms for Lower-order Regularization Problems," *Journal of Global Optimization*, 79, 853–883.

Huang, J., Jiao, Y., Liu, Y., and Lu, X. (2018), "A Constructive Approach to L_0 Penalized Regression," *The Journal of Machine Learning Research*, 19, 403–439.

Huang, L., Jia, J., Yu, B., Chun, B.-G., Maniatis, P., and Naik, M. (2010), "Predicting Execution Time of Computer Programs Using Sparse Polynomial Regression," *Advances in Neural Information Processing Systems*, 23, 883–891.

Huang, X. and Yang, X. (2003), "A Unified Augmented Lagrangian Approach to Duality and Exact Penalization," *Mathematics of Operations Research*, 28, 533–552.

Ito, K. and Kunisch, K. (2013), "A Variational Approach to Sparsity Optimization Based on Lagrange Multiplier Theory," *Inverse Problems*, 30, 015001.

Jadon, S. (2020), “A survey of loss functions for semantic segmentation,” in *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*, pp. 1–7, IEEE.

Jewell, S. and Witten, D. (2018), “Exact Spike Train Inference via ℓ_0 Optimization,” *The Annals of Applied Statistics*, 12, 2457–2482.

Jewell, S. W., Hocking, T. D., Fearnhead, P., and Witten, D. M. (2020), “Fast Nonconvex Deconvolution of Calcium Imaging Data,” *Biostatistics*, 21, 709–726.

Jia, X., Kanzow, C., and Mehlitz, P. (2023), “Convergence Analysis of the Proximal Gradient Method in the Presence of the Kurdyka–Łojasiewicz Property Without Global Lipschitz Assumptions,” *SIAM Journal on Optimization*, 33, 3038–3056.

Jiang, H., Luo, S., and Dong, Y. (2021), “Simultaneous Feature Selection and Clustering Based on Square Root Optimization,” *European Journal of Operational Research*, 289, 214–231.

Jiao, Y., Jin, B., and Lu, X. (2015), “A Primal Dual Active Set with Continuation Algorithm for the ℓ_0-regularized Optimization Problem,” *Applied and Computational Harmonic Analysis*, 39, 400–426.

Kanzow, C. and Lechner, T. (2021), “Globalized Inexact Proximal Newton-type Methods for Nonconvex Composite Functions,” *Computational Optimization and Applications*, 78, 377–410.

Kanzow, C. and Lechner, T. (2022), “Efficient Regularized Proximal Quasi-Newton Methods for Large-scale Nonconvex Composite Optimization Problems,” *arXiv preprint arXiv:2210.07644*.

Kanzow, C. and Mehlitz, P. (2022), “Convergence Properties of Monotone and Nonmonotone Proximal Gradient Methods Revisited,” *Journal of Optimization Theory and Applications*, 195, 624–646.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012), “Optimal Detection of Changepoints with a Linear Computational Cost,” *Journal of the American Statistical Association*, 107, 1590–1598.

Kurdyka, K. (1998), “On Gradients of Functions Definable in o-minimal Structures,” in *Annales de l'institut Fourier*, vol. 48, pp. 769–783.

Lai, M.-J. and Wang, J. (2011), “An Unconstrained ℓ_q Minimization with $0 < q \leq 1$ for Sparse Solution of Underdetermined Linear Systems,” *SIAM Journal on Optimization*, 21, 82–101.

Lai, M.-J., Xu, Y., and Yin, W. (2013), "Improved Iteratively Reweighted Least Squares for Unconstrained Smoothed ℓ_q Minimization," *SIAM Journal on Numerical Analysis*, 51, 927–957.

Land, S. R. and Friedman, J. H. (1997), "Variable Fusion: A New Adaptive Signal Regression Method," Tech. rep., Technical Report 656, Department of Statistics, Carnegie Mellon University.

Lee, J. D., Sun, Y., and Saunders, M. A. (2014), "Proximal Newton-type Methods for Minimizing Composite Functions," *SIAM Journal on Optimization*, 24, 1420–1443.

Li, G. and Pong, T. K. (2015), "Global Convergence of Splitting Methods for Nonconvex Composite Optimization," *SIAM Journal on Optimization*, 25, 2434–2460.

Li, G. and Pong, T. K. (2016), "Douglas–Rachford Splitting for Nonconvex Optimization with Application to Nonconvex Feasibility Problems," *Mathematical Programming*, 159, 371–401.

Li, G. and Pong, T. K. (2018), "Calculus of the Exponent of Kurdyka–Łojasiewicz Inequality and its Applications to Linear Convergence of First-order Methods," *Foundations of Computational Mathematics*, 18, 1199–1232.

Li, J., So, A. M.-C., and Ma, W.-K. (2020), "Understanding notions of stationarity in nonsmooth optimization: A guided tour of various constructions of subdifferential for nonsmooth functions," *IEEE Signal Processing Magazine*, 37, 18–31.

Li, M., Meng, K., and Yang, X. (2023), "Variational Analysis of Kurdyka-Lojasiewicz Property by Way of Outer Limiting Subgradients," *arXiv preprint arXiv:2308.15760*.

Li, X., Sun, D., and Toh, K.-C. (2018), "On Efficiently Solving the Subproblems of a Level-set Method for Fused Lasso Problems," *SIAM Journal on Optimization*, 28, 1842–1866.

Liu, J., Ji, S., and Ye, J. (2009), "SLEP: Sparse Learning with Efficient Projections," *Arizona State University*, 6, 7.

Liu, J., Yuan, L., and Ye, J. (2010), "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*, pp. 323–332.

Liu, R., Pan, S., Wu, Y., and Yang, X. (2022), "An Inexact Regularized Proximal Newton Method for Nonconvex and Nonsmooth Optimization," *arXiv preprint arXiv:2209.09119*.

Liu, R., Pan, S., Wu, Y., and Yang, X. (2024), "An Inexact Regularized Proximal Newton Method for Nonconvex and Nonsmooth Optimization," *Computational Optimization and Applications.*

Liu, Z., Wu, C., and Zhao, Y. (2019), "A New Globally Convergent Algorithm for Non-Lipschitz $\ell_p - \ell_q$ Minimization," *Advances in Computational Mathematics*, 45, 1369–1399.

Lojasiewicz, S. (1963), "Une Propriété Topologique des Sous-ensembles Analytiques Réels," *Les équations aux dérivées partielles*, 117, 87–89.

Lojasiewicz, S. (1984), "Sur les trajectoires du gradient d'une fonction analytique.(French)[Trajectories of the gradient of an analytic function] Geometry seminars, 1982–1983 (Bologna, 1982/1983), 115–117, Univ," *Stud. Bologna, Bologna.*

Lu, Z. (2014a), "Iterative Hard Thresholding Methods for ℓ_0 Regularized Convex Cone Programming," *Mathematical Programming*, 147, 125–154.

Lu, Z. (2014b), "Iterative Reweighted Minimization Methods for ℓ_p Regularized Unconstrained Nonlinear Programming," *Mathematical Programming*, 147, 277–307.

Lu, Z. and Zhang, Y. (2013), "Sparse Approximation via Penalty Decomposition Methods," *SIAM Journal on Optimization*, 23, 2448–2478.

Luo, Z.-Q. and Tseng, P. (1992), "Error Bound and Convergence Analysis of Matrix Splitting Algorithms for the Affine Variational Inequality Problem," *SIAM Journal on Optimization*, 2, 43–54.

Luo, Z.-Q., Pang, J.-S., and Ralph, D. (1996), *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press.

Molinari, C., Liang, J., and Fadili, J. (2019), "Convergence Rates of Forward–Douglas–Rachford Splitting Method," *Journal of Optimization Theory and Applications*, 182, 606–639.

Mordukhovich, B. S., Yuan, X., Zeng, S., and Zhang, J. (2023), "A Globally Convergent Proximal Newton-type Method in Nonsmooth Convex Optimization," *Mathematical Programming*, 198, 899–936.

Nikolova, M., Ng, M. K., Zhang, S., and Ching, W.-K. (2008), "Efficient Reconstruction of Piecewise Constant Images Using Nonsmooth Nonconvex Minimization," *SIAM Journal on Imaging Sciences*, 1, 2–25.

Nocedal, J. and Wright, S. (2006), *Numerical Optimization*, Springer Science & Business Media.

Noll, D. (2014), "Convergence of Non-smooth Descent Methods Using the Kurdyka–Łojasiewicz Inequality," *Journal of Optimization Theory and Applications*, 160, 553–572.

Ochs, P. (2018), "Local Convergence of the Heavy-ball Method and iPiano for Non-convex Optimization," *Journal of Optimization Theory and Applications*, 177, 153–180.

Ochs, P. (2019), "Unifying Abstract Inexact Convergence Theorems and Block Coordinate Variable Metric iPiano," *SIAM Journal on Optimization*, 29, 541–570.

Ochs, P., Chen, Y., Brox, T., and Pock, T. (2014), "iPiano: Inertial Proximal Algorithm for Nonconvex Optimization," *SIAM Journal on Imaging Sciences*, 7, 1388–1419.

Pan, S. and Liu, Y. (2018), "Subregularity of Subdifferential Mappings Relative to the Critical Set and KL Property of Exponent 1/2," *arXiv preprint arXiv:1812.00558*.

Pan, S., Liang, L., and Liu, Y. (2023), "Local Optimality for Stationary Points of Group Zero-norm Regularized Problems and Equivalent Surrogates," *Optimization*, 72, 2311–2343.

Patrinos, P. and Bemporad, A. (2013), "Proximal Newton Methods for Convex Composite Optimization," in *52nd IEEE Conference on Decision and Control*, pp. 2358–2363, IEEE.

Poliquin, R. A. and Rockafellar, R. T. (2010), "A Calculus of Prox-regularity," *Journal of Convex Analysis*, 17, 203–210.

Qian, Y. and Pan, S. (2023), "Convergence of a Class of Nonmonotone Descent Methods for Kurdyka–Łojasiewicz Optimization Problems," *SIAM Journal on Optimization*, 33, 638–651.

Rockafellar, R. T. (1970), *Convex Analysis*, Princeton University Press.

Rockafellar, R. T. and Wets, R. J.-B. (2009), *Variational Analysis*, vol. 317, Springer.

Rudin, L. I., Osher, S., and Fatemi, E. (1992), "Nonlinear Total Variation Based Noise Removal Algorithms," *Physica D: Nonlinear Phenomena*, 60, 259–268.

Saab, R., Chartrand, R., and Yilmaz, O. (2008), "Stable Sparse Approximations via Nonconvex Optimization," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3885–3888, IEEE.

Salzo, S. (2017), "The Variable Metric Forward-backward Splitting Algorithm under Mild Differentiability Assumptions," *SIAM Journal on Optimization*, 27, 2153–2181.

Sharir, M. (1988), "Davenport-Schinzel Sequences and Their Geometric Applications," in *Theoretical Foundations of Computer Graphics and CAD*, pp. 253–278, Springer.

Sra, S. (2012), "Scalable Nonconvex Inexact Proximal Splitting," *Advances in Neural Information Processing Systems*, 25.

Stella, L., Themelis, A., and Patrinos, P. (2017), "Forward–backward Quasi-Newton Methods for Nonsmooth Optimization Problems," *Computational Optimization and Applications*, 67, 443–487.

Stewart, G. W. (2002), "A Krylov–Schur Algorithm for Large Eigenproblems," *SIAM Journal on Matrix Analysis and Applications*, 23, 601–614.

Sun, Y., Jeong, H., Nutini, J., and Schmidt, M. (2019), "Are We There yet? Manifold Identification of Gradient-related Proximal Methods," in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1110–1119, PMLR.

Themelis, A., Stella, L., and Patrinos, P. (2018), "Forward-backward Envelope for the Sum of Two Nonconvex Functions: Further Properties and Nonmonotone Line-search Algorithms," *SIAM Journal on Optimization*, 28, 2274–2303.

Themelis, A., Ahookhosh, M., and Patrinos, P. (2019), "On the Acceleration of forward-backward Splitting via an Inexact Newton Method," in *Splitting Algorithms, Modern Operator Theory, and Applications*, pp. 363–412, Springer.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 91–108.

Ueda, K. and Yamashita, N. (2010), "Convergence Properties of the Regularized Newton Method for the Unconstrained Nonconvex Optimization," *Applied Mathematics and Optimization*, 62, 27–46.

Van den Dries, L. (1998), *Tame Topology and o-minimal Structures*, vol. 248, Cambridge University Press.

Van den Dries, L. and Miller, C. (1996), "Geometric Categories and o-minimal Structures," *Duke Mathematical Journal*, 84.

Wang, H., Zeng, H., Wang, J., and Wu, Q. (2021a), "Relating ℓ_p Regularization and Reweighted ℓ_1 Regularization," *Optimization Letters*, 15, 2639–2660.

Wang, H., Zeng, H., and Wang, J. (2022), "An Extrapolated Iteratively Reweighted ℓ_1 Method with Complexity Analysis," *Computational Optimization and Applications*, 83, 967–997.

Wang, H., Zeng, H., and Wang, J. (2023), "Convergence Rate Analysis of Proximal Iteratively Reweighted ℓ_1 Methods for ℓ_p Regularization Problems," *Optimization Letters*, 17, 413–435.

Wang, X. and Wang, Z. (2023), "Calculus Rules of the Generalized Concave Kurdyka–Łojasiewicz Property," *Journal of Optimization Theory and Applications*, 197, 839–854.

Wang, X., Ye, J. J., Yuan, X., Zeng, S., and Zhang, J. (2021b), "Perturbation Techniques for Convergence Analysis of Proximal Gradient Method and Other First-order Algorithms via Variational Analysis," *Set-Valued and Variational Analysis*, pp. 1–41.

Wang, Y. and Yin, W. (2010), "Sparse Signal Reconstruction via Iterative Support Detection," *SIAM Journal on Imaging Sciences*, 3, 462–491.

Wang, Y., Yin, W., and Zeng, J. (2019), "Global Convergence of ADMM in Nonconvex Nonsmooth Optimization," *Journal of Scientific Computing*, 78, 29–63.

Wen, B., Chen, X., and Pong, T. K. (2017), "Linear Convergence of Proximal Gradient Algorithm with Extrapolation for a Class of Nonconvex Nonsmooth Minimization Problems," *SIAM Journal on Optimization*, 27, 124–145.

Wright, S. J., Nowak, R. D., and Figueiredo, M. A. (2009), "Sparse Reconstruction by Separable Approximation," *IEEE Transactions on Signal Processing*, 57, 2479–2493.

Wu, F. and Bian, W. (2020), "Accelerated Iterative Hard Thresholding Algorithm for l_0 Regularized Regression Problem," *Journal of Global Optimization*, 76, 819–840.

Wu, Y., Pan, S., and Bi, S. (2021), "Kurdyka–Łojasiewicz Property of Zero-norm Composite Functions," *Journal of Optimization Theory and Applications*, 188, 94–112.

Wu, Y., Pan, S., and Yang, X. (2023a), "An Inexact Projected Regularized Newton Method for Fused Zero-norms Regularization Problems," *arXiv preprint arXiv:2312.15718*.

Wu, Y., Pan, S., and Yang, X. (2023b), "A Regularized Newton Method for ℓ_q-Norm Composite Optimization Problems," *SIAM Journal on Optimization*, 33, 1676–1706.

Xu, Z., Zhang, H., Wang, Y., Chang, X., and Liang, Y. (2010), "$L_{1/2}$ Regularization," *Science China Information Sciences*, 53, 1159–1169.

Xu, Z., Chang, X., Xu, F., and Zhang, H. (2012), "$\ell_{1/2}$ Regularization: A Thresholding Representation Theory and a Fast Solver," *IEEE Transactions on Neural Networks and Learning Systems*, 23, 1013–1027.

Yang, X. and Huang, X. (2001), "A Nonlinear Lagrangian Approach to Constrained Optimization Problems," *SIAM Journal on Optimization*, 11, 1119–1144.

Yu, P., Li, G., and Pong, T. K. (2022), "Kurdyka–Łojasiewicz Exponent via Inf-projection," *Foundations of Computational Mathematics*, 22, 1171–1217.

Yue, M.-C., Zhou, Z., and So, A. M.-C. (2019), "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, 327–358.

Zeng, J., Lin, S., and Xu, Z. (2016), "Sparse regularization: Convergence of iterative jumping thresholding algorithm," *IEEE Transactions on Signal Processing*, 64, 5106–5118.

Zhang, T., Pan, S., and Liu, R. (2023), "A VMiPG Method for Composite Optimization with Nonsmooth Term Having no Closed-form Proximal Mapping," *arXiv preprint arXiv:2308.13776*.

Zhou, R. and Palomar, D. P. (2021), "Solving high-order portfolios via successive convex approximation algorithms," *IEEE Transactions on Signal Processing*, 69, 892–904.

Zhou, S., Pan, L., and Xiu, N. (2021), "Newton Method for ℓ_0-Regularized Optimization," *Numerical Algorithms*, 88, 1541–1570.

Zhou, Z. and So, A. M.-C. (2017), "A Unified Approach to Error Bounds for Structured Convex Optimization Problems," *Mathematical Programming*, 165, 689–728.