PROXIMAL ALGORITHMS WITH EXTRAPOLATION FOR NONCONVEX NONSMOOTH OPTIMIZATION PROBLEMS

BO WEN

Ph.D

The Hong Kong Polytechnic University

This programme is jointly offered by The Hong Kong Polytechnic University and Harbin Institute of Technology

2017

The Hong Kong Polytechnic University
Department of Applied Mathematics
Harbin Institute of Technology
Department of Mathematics

Proximal Algorithms with Extrapolation for Nonconvex Nonsmooth Optimization Problems

Bo Wen

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

April 2017

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)

________WEN Bo________(Name of student)

Dedicate to my parents.

Abstract

In this thesis, we consider the proximal algorithms with extrapolation for solving nonconvex nonsmooth optimization problems. This class of optimization problems arise in many application areas of engineering, computer science, economic field, see [18, 20, 23, 27, 40, 54, 64]. Due to the importance of these problems, a plenty of algorithms are proposed for solving them. This thesis mainly studies two classes of the proximal algorithms with extrapolation for solving different structured nonconvex and nonsmooth optimization problems. The details are as follows:

1. We first consider the proximal gradient algorithm with extrapolation for minimizing the sum of a Lipschitz differentiable function and a proper closed convex function. Using the error bound condition studied in the literature [38] for analyzing the convergence properties of proximal gradient algorithm, we show that there exists a threshold such that if the extrapolation coefficients are chosen below this threshold, then the sequence generated converges R-linearly to a stationary point of the problem. Moreover, the corresponding sequence of objective values is also R-linearly convergent. In addition, the threshold reduces to 1 for convex problems and, as a consequence, we obtain the R-linear convergence of the sequence generated by FISTA with fixed restart. Again for convex problems, we show that the successive changes of the iterates vanish for many choices of sequences of extrapolation coefficients that approach the threshold. In particular, we prove that this conclusion also holds for the sequence generat-

ed by the FISTA. Finally, we present some numerical experiments to illustrate our results.

2. Difference-of-convex (DC) optimization problems attract many researchers' attention in recent years. Many numerical algorithms are proposed for solving them. Among these algorithms, difference-of-convex algorithm (DCA) is a fundamental and classical one. We consider a class of DC optimization problems whose objective is level-bounded and is the sum of a smooth convex function with Lipschitz gradient, a proper closed convex function and a continuous concave function. This kind of DC problems can be solved by the aforementioned DCA, however, a direct application of DCA may lead to difficult subproblems. To overcome this difficulty, proximal DCA has been proposed. While the subproblems involved in the proximal DCA are simpler, proximal DCA maybe slow in practice. This is because proximal DCA reduces to the proximal gradient algorithm when the concave part of the objective is void. In this theis, motivated by the extrapolation techniques for accelerating the proximal gradient algorithm in the convex settings, we consider a proximal difference-of-convex algorithm with extrapolation to possibly accelerate the proximal DCA. We show that any accumulation point of the sequence generated by our algorithm is a stationary point of the DC optimization problem for a fairly general choice of extrapolation parameters: in particular, the parameters can be chosen as in FISTA with fixed restart [26]. Moreover, by assuming the Kurdyka-Łojasiewicz property of an auxiliary function and the differentiability of the concave part, we establish global convergence of the sequence generated by our algorithm and analyze its convergence rate. From the results in our numerical experiments on two difference-of-convex regularized least squares models, proximal difference-of-convex algorithm with extrapolation usually outperforms the proximal DCA

with nonmonotone linesearch.

This thesis is based on the following papers written during the period of my study at Department of Applied Mathematics, The Hong Kong Polytechnic University as a Joint-PhD student with Harbin Institute of Technology.

1. Bo Wen, Xiaojun Chen, Ting Kei Pong. Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. SIAM Journal on Optimization, 27: 124-145, 2017.

2. Bo Wen, Xiaojun Chen, Ting Kei Pong. A proximal difference-of-convex algorithm with extrapolation. submitted to Computational Optimization and Applications. December 2016.

Acknowledgements

The last few years were really very important and memorable period in my whole life. I would like to acknowledge many people's assistance during my Joint-PhD program. Without their support, this thesis would not have been possible.

First and foremost, I would like to express my sincere gratitude to my supervisor Professor Chen Xiaojun for her valuable advice, professional guidance and generous support throughout the years. She always gave me many encouragement and insightful suggestions, which inspired me with new ideas and widen my mind. Her expansive knowledge of optimization and inexhaustible enthusiasm for research have impressed me profoundly.

I wish to express my deep thanks to my supervisor in Harbin Institute of Technology, Professor Xue Xiaoping for his enlightening guidance and insightful ideas. His amazing depth of mathematical knowledge enriched my mind and broadened my horizons.

Furthermore, I would like to thank Dr. Pong Ting Kei for his kindly help and invaluable discussions throughout the years. I learnt a lot of optimization algorithms and writing techniques from him. His depth of knowledge and rigorous and diligent attitude to academic research will influence me all my life.

I would like to thank Professor Lu Zhaosong, Professor Xiang Shuhuang, Professor Liu Xin and Dr. Liu Yafeng for their useful advice and great encouragement. I am very grateful to Professor Bian Wei for her selfless help on my research and life.

Many thanks to all the staff of Department of Applied Mathematics at The Hong Kong Polytechnic University for their help during my study here.

I want to thank all the members in my research group, Dr. Zhang Chao, Dr. An Congpei, Dr. Zhang Yanfang, Dr. Zhou Weijun, Dr. Sun Hailin, Dr. Wu Shulin, Dr. Zhang Zaikun, Dr. Liu Tianxiang, Dr. Wang Hong, Ms. Wang Qiyu, Mr. Yang Lei, Mr. Shi Yun, Mr. Liu Guidong, Ms. Pan Lili, Mr. Jiang Jie for their help and support in my research and life. To my good friends, Dr. Meng Kaiwen, Dr. Li Zhibao, Dr. Hao Meiling, Dr. Wei Yan, Mr. Yang Jin, Mr. Dong Zhilong, Mr. Yan Xiaodong, Mr. Wang Qingzheng, Ms. Fang Fei, Ms. Zhang Huili, Ms. Shi Yue, I am very grateful for their help and accompany during my study life, they make my life rich and colorful. And the friends in TU834 and P115 are deeply acknowledged.

I would like to thank The Hong Kong Polytechnic University and Harbin Institute of Technology for giving me this chance to enjoy the resources on both universities. This chance and support made me broaden my horizon and enrich my experience.

Last but not least, I want to express my special thanks to my parents and my sister for their unconditional love, constant support and encouragement to me.

Contents

List of Figures

List of Tables

List of Notations

$\mathbb{R}$	set of real numbers
$\mathbb{R}^n$	n-dimensional Euclidean space
$\mathbb{R}^{m\times n}$	set of $m \times n$ matrices
$\langle \cdot, \cdot \rangle$	standard inner product in Euclidean space
$\| \cdot \|$	the Euclidean norm
$\| \cdot \|_1$	the ℓ_1 norm
$\| \cdot \|_\infty$	the ℓ_∞ norm
e	the vector of all ones
A^T	the transpose of matrix A
$\lambda_{\max}(A)$	the largest eigenvalue of a a symmetric matrix $A \in \mathbb{R}^{n\times n}$
$\lambda_{\min}(A)$	the smallest eigenvalue of a a symmetric matrix $A \in \mathbb{R}^{n\times n}$
∇	gradient operator
∂	subgradient operator

Chapter 1

Introduction

In recent years, nonconvex problems arise in many application areas such as matrix completion [17], image processing [20, 48], portfolio selection [40] and so on. Since the nonconvex problems maybe from different practical applications, the structures of these nonconvex problems are potentially different. For example, both the quadratic problems and DC problems are nonconvex problems, but they have different structures. And there are many other different nonconvex problems whose structures are different. Based on the special structures of these nonconvex optimization problems, many algorithms including proximal gradient algorithm, DCA, and ipiano are proposed for solving them. However, some algorithms among them, in their original forms, are not efficient enough to solve the nonconvex problems. Thus various extrapolation techniques are used for possibly accelerating these algorithms.

In this thesis, we study the proximal algorithms with extrapolation for minimizing two classes of nonconvex optimization problems. Concretely, we first study the proximal gradient algorithm with extrapolation for a minimizing problem, whose objective is the sum of a Lipschitz differentiable function and a proper closed convex function. And then we consider a proximal difference-of-convex algorithm with extrapolation for minimizing a DC optimization problem.

1.1 Motivation and related algorithms

1.1.1 Problem description and proximal gradient algorithms

As we all know, the gradient algorithm is a very simple but fundamental algorithm for smooth and convex problems. During the past few years, many variants of gradient algorithm including heavy ball method [51] are proposed for speeding up the original gradient algorithm. Recently, due to the needs of some practical applications, many researchers pay attention to a kind of composite functions, which are the sum of a differentiable function and a nondifferentiable function whose proximal map is easy to compute. To solve this kind of optimization problems, proximal gradient algorithm [35] was proposed and further studied in many literatures [3, 43, 44, 45, 48].

In the first part of this thesis, we consider the following minimization problem:

$$\min_{x\in\mathbb{R}^n} F(x) := f(x) + g(x), \tag{1.1}$$

where g is a proper closed convex function and f is a possibly nonconvex function that has a Lipschitz continuous gradient. We also assume that the proximal operator of μg is easy to compute, where $\mu > 0$ is a fixed number. Moreover, we assume the optimal value of (1.1) is finite and attained.

Problem (1.1) arises in many real applications, such as image restoration [20, 48], compressed sensing [18, 27], matrix completion [17] and so on. Since the scales of these problems are very large, a plenty of first order algorithms are proposed for solving them. Among these first order methods, proximal gradient algorithm [35] is a fundamental and commonly used one, whose computational efforts in each iteration are the evaluations of ∇f and the proximal mapping of μg. When f in (1.1) is convex, then using the proximal gradient algorithm to solve (1.1), we obtain from [61, Theorem 1(a)] that

$$F(x^t) - \inf_{x\in\mathbb{R}^n} F(x) = O\left(\frac{1}{t}\right),$$

where $\{x^t\}$ is generated by the proximal gradient algorithm. However, the original proximal gradient algorithm can be slow when it is used for solving practical problems; see, for example, [26, Section 5].

Hence, many mathematicians and experts from different areas have tried different ways to accelerate the proximal gradient algorithm. Among those methods, performing extrapolation, which means adding *momentum* terms involving the previous iterations to the current iteration, is an efficient and simple strategy. We give a prototypical extrapolation algorithm as follows

$$\begin{cases} y^t = x^t + \beta_t(x^t - x^{t-1}), \\ x^{t+1} = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \langle \nabla f(y^t), x\rangle + \frac{1}{2\mu}\|x - y^t\|^2 + g(x) \right\}, \end{cases} \tag{1.2}$$

where $\mu > 0$ is a positive constant that depends on the Lipschitz continuity modulus of ∇f, and the extrapolation parameters β_t satisfy $0 \leq \beta_t \leq 1$ for all t. One well known example of the extrapolation methods is the fast iterative shrinkage-thresholding algorithm (FISTA) proposed by Beck and Teboulle [8], which is based on Nesterov's extrapolation techniques [43, 44, 46, 47] and is designed for solving (1.1) with f being convex and g being continuous. Their analysis can be directly extended to the case when g is a proper closed convex function. Another multistep version of accelerated gradient-like algorithm for solving problem (1.1) was proposed and studied by Nesterov [45], but the theoretical analysis and proof techniques were completely different. Since FISTA is a special extrapolation algorithm, it also takes the form (1.2) and requires $\{\beta_t\}$ to satisfy a certain recurrence relation. We can see from [8, 45] that FISTA displays a faster convergence rate than the original proximal gradient algorithm, which is

$$F(x^t) - \inf_{x\in\mathbb{R}^n} F(x) = O\left(\frac{1}{t^2}\right),$$

where $\{x^t\}$ is generated by FISTA. Since then, many other accelerated proximal gradient algorithms which are based on Nesterov's extrapolation techniques have been proposed and studied. We refer the readers to see [10, 11, 61] for more details.

Due to the fast convergence rate of FISTA in term of objective values, many literatures further study the extrapolation method (1.2), see, for example, [6, 21, 26, 33, 55]. Among them, O'Donoghue and Candès [26] proposed restart schemes for the extrapolation coefficients β_t based on FISTA for solving (1.1) with a convex f and a void g. They mainly reset the extrapolation coefficients $\beta_t = \beta_0$ every T iterations instead of using the recurrence relation of β_t in FISTA for all t. When f is strongly convex, they showed that the sequence of objective values was *globally* linearly convergent to the optimal value of the problems for sufficient large T. We see from the discussion in [26, Section 2.1] that FISTA with restart schemes is robust against errors in the estimation of the strong convexity modulus of f. Later, Chambolle and Dossal [21] showed that the whole sequence generated by (1.2) with $\beta_t = \frac{t-1}{t+\alpha-1}$ ($\alpha > 3$) for solving (1.1) with a convex f is convergent in Hilbert space. Recently, Attouch and Chabani [6] extended their results to allow errors in gradient computation. More recently, Tao, Boley and Zhang [55] proved local linear convergence of FISTA, when it is applied to solving the LASSO (i.e., g is a positive multiple of the ℓ_1 norm and f is a least squares loss function) by assuming that the problem has a unique solution which satisfies strict complementarity condition. Johnstone and Moulin [33] considered (1.1) with f being convex, and showed that the whole sequence generated by (1.2) is convergent by assuming that the extrapolation coefficients β_t satisfy $0 \leq \beta_t \leq \bar{\beta}$ for some $\bar{\beta} < 1$. Moreover, by imposing uniqueness of the optimal solution together with a technical assumption, they showed that the sequence generated by (1.2) is locally linearly convergent when applied to the LASSO for a particular choice of $\{\beta_t\}$.

Noting from the above literatures that the local linear convergence of (1.2) is only established for convex problems whose optimal solution is unique for some specific choices of extrapolation coefficients $\{\beta_t\}$. These conditions are too restrictive for many real application problems. In addition, for convex problems, nothing is known concerning the convergence behavior of $\{x^t\}$ when $\sup_t \beta_t = 1$ for a choice of $\{\beta_t\}$ other than $\beta_t = \frac{t-1}{t+\alpha-1}$ with $\alpha > 3$. Thus, we further consider the convergence behavior of the sequence $\{x^t\}$ generated by (1.2) in this thesis. In particular, we discuss local linear convergence under more general conditions in the possibly nonconvex case. We also study the convergence behavior of $\{x^t\}$ in the convex case when $\sup_t \beta_t = 1$.

1.1.2 DC problems and DCA

In this subsection, we introduce the difference-of-convex (DC) optimization problems and some well known algorithms which are usually applied to solving them.

DC optimization problems are very common in our real life. For example, they arise in compressed sensing [64], digital communication system [2] and assignment and power allocation [54]; see more applications in the recent monograph [62, Chapter 7]. DC optimization problems are problems whose objective can be written as the difference of a proper closed convex function and a continuous convex function. Hence, making use of the special structures of the DC problems, many algorithms have been proposed for solving them. Among them, DC algorithm (DCA) proposed by Tao and An [50] becomes a fundamental and classical algorithm for solving DC optimization problems, see more details in [7, 31, 41, 56, 57, 58]. This algorithm mainly uses a linear majorant to replace the concave part of the objective in DC problems and then solves the resulting convex problems. Despite the simple framework of this algorithm, a directly use of the original DCA may lead to difficult subproblems. In view of this, recently, Gotoh, Takeda, and Tono [31] proposed a proximal DCA[1] for solving DC optimization problems whose objective can be written as the sum of a

[1] This algorithm was called "the proximal difference-of-convex decomposition algorithm" in [31].

smooth convex function with Lipschitz gradient, a proper closed convex function and a continuous concave function. Their algorithm not only majorizes the concave part in the objective by a linear majorant, but also majorizes the smooth convex part by a quadratic majorant in each iteration. Then they showed that when the proximal mapping of the proper closed convex function is simple, which means it can be easily computed, the subproblems of their algorithm can be solved efficiently. However, we note that when the concave part of the objective is void, their proximal DCA is the same as the original proximal gradient algorithm for solving convex problems, which can be slow in practice [26, Section 5]. Hence, proximal DCA may be also slow in practice.

Then we want to incorporate some techniques to accelerate the proximal DCA. Performing extrapolation is a commonly used technique. Indeed, such technique can date back to Polyak's heavy ball method [51] for solving convex optimization problems. We refer readers to see Subsection 1.1.1 for a detailed overview of extrapolation technique used recently.

Inspired by the success of using extrapolation technique in the proximal gradient algorithm to accelerate the original proximal gradient algorithm, and in view of the fact that the proximal gradient algorithm and the proximal DCA are the same when applied to convex problems, in this thesis, we mainly use the extrapolation techniques to possibly accelerate the proximal DCA for solving the same DC optimization problems stated in proximal DCA [31].[2]

1.2 Contributions of this thesis

The contributions of this thesis are as follows:

- First, under the same error bound condition used in [38] for analyzing conver-

[2] In the numerical section of [31], the authors also state that incorporating extrapolation techniques suitably into the proximal DCA can accelerate the algorithm empirically.

gence of the proximal gradient algorithm, we show that there is a threshold $\widetilde{\beta}$ depending on f so that if $\sup_t \beta_t < \widetilde{\beta}$, then the sequence $\{x^t\}$ generated by (1.2) converges R-linearly to a stationary point of (1.1) and the sequence of the objective value $\{F(x^t)\}$ is also R-linearly convergent. In particular, if f is in addition convex, then $\widetilde{\beta}$ reduces to 1 and we can conclude that the sequence $\{x^t\}$ generated by the FISTA with fixed restart is R-linearly convergent to an optimal solution of (1.1); see Section 3.2.3. The error bound condition is satisfied for a wide range of problems including the LASSO, and hence our linear convergence result concerning (1.2) with a fixed μ is more general than those discussed in [33].

- Second, when f in (1.1) is convex and F is level-bounded, we show that if $\sup_t \beta_t = 1$, $\{\beta_t\}$ is nondecreasing and $\sum_{t=1}^{\infty}(1-\beta_t) = \infty$, then the successive changes $\|x^{t+1} - x^t\|$ go to 0 as $t \to \infty$. As a corollary, we deduce that $\lim\limits_{t\to\infty} \|x^{t+1} - x^t\| = 0$ for the sequence generated by the FISTA.

- For the DC problems, we propose a proximal difference-of-convex algorithm with extrapolation (pDCA$_e$) for solving them. We prove that, for a fairly general choice of extrapolation parameters, if the objective is level-bounded, then any accumulation point of the sequence generated by our algorithm is a stationary point of the DC problem we considered. The choice of parameters is general enough to cover those used in FISTA with fixed restart [26]. Additionally, by assuming that the objective is a level-bounded Kurdyka-Łojasiewicz function (see, for example, [4]) and the concave part is differentiable, we show that the whole sequence generated by our algorithm is globally convergent. We also establish the convergence rate of the algorithm by using the Kurdyka-Łojasiewicz exponent of an auxiliary function. Finally, we perform numerical experiments on ℓ_{1-2} [64] and logarithmic [19] regularized least squares problems. Our numerical experiments show that the pDCA$_e$ usually outperforms the proximal DCA with nonmonotone linesearch.

1.3 Outline of this thesis

Chapter 2 gives some preliminary materials which will be used in the following chapters. We first give an overview of some basic definitions and lemmas about nonsmooth analysis. Next, we introduce the definitions of Q-linear convergence and R-linear convergence, give a lemma which implies the relationship between them. Finally, we introduce the Kurdyka-Łojasiewicz (KL) property and uniformed KL property, which are very important for the convergence analysis of many algorithms.

Chapter 3 focuses on the proximal gradient algorithm with extrapolation for solving problem (1.1). We present the framework of proximal gradient algorithm with extrapolation for solving (1.1). Moreover, we show that any accumulation point of the sequence generated by the proximal gradient algorithm with extrapolation is a stationary point of the objective function F. Next, we introduce the error bound condition, and using this condition, we show that both the objective sequence and the iterate sequence are R-linearly convergent if the extrapolation coefficients are below a certain threshold. We further demonstrate that our theory can be applied to analyzing the convergence of FISTA with the fixed restart scheme for convex problems. In addtion, again for convex problems, we prove that the successive changes of the iterates go to zero for many choices of $\{\beta_t\}$ that approach the threshold: the choices are flexible enough to cover the choice of $\{\beta_t\}$ used in the FISTA. Finally, some numerical experiments are performed to illustrate our results.

Chapter 4 is devoted to analyzing the convergence behavior of the Proximal DCA with extrapolation (pDCA$_\mathrm{e}$) for solving a class of the DC problems. We first describe the DC problems we studied and present the proximal DCA with extrapolation for solving them. After that, we establish global subsequential convergence of the sequence generated by $\mathrm{pDCA_e}$. In addition, by assuming that an auxiliary function satisfies the Kurdyka-Łojasiewicz property and the the concave part is differentiable, we establish global convergence of the sequence generated by our algorithm and analyze its convergence rate. In the end of this chapter, we perform numerical experiments on ℓ_{1-2} [64] and logarithmic [19] regularized least squares problems.

Chapter 2

Preliminary materials

Before we start to study the proximal algorithms for solving the nonconvex nonsmooth problems, some basic notation and preliminary definitions and lemmas are given. We first review some knowledge of nonsmooth optimization we used in this thesis. Next, we give the definitions of two important classes of linear convergence in optimization and numerical analysis, which are Q-linear convergence and R-linear convergence. Finally, we recall the Kurdyka-Łojasiewicz (KL) property, which holds for many functions. And then, we give the definition of KL functions and uniformed KL property. This property will be used in our thesis for analyzing the convergence of proximal difference-of-convex algorithm with extrapolation.

2.1 Basic knowledge in nonsmooth optimization

We recall some basic definitions and notions of nonsmooth optimization in this subsection.

For a nonempty closed set $\mathcal{C} \subseteq \mathbb{R}^n$, its indicator function is defined by

$$\delta_{\mathcal{C}}(x) = \begin{cases} 0 & \text{if } \ x \in \mathcal{C}, \\ +\infty & \text{if } \ x \notin \mathcal{C}. \end{cases}$$

Moreover, we use $\operatorname{dist}(x, \mathcal{C})$ to denote the distance from x to $\mathcal{C}$, where $\operatorname{dist}(x, \mathcal{C}) = \inf_{y \in \mathcal{C}} \|x - y\|$. When $\mathcal{C}$ is in addition convex, we use $\operatorname{Proj}_{\mathcal{C}}(x)$ to denote the unique closest point on C to x.

The domain of an extended-real-valued function $h : \mathbb{R}^n \to [-\infty, \infty]$ is defined as $\mathrm{dom}\, h = \{x \in \mathbb{R}^n : \ h(x) < +\infty\}$. We say that h is proper if it never equals $-\infty$ and $\mathrm{dom}\, h \neq \emptyset$. Such a function is closed if it is lower semicontinuous. A proper closed function h is said to be level bounded if the lower level sets of h are bounded, i.e., the set $\{x \in \mathbb{R}^n : \ h(x) \leq r\}$ is bounded for any $r \in \mathbb{R}$. For a proper closed function $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, the (limiting) subdifferential of h at $x \in \mathrm{dom}\, h$ is given by

$$\partial h(x) = \left\{ v \in \mathbb{R}^n : \exists\ x^t \xrightarrow{h} x, v^t \to v \text{ with } \liminf_{y \to x^t} \frac{h(y) - h(x^t) - \langle v^t, y - x^t \rangle}{\|y - x^t\|} \geq 0 \ \ \forall t \right\}, \tag{2.1}$$

where $z \xrightarrow{h} x$ means $z \to x$ and $h(z) \to h(x)$. We also write $\mathrm{dom}\, \partial h := \{x \in \mathbb{R}^n : \ \partial h(x) \neq \emptyset\}$. The aforementioned subdifferential (2.1) is the same as the subdifferential defined in convex analysis when h is convex, i.e.,

$$\partial h(x) = \{v \in \mathbb{R}^n : \ h(y) - h(x) - \langle v, y - x \rangle \geq 0, \ \forall y \in \mathbb{R}^n\};$$

see, for example, [53, Proposition 8.12]. In addition, if h is continuously differentiable, then the subdifferential (2.1) is just ∇h. The partial gradient of a continuously differentiable h with respect to the i-th component of x is denoted by $\nabla_i h$.

For a proper closed convex function h, we use $\mathrm{Prox}_h(v)$ to denote the proximal operator of h at any $v \in \mathbb{R}^n$, i.e.,

$$\mathrm{Prox}_h(v) = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ h(x) + \frac{1}{2}\|x - v\|^2 \right\}.$$

We note that this operator is well defined for any $v \in \mathbb{R}^n$. One often used proximal operator of μh instead of h at $v \in \mathbb{R}^n$ is as follows:

$$\mathrm{Prox}_h(v) = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ h(x) + \frac{1}{2\mu}\|x - v\|^2 \right\},$$

where $\mu > 0$ is a constant. We refer the readers to [42] for more properties of this proximal operator.

In view of [53, Exercise 8.8(c)], we obtain that the following first-order necessary condition always holds for an optimal solution $\hat{x}$ of (1.1),

$$0 \in \nabla f(\hat{x}) + \partial g(\hat{x}). \tag{2.2}$$

One point $\tilde{x} \in \mathbb{R}^n$ is called a stationary point of (1.1) if $\tilde{x}$ satisfies (2.2). In particular, from the relation, we see that any optimal solution $\hat{x}$ of (1.1) is a stationary point of (1.1). Moreover, the set of stationary points of F in (1.1) is denoted by $\mathcal{X}$.

2.2 Some definitions about the linear convergence

In this section, we recall two notions of (local) linear convergence, which will be used in our convergence analysis in this thesis.

Definition 2.2.1. *For a sequence $\{x^t\}$, we say that $\{x^t\}$ converges Q-linearly to x^* if there exist $c \in (0,1)$ and $t_0 > 0$ such that*

$$\|x^{t+1} - x^*\| \leq c\|x^t - x^*\|, \quad \forall t \geq t_0;$$

Similarly, we give the following R-linear convergence of sequence $\{x^t\}$.

Definition 2.2.2. *We say that $\{x^t\}$ converges R-linearly to x^* if*

$$\limsup_{t\to\infty} \|x^t - x^*\|^{\frac{1}{t}} < 1.$$

The following fact states the relationship between the two notions of linear convergence, which will be used in our convergence analysis below.

Lemma 2.2.1. *Suppose that $\{a_t\}$ and $\{b_t\}$ are two sequences in $\mathbb{R}$ and $0 \leq b_t \leq a_t$ for all t. Suppose further that $\{a_t\}$ is Q-linearly convergent. Then $\{b_t\}$ is R-linearly convergent.*

2.3 KL property and uniformized KL property

This section introduces the Kurdyka-Łojasiewicz (KL) property [3, 4, 5, 13], which is an important tool in establishing the convergence of many first-order methods; see, for example, [4, 5]. Indeed, many functions including the semialgebraic functions satisfy this property.

Definition 2.3.1. (KL property) *For a proper closed function h, we say that the KL property holds for h at point $\hat{x} \in \operatorname{dom} \partial h$ if there exist a neighborhood $\mathcal{O}$ of $\hat{x}$, a positive number a, and a continuous concave function $\phi : [0, a) \to \mathbb{R}_+$ with $\phi(0) = 0$ such that:*

(i) *ϕ is continuously differentiable on $(0, a)$;*

(ii) *$\phi' > 0$ on $(0, a)$;*

(iii) *Take any $x \in \mathcal{O}$ which satisfies $h(\hat{x}) < h(x) < h(\hat{x}) + a$, one has*

$$1 \leq \phi'(h(x) - h(\hat{x})) \operatorname{dist}(0, \partial h(x)). \tag{2.3}$$

A function h is called a KL function, if it satisfies the KL property at all points in $\operatorname{dom} \partial h$ *.*

We next recall the following result provided in [14, Lemma 6] concerning the uniformized KL property. For notational simplicity, the set containing all concave continuous functions $\phi : [0, a) \to \mathbb{R}_+$ that are continuously differentiable on $(0, a)$ with $\phi' > 0$ and $\phi(0) = 0$ is denoted by Ξ_a.

Lemma 2.3.1. (Uniformized KL property) *Suppose that the function h is proper closed and the set Γ is compact. If h is a constant on Γ and satisfies the KL property at each point of Γ, then there exist $\epsilon, a > 0$ and $\phi \in \Xi_a$ such that for any $\hat{x} \in \Gamma$ and any x satisfying* $\operatorname{dist}(x, \Gamma) < \epsilon$ *and $h(\hat{x}) < h(x) < h(\hat{x}) + a$,*

$$1 \leq \phi'(h(x) - h(\hat{x}))\operatorname{dist}(0, \partial h(x)).$$

Chapter 3

Linear convergence of proximal gradient with extrapolation for solving a class of nonconvex nonsmooth problems

In this chapter, we mainly analyze the convergence behavior of proximal gradient algorithm with extrapolation for solving optimization problem (1.1).

We first present a very important fact for the differentiable function f in (1.1), and this fact will play an important role in establishing our convergence results. Then we present the proximal gradient algorithm with extrapolation. Next we construct an auxiliary sequence, which will be applied to proving the subsequential convergence of proximal gradient algorithm with extrapolation. Moreover, we introduce the error bound condition. Under this error bound condition, we obtain the R-linear convergence of sequence $\{x^t\}$ and $\{F(x^t)\}$ under the assumption that the extrapolation coefficients are below a certain threshold, where $\{x^t\}$ is generated by the proximal gradient algorithm with extrapolation. In addition, we show that when f is convex, the threshold reduces to 1, and FISTA with fixed restart is a special case of our algorithm. Furthermore, we show that the successive changes of the iterates go to zero for many choices of $\{\beta_t\}$ that approach the threshold 1: the choices are flexible enough to cover the choice of $\{\beta_t\}$ used in the FISTA. Finally, we perform some numerical experiments to illustrate our results.

3.1 Proximal gradient algorithm with extrapolation

In this section, we first give a fact of the differentiable function f and then present the proximal gradient algorithm with extrapolation for solving (1.1). After that, we introduce an auxiliary sequence which will be used for the convergence analysis below.

We recall that in our problem (1.1), the function g is proper closed convex and f has a Lipschitz continuous gradient; moreover, $\inf F > -\infty$ and $\mathcal{X} \neq \emptyset$. Furthermore, we observe that any function f whose gradient is Lipschitz continuous can be written as $f = f_1 - f_2$, where f_1 and f_2 are two convex functions with Lipschitz continuous gradients. For instance, one can decompose f as

$$f(x) = \underbrace{f(x) + \frac{c}{2}\|x\|^2}_{f_1(x)} - \underbrace{\frac{c}{2}\|x\|^2}_{f_2(x)},$$

for any $c \geq L_f$, where L_f is a Lipschitz continuity modulus of ∇f. It is then routine to show that both f_1 and f_2 are convex functions with Lipschitz continuous gradients.

Thus, without loss of generality, from now on, we assume that $f = f_1 - f_2$ for some convex functions f_1 and f_2 with Lipschitz continuous gradients. For concreteness, we denote a Lipschitz continuity modulus of ∇f_1 by $L > 0$, and a Lipschitz continuity modulus of ∇f_2 by $l \geq 0$. Moreover, by taking a larger L if necessary, we assume throughout that $L \geq l$. Then it is not hard to show that ∇f is Lipschitz continuous with a modulus L.

We are now ready to present our algorithm studied in this chapter.

```
PG_e: Proximal gradient algorithm with extrapolation
Input: $x^0 \in \operatorname{dom} g$, $\{\beta_t\} \subseteq \left[0, \sqrt{\frac{L}{L+l}}\right]$. Set $x^{-1} = x^0$.
    for $t = 0, 1, 2, \cdots$ do
        $y^t = x^t + \beta_t(x^t - x^{t-1}),$
        $x^{t+1} = \operatorname{Prox}_{\frac{1}{L}g}\left(y^t - \frac{1}{L}\nabla f(y^t)\right).$        (3.1)
    end for
```

In the following contexts of this chapter, we shall discuss the convergence behavior of PG$_{\mathrm{e}}$. According to the definition of proximal operator presented in Subsection 2.1, the x-update in (3.1) is immediately given by

$$x^{t+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \langle \nabla f(y^t), x \rangle + \frac{L}{2}\|x - y^t\|^2 + g(x) \right\}. \tag{3.2}$$

We will use this relation repeatedly in our convergence analysis below. Next, we introduce an auxiliary sequence $\{H_{t,\alpha}\}$,

$$H_{t,\alpha} = F(x^t) + \alpha\|x^t - x^{t-1}\|^2, \tag{3.3}$$

where α is a positive number in the interval $[\frac{L+l}{2}\bar{\beta}^2, \frac{L}{2}]$, $\bar{\beta} := \sup_t \beta_t$ and $\{x^t\}$ is generated by PG$_{\mathrm{e}}$. In the next Subsection 3.2.1, we mainly consider the convergence properties of $\{H_{t,\alpha}\}$. The corresponding results obtained will then be applied to establishing the convergence of $\{x^t\}$ and $\{F(x^t)\}$. Similar auxiliary sequences (3.3) were also used in [6, 21, 33] for analyzing (1.2) with different extrapolation coefficients.

3.2 Convergence analysis of PG$_{\mathrm{e}}$

We first give some lemmas about the auxiliary sequence $\{H_{t,\alpha}\}$ which is defined as (3.3).

3.2.1 Some lemmas

Lemma 3.2.1. *Let $\{x^t\}$ be a sequence generated by PG$_\mathrm{e}$ and $\alpha \in [\frac{L+l}{2}\bar{\beta}^2, \frac{L}{2}]$. Then we have the following results.*

(i) *Given any fixed point $z \in \mathrm{dom}\ g$, we obtain that*

$$F(x^{t+1}) \le F(z) + \frac{L+l}{2}\|z - y^t\|^2 - \frac{L}{2}\|x^{t+1} - z\|^2. \tag{3.4}$$

(ii) *It holds that for all t,*

$$H_{t+1,\alpha} - H_{t,\alpha} \le \left(-\frac{L}{2} + \alpha\right)\|x^{t+1} - x^t\|^2 + \left(\frac{L+l}{2}\beta_t^2 - \alpha\right)\|x^t - x^{t-1}\|^2. \tag{3.5}$$

(iii) *The sequence $\{H_{t,\alpha}\}$ is nonincreasing.*

Proof. We first prove (i). Take any fixed $z \in \mathrm{dom}\ g$. From the definition of x^{t+1} in (3.2) and the fact that the objective function in the minimization problem (3.2) is strongly convex, we obtain that

$$\begin{aligned}&\langle \nabla f(y^t), x^{t+1}\rangle + \frac{L}{2}\|x^{t+1} - y^t\|^2 + g(x^{t+1})\\ &\le \langle \nabla f(y^t), z\rangle + \frac{L}{2}\|z - y^t\|^2 + g(z) - \frac{L}{2}\|x^{t+1} - z\|^2.\end{aligned}$$

By rearranging terms, we see further that

$$\begin{aligned}g(x^{t+1}) \le g(z) + \langle -\nabla f(y^t), x^{t+1} - z\rangle + \frac{L}{2}\|z - y^t\|^2\\ - \frac{L}{2}\|x^{t+1} - y^t\|^2 - \frac{L}{2}\|x^{t+1} - z\|^2.\end{aligned} \tag{3.6}$$

On the other hand, from the fact that f has a Lipschitz continuous gradient with a Lipschitz continuity modulus L, we obtain that

$$f(x^{t+1}) \le f(y^t) + \langle \nabla f(y^t), x^{t+1} - y^t\rangle + \frac{L}{2}\|x^{t+1} - y^t\|^2. \tag{3.7}$$

Summing (3.6) and (3.7), we see further that

$$\begin{aligned} f(x^{t+1}) + g(x^{t+1}) \le f(y^t) + g(z) + \langle \nabla f(y^t), z - y^t \rangle \\ + \frac{L}{2}\|z - y^t\|^2 - \frac{L}{2}\|x^{t+1} - z\|^2. \end{aligned} \tag{3.8}$$

Next, using the fact that $f = f_1 - f_2$, we have

$$\begin{aligned} & f(y^t) + \langle \nabla f(y^t), z - y^t \rangle \\ & = f_1(y^t) - f_2(y^t) + \langle \nabla f_1(y^t), z - y^t \rangle - \langle \nabla f_2(y^t), z - y^t \rangle. \end{aligned} \tag{3.9}$$

Since f_1 is convex and continuously differentiable, we immediately obtain that

$$f_1(y^t) + \langle \nabla f_1(y^t), z - y^t \rangle \le f_1(z). \tag{3.10}$$

Using the fact that ∇f_2 is Lipschitz continuous with a modulus l, we have

$$f_2(z) - f_2(y^t) - \langle \nabla f_2(y^t), z - y^t \rangle \le \frac{l}{2}\|z - y^t\|^2. \tag{3.11}$$

Combining (3.10), (3.11) with (3.9) and recalling that $f = f_1 - f_2$, we see further that

$$f(y^t) + \langle \nabla f(y^t), z - y^t \rangle \le f(z) + \frac{l}{2}\|z - y^t\|^2. \tag{3.12}$$

Summing (3.8) and (3.12), and recalling that $F = f + g$, we obtain (3.4) immediately. This proves (i).

We now prove (ii). From the definition of the y-update in (3.1), we see that $y^t - x^t = \beta_t(x^t - x^{t-1})$. Using this and (3.4) with $z = x^t$, we obtain that

$$F(x^{t+1}) - F(x^t) \le \frac{L + l}{2}\beta_t^2\|x^t - x^{t-1}\|^2 - \frac{L}{2}\|x^{t+1} - x^t\|^2.$$

Combining this with the definition of $H_{t,\alpha}$ from (3.3), we see further that

$$\begin{aligned}
&H_{t+1,\alpha} - H_{t,\alpha} = F(x^{t+1}) + \alpha\|x^{t+1} - x^t\|^2 - F(x^t) - \alpha\|x^t - x^{t-1}\|^2 \\
&= F(x^{t+1}) - F(x^t) + \alpha\|x^{t+1} - x^t\|^2 - \alpha\|x^t - x^{t-1}\|^2 \\
&\le -\frac{L}{2}\|x^{t+1} - x^t\|^2 + \frac{L+l}{2}\beta_t^2\|x^t - x^{t-1}\|^2 + \alpha\|x^{t+1} - x^t\|^2 - \alpha\|x^t - x^{t-1}\|^2 \\
&= \left(-\frac{L}{2} + \alpha\right)\|x^{t+1} - x^t\|^2 + \left(\frac{L+l}{2}\beta_t^2 - \alpha\right)\|x^t - x^{t-1}\|^2,
\end{aligned}$$

which is just (3.5). This proves (ii). Finally, using the assumption that $\frac{L+l}{2}\bar{\beta}^2 \le \alpha \le \frac{L}{2}$, we have

$$-\frac{L}{2} + \alpha \le 0, \text{ and } \frac{L+l}{2}\beta_t^2 - \alpha \le \frac{L+l}{2}\bar{\beta}^2 - \alpha \le 0 \ \ \forall t.$$

Consequently, $H_{t+1,\alpha} - H_{t,\alpha} \le 0$, i.e., $\{H_{t,\alpha}\}$ is nonincreasing, which completes the proof. □

In view of Lemma 3.2.1, we immediately obtain the following corollary.

Corollary 3.2.2. *Suppose that F in* (1.1) *is level bounded. Then the sequence $\{x^t\}$ generated by PG_{e} is bounded.*

Proof. Take $\alpha = \frac{L}{2}$ in the sequence $\{H_{t,\alpha}\}$. From Lemma 3.2.1, we see that the sequence $\{H_{t,\frac{L}{2}}\}$ is nonincreasing. From this and the definition of $H_{t,\frac{L}{2}}$, we have

$$F(x^t) \le H_{t,\frac{L}{2}} \le H_{0,\frac{L}{2}} < \infty.$$

Since F is level bounded by assumption, we conclude that $\{x^t\}$ is bounded. □

Lemma 3.2.3. *Let $\{x^t\}$ be a sequence generated by PG_{e}, and $\alpha \in [\frac{L+l}{2}\bar{\beta}^2, \frac{L}{2}]$. Then we have the following results.*

(i) *The sequence $\{H_{t,\alpha}\}$ is convergent.*

(ii) $\sum_{t=0}^{\infty}\left(\alpha - \frac{L+l}{2}\beta_{t+1}^2\right)\|x^{t+1} - x^t\|^2 < \infty$.

Proof. Since $\inf F > -\infty$ from our assumption, we immediately deduce that $H_{t,\alpha} = F(x^t) + \alpha\|x^t - x^{t-1}\|^2$ is bounded from below. This together with the fact that $\{H_{t,\alpha}\}$ is nonincreasing from Lemma 3.2.1 implies that $\{H_{t,\alpha}\}$ is convergent, which proves (i).

We now prove (ii). using the fact $-\frac{L}{2} + \alpha \leq 0$, we have from (3.5) that

$$H_{t+1,\alpha} - H_{t,\alpha} \leq -\left(\alpha - \frac{L+l}{2}\beta_t^2\right)\|x^t - x^{t-1}\|^2. \tag{3.13}$$

Summing both sides of (3.13) from 1 to N, we have

$$0 \leq \sum_{t=1}^{N}\left(\alpha - \frac{L+l}{2}\beta_t^2\right)\|x^t - x^{t-1}\|^2 \leq \sum_{t=1}^{N}(H_{t,\alpha} - H_{t+1,\alpha}) = H_{1,\alpha} - H_{N+1,\alpha}, \tag{3.14}$$

where the nonnegativity is a consequence of the fact that $\alpha \geq \frac{L+l}{2}\bar{\beta}^2 \geq \frac{L+l}{2}\beta_t^2$ for all t. From (i), we see that $\{H_{t,\alpha}\}$ is convergent. Hence, letting $N \to \infty$ in (3.14), we conclude that

$$\begin{aligned}&\sum_{t=1}^{\infty}\left(\alpha - \frac{L+l}{2}\beta_t^2\right)\|x^t - x^{t-1}\|^2\\ &\leq H_{1,\alpha} - \lim_{N\to\infty} H_{N+1,\alpha} < \infty,\end{aligned}$$

which completes the proof. □

In the following lemma, we show that when $\{\beta_t\}$ is chosen below a certain threshold, any accumulation point of the sequence $\{x^t\}$ generated by $\mathrm{PG_e}$, if exists, is a stationary point of F. This result has been proved in [33] when the function f is convex. Indeed, in the convex case, it was shown in [33, Theorem 4.1] that the whole sequence $\{x^t\}$ is convergent. However, the following convergence result is new when the function f is nonconvex.

Lemma 3.2.4. *Suppose that $\bar{\beta} < \sqrt{\frac{L}{L+l}}$ and $\{x^t\}$ is a sequence generated by PG$_{\mathrm{e}}$. Then we have the following results.*

(i) $\sum_{k=0}^{\infty} \|x^{t+1} - x^t\|^2 < \infty$.

(ii) *Any accumulation point of $\{x^t\}$ is a stationary point of F.*

Proof. Since $\bar{\beta} < \sqrt{\frac{L}{L+l}}$, one can choose a fixed $\alpha \in (\frac{L+l}{2}\bar{\beta}^2, \frac{L}{2})$. Hence, $\frac{L+l}{2}\beta_t^2 \leq \frac{L+l}{2}\bar{\beta}^2 < \alpha$ for all t. Combining this with Lemma 3.2.3 (ii), we see that

$$
\begin{aligned}
0 &< \left(\alpha - \frac{L+l}{2}\bar{\beta}^2\right) \sum_{t=0}^{\infty} \|x^{t+1} - x^t\|^2 \\
&= \sum_{t=0}^{\infty} \left(\alpha - \frac{L+l}{2}\bar{\beta}^2\right) \|x^{t+1} - x^t\|^2 \\
&\leq \sum_{t=0}^{\infty} \left(\alpha - \frac{L+l}{2}\beta_{t+1}^2\right) \|x^{t+1} - x^t\|^2 < \infty,
\end{aligned}
$$

which immediately implies that the conclusion in (i) holds.

We next prove (ii). Choose any fixed accumulation point $\bar{x}$ of the sequence $\{x^t\}$, hence there exists a subsequence $\{x^{t_i}\}$ such that $\lim\limits_{i\to\infty} x^{t_i} = \bar{x}$. From the x-update (3.2), we have

$$x^{t_i+1} = \operatorname*{argmin}_{x\in\mathbb{R}^n} \left\{ \langle \nabla f(y^{t_i}), x\rangle + \frac{L}{2}\|x - y^{t_i}\|^2 + g(x) \right\}. \tag{3.15}$$

Using the first-order optimality condition of the minimization problem (3.15), we see further that

$$-L(x^{t_i+1} - y^{t_i}) \in \nabla f(y^{t_i}) + \partial g(x^{t_i+1}).$$

From this and the definition of y^{t_i}, which is $y^{t_i} = x^{t_i} + \beta_{t_i}(x^{t_i} - x^{t_i-1})$, we see further that

$$- L[(x^{t_i+1} - x^{t_i}) - \beta_{t_i}(x^{t_i} - x^{t_i-1})] \in \nabla f(y^{t_i}) + \partial g(x^{t_i+1}). \tag{3.16}$$

Letting i go to ∞ on both sides of (3.16), and recalling that $\|x^{t_i+1} - x^{t_i}\| \to 0$ from (i) together with the facts that ∇f is continuous and ∂g is closed (see, for example, [15, Page 80]), we obtain that

$$0 \in \nabla f(\bar{x}) + \partial g(\bar{x}),$$

which completes the proof. □

Define Ω as the set of accumulation points of the sequence $\{x^t\}$ generated by PG$_\mathrm{e}$. We note from Corollary 3.2.2 and Lemma 3.2.4 (ii) that $\emptyset \neq \Omega \subseteq \mathcal{X}$ when F is level bounded. In the next proposition, we show that F is constant over Ω if $\{\beta_t\}$ is chosen below a certain threshold. Since F is only assumed to be lower semicontinuous, this result is nontrivial when F has stationary points that are not globally optimal.

Proposition 3.2.5. *Suppose that $\bar{\beta} < \sqrt{\frac{L}{L+l}}$ and $\{x^t\}$ is a sequence generated by PG$_\mathrm{e}$ with its set of accumulation points denoted by Ω. Then $\zeta := \lim\limits_{t\to\infty} F(x^t)$ exists and $F \equiv \zeta$ on Ω.*

Proof. Since $\bar{\beta} < \sqrt{\frac{L}{L+l}}$, we take any fixed $\alpha \in (\frac{L+l}{2}\bar{\beta}^2, \frac{L}{2})$. According to Lemmas 3.2.3 and 3.2.4, the sequence $\{H_{t,\alpha}\}$ is convergent and $\|x^{t+1} - x^t\| \to 0$. Using these results and recalling that the definition of $H_{t,\alpha}$ is that $H_{t,\alpha} = F(x^t) + \alpha\|x^t - x^{t-1}\|^2$, we can easily obtain that $\lim\limits_{t\to\infty} F(x^t)$ exists. We denote this limit by ζ.

We now prove the second part, i.e., $F \equiv \zeta$ on Ω. If $\Omega = \emptyset$, then the conclusion holds trivially. Otherwise, take any $\hat{x} \in \Omega$, by the definition, there must exist a subsequence $\{x^{t_i}\}$ such that $\lim\limits_{i\to\infty} x^{t_i} = \hat{x}$. Then we conclude from the definition of ζ and the fact that F is lower semicontinuous that

$$F(\hat{x}) \leq \liminf_{i\to\infty} F(x^{t_i}) = \zeta. \tag{3.17}$$

On the other hand, using the definition of x^{t_i} as the minimizer in (3.2) and rearranging terms, we have

$$g(x^{t_i}) + \langle \nabla f(y^{t_i-1}), x^{t_i} - \hat{x} \rangle + \frac{L}{2}\|x^{t_i} - y^{t_i-1}\|^2 \le g(\hat{x}) + \frac{L}{2}\|\hat{x} - y^{t_i-1}\|^2. \tag{3.18}$$

Adding $f(x^{t_i})$ to both sides of (3.18), we obtain further that

$$f(x^{t_i}) + g(x^{t_i}) + \langle \nabla f(y^{t_i-1}), x^{t_i} - \hat{x} \rangle + \frac{L}{2}\|x^{t_i} - y^{t_i-1}\|^2 \le f(x^{t_i}) + g(\hat{x}) + \frac{L}{2}\|\hat{x} - y^{t_i-1}\|^2. \tag{3.19}$$

Next, recall that $y^{t_i-1} = x^{t_i-1} + \beta_{t_i-1}(x^{t_i-1} - x^{t_i-2})$. Thus, we have

$$\begin{aligned} \|x^{t_i} - y^{t_i-1}\| &= \|x^{t_i} - x^{t_i-1} - \beta_{t_i-1}(x^{t_i-1} - x^{t_i-2})\| \\ &\le \|x^{t_i} - x^{t_i-1}\| + \bar{\beta}\|x^{t_i-1} - x^{t_i-2}\|, \end{aligned} \tag{3.20}$$

where the inequality is obtained by using the triangle inequality and the fact $\beta_{t_i-1} \le \bar{\beta} = \sup_t \beta_t$. Similarly, we have

$$\begin{aligned} \|\hat{x} - y^{t_i-1}\| &= \|\hat{x} - x^{t_i} + x^{t_i} - y^{t_i-1}\| \\ &\le \|\hat{x} - x^{t_i}\| + \|x^{t_i} - y^{t_i-1}\|. \end{aligned} \tag{3.21}$$

Since $\|x^{t+1} - x^t\| \to 0$ and $\lim\limits_{i\to\infty} x^{t_i} = \hat{x}$, it follows from (3.20) and (3.21) that

$$\|x^{t_i} - y^{t_i-1}\| \to 0 \text{ and } \|\hat{x} - y^{t_i-1}\| \to 0,$$

and hence $\nabla f(y^{t_i-1}) \to \nabla f(\hat{x})$, which follows from the continuity of ∇f. From these and (3.19), we obtain that

$$\zeta = \limsup_{i\to\infty} F(x^{t_i}) \le F(\hat{x}). \tag{3.22}$$

Thus $F(\hat{x}) = \lim\limits_{i\to\infty} F(x^{t_i}) = \zeta$ from (3.17) and (3.22). Since $\hat{x} \in \Omega$ is arbitrary, we see that $F \equiv \zeta$ on Ω, which completes the proof. □

3.2.2 R-linear convergence of $\{x^t\}$ and $\{F(x^t)\}$

This subsection establishes the R-linear convergence of $\{x^t\}$ and $\{F(x^t)\}$ under the following assumption. We first introduce this assumption.

Assumption 1. (i) **(Error bound condition)** *Given any fixed* $\xi \geq \inf_{x\in\mathbb{R}^n} F(x)$*, there exist* $\epsilon > 0$ *and* $\tau > 0$ *such that*

$$\operatorname{dist}(x, \mathcal{X}) \leq \tau \left\| \operatorname{Prox}_{\frac{1}{L}g}\left(x - \frac{1}{L}\nabla f(x)\right) - x \right\|,$$

whenever $\|\operatorname{Prox}_{\frac{1}{L}g}(x - \frac{1}{L}\nabla f(x)) - x\| < \epsilon$ *and* $F(x) \leq \xi$.

(ii) *There exists* $\delta > 0$*, such that* $\|x - y\| \geq \delta$ *whenever* $x, y \in \mathcal{X}$*,* $F(x) \neq F(y)$.

The above assumption has been used in the convergence analysis of many algorithms, including the gradient projection and block coordinate gradient descent method, etc; see, for example, [9, 37, 38, 39, 59, 60, 61] and the references therein. The assumption consists of two parts: the first part is an error bound condition, while the second part states that when restricted to $\mathcal{X}$, the isocost surfaces of F are properly separated.

Under our blanket assumptions on F, Assumption 1 is known to be satisfied for interesting choices of f and g, including:

- $f(x) = h(Ax)$, g is a polyhedral function, where h is a twice continuously differentiable function on $\mathbb{R}^n$ and ∇h is Lipschitz continuous, and on any compact convex set, h is strongly convex; see, [37, Theorem 2.1] and [60, Lemma 6]. This covers the well-known LASSO;

- f is a quadratic function (possibly nonconvex), g is a polyhedral function; see, for example, [60, Theorem 4].

The first example is a convex problem, while the second one is possibly nonconvex. We refer the readers to [60, 61, 67] and the references therein for more examples and discussions on the error bound condition.

We next show that the auxiliary sequence $\{H_{t,\alpha}\}$ is Q-linearly convergent under Assumption 1. Similar idea has been used for analyzing the convergence behavior of a class of block coordinate gradient descent methods in [60, Theorem 2].

Lemma 3.2.6. *Suppose that $\bar\beta < \sqrt{\frac{L}{L+l}}$, $\alpha \in (\frac{L+l}{2}\bar\beta^2, \frac{L}{2})$ and that Assumption 1 holds. Suppose further that $\{x^t\}$ is a sequence generated by PG$_{\rm e}$. Then we have the following results.*

(i) $\lim\limits_{t\to\infty} \operatorname{dist}(x^t, \mathcal{X}) = 0$.

(ii) *The sequence $\{H_{t,\alpha}\}$ is Q-linearly convergent.*

Proof. We first prove (i). Using the triangle inequality, we observe that

$$\begin{aligned}
&\left\|\operatorname{Prox}_{\frac{1}{L}g}\left(x^t - \frac{1}{L}\nabla f(x^t)\right) - x^t\right\| \\
&\le \left\|\operatorname{Prox}_{\frac{1}{L}g}\left(x^t - \frac{1}{L}\nabla f(x^t)\right) - \operatorname{Prox}_{\frac{1}{L}g}\left(y^t - \frac{1}{L}\nabla f(y^t)\right)\right\| \\
&\quad + \left\|\operatorname{Prox}_{\frac{1}{L}g}\left(y^t - \frac{1}{L}\nabla f(y^t)\right) - x^t\right\| \qquad (3.23)\\
&\le \left\|\operatorname{Prox}_{\frac{1}{L}g}\left(x^t - \frac{1}{L}\nabla f(x^t)\right) - \operatorname{Prox}_{\frac{1}{L}g}\left(y^t - \frac{1}{L}\nabla f(y^t)\right)\right\| \\
&\quad + \left\|\operatorname{Prox}_{\frac{1}{L}g}\left(y^t - \frac{1}{L}\nabla f(y^t)\right) - y^t\right\| + \|y^t - x^t\|.
\end{aligned}$$

We now derive an upper bound for the first term on the right hand side of (3.23). To this end, using the nonexpansiveness property of the proximal operator (see, for example, [52, Page 340]), we have

$$\begin{aligned}
&\left\|\operatorname{Prox}_{\frac{1}{L}g}\left(x^t - \frac{1}{L}\nabla f(x^t)\right) - \operatorname{Prox}_{\frac{1}{L}g}\left(y^t - \frac{1}{L}\nabla f(y^t)\right)\right\| \\
&\le \left\|x^t - \frac{1}{L}\nabla f(x^t) - y^t + \frac{1}{L}\nabla f(y^t)\right\| \qquad (3.24) \\
&\le \|x^t - y^t\| + \frac{1}{L}\|\nabla f(x^t) - \nabla f(y^t)\| \le 2\|x^t - y^t\|,
\end{aligned}$$

where the last inequality follows from the Lipschitz continuity of ∇f with a modulus L. Combining (3.23), (3.24) and noticing that the $x^{t+1} = \operatorname{Prox}_{\frac{1}{L}g}\left(y^t - \frac{1}{L}\nabla f(y^t)\right)$ from $\mathrm{PG_e}$, we see further that

$$\begin{aligned}
&\left\|\operatorname{Prox}_{\frac{1}{L}g}\left(x^t - \frac{1}{L}\nabla f(x^t)\right) - x^t\right\| \\
&\le 3\|x^t - y^t\| + \|x^{t+1} - y^t\| \\
&\le 3\|x^t - y^t\| + \|x^{t+1} - x^t\| + \|x^t - y^t\| \qquad (3.25) \\
&= 4\|x^t - y^t\| + \|x^{t+1} - x^t\| \\
&= 4\beta_t\|x^t - x^{t-1}\| + \|x^{t+1} - x^t\| \\
&\le 4\bar{\beta}\|x^t - x^{t-1}\| + \|x^{t+1} - x^t\|,
\end{aligned}$$

where the second equality and the last inequality follows from the definition of y^t in (3.1) and the definition of $\bar{\beta}$ repectively. Since $\|x^{t+1} - x^t\| \to 0$ by Lemma 3.2.4, we conclude from (3.25) that

$$\left\|\operatorname{Prox}_{\frac{1}{L}g}\left(x^t - \frac{1}{L}\nabla f(x^t)\right) - x^t\right\| \to 0. \qquad (3.26)$$

Let $\xi = H_{0,\alpha}$. Since $\{H_{t,\alpha}\}$ is nonincreasing from Lemma 3.2.1, we must have $H_{t,\alpha} \le \xi$ for all t. And recalling the definition of $\{H_{t,\alpha}\}$, we consequently obtain that $F(x^t) \le \xi$ for all t. In view of this, (3.26) and Assumption 1 (i), we see that for $\xi = H_{0,\alpha}$, there exist $\tau > 0$ and a positive integer T so that for all $t \ge T$, we deduce that

$$\operatorname{dist}(x^t, \mathcal{X}) \le \tau \left\|\operatorname{Prox}_{\frac{1}{L}g}\left(x^t - \frac{1}{L}\nabla f(x^t)\right) - x^t\right\|. \qquad (3.27)$$

Thus from (3.26) and (3.27), we immediately obtain the conclusion in (i).

We now prove (ii). Take an arbitrary $z \in \mathcal{X}$, we have from (3.4) that

$$\begin{aligned} F(x^{t+1}) &\leq F(z) + \frac{L+l}{2}\|z - y^t\|^2 - \frac{L}{2}\|x^{t+1} - z\|^2 \\ &\leq F(z) + \frac{L+l}{2}\|z - y^t\|^2 \\ &= F(z) + \frac{L+l}{2}\|z - x^t + x^t - y^t\|^2 \\ &\leq F(z) + (L+l)\|z - x^t\|^2 + (L+l)\|x^t - y^t\|^2. \end{aligned} \tag{3.28}$$

Choose z in (3.28) as an $\bar{x}^t \in \mathcal{X}$ so that $\|\bar{x}^t - x^t\| = \mathrm{dist}(x^t, \mathcal{X})$. Then using this and (3.28), we see further that

$$F(x^{t+1}) - F(\bar{x}^t) \leq (L+l)\mathrm{dist}^2(x^t, \mathcal{X}) + (L+l)\|x^t - y^t\|^2. \tag{3.29}$$

In addition, we observe from the definition of $\bar{x}^t$ that

$$\begin{aligned} \|\bar{x}^{t+1} - \bar{x}^t\| &\leq \|\bar{x}^{t+1} - x^{t+1}\| + \|x^{t+1} - x^t\| + \|x^t - \bar{x}^t\| \\ &= \mathrm{dist}(x^{t+1}, \mathcal{X}) + \mathrm{dist}(x^t, \mathcal{X}) + \|x^{t+1} - x^t\|. \end{aligned} \tag{3.30}$$

Recalling that $\|x^{t+1} - x^t\| \to 0$ by Lemma 3.2.4. This together with (3.26), (3.27) and (3.30) shows that $\|\bar{x}^{t+1} - \bar{x}^t\| \to 0$. From this and Assumption 1 (ii), it must then hold true that $F(\bar{x}^t) \equiv \zeta$ for all sufficiently large t, where ζ is a positive constant. Thus, for all sufficiently large t, we obtain from (3.29) that

$$F(x^{t+1}) - \zeta \leq (L+l)\mathrm{dist}^2(x^t, \mathcal{X}) + (L+l)\|x^t - y^t\|^2. \tag{3.31}$$

On the other hand, in view of the fact that $\bar{x}^t$ is a stationary point of (1.1), we immediately have $-\nabla f(\bar{x}^t) \in \partial g(\bar{x}^t)$. Using the convexity of g, we see further that for all t,

$$g(\bar{x}^t) - g(x^t) \leq \langle -\nabla f(\bar{x}^t), \bar{x}^t - x^t \rangle.$$

Combining the above relation with the definitions of F, $H_{t,\alpha}$ and ζ, we see that

$$\begin{aligned}
\zeta - H_{t,\alpha} &= F(\bar{x}^t) - F(x^t) - \alpha\|x^t - x^{t-1}\|^2 \\
&= f(\bar{x}^t) + g(\bar{x}^t) - f(x^t) - g(x^t) - \alpha\|x^t - x^{t-1}\|^2 \\
&\le f(\bar{x}^t) - f(x^t) + \langle -\nabla f(\bar{x}^t), \bar{x}^t - x^t\rangle - \alpha\|x^t - x^{t-1}\|^2 \\
&= -f(x^t) - [-f(\bar{x}^t)] - \langle -\nabla f(\bar{x}^t), x^t - \bar{x}^t\rangle - \alpha\|x^t - x^{t-1}\|^2 \\
&\le \frac{L}{2}\|x^t - \bar{x}^t\|^2 - \alpha\|x^t - x^{t-1}\|^2
\end{aligned}$$

for all sufficiently large t, the last inequality follows from the fact that $-\nabla f$ is Lipschitz continuous with a modulus L. Using this, the fact that $\|x^{t+1} - x^t\| \to 0$ by Lemma 3.2.4 and the conclusion $\|\bar{x}^t - x^t\| = \mathrm{dist}(x^t, \mathcal{X}) \to 0$ by (i), we deduce that

$$\zeta \le \lim_{k\to\infty} H_{t,\alpha} = \inf_k H_{t,\alpha}, \tag{3.32}$$

where the equality follows from Lemma 3.2.1 (iii).

Now, combining (3.25), (3.27) with (3.31) together, we see that for all sufficiently large t,

$$\begin{aligned}
F(x^{t+1}) - \zeta &\le (L+l)\mathrm{dist}^2(x^t, \mathcal{X}) + (L+l)\|x^t - y^t\|^2 \\
&\le (L+l)\tau^2(4\bar{\beta}\|x^t - x^{t-1}\| + \|x^{t+1} - x^t\|)^2 + (L+l)\|x^t - y^t\|^2 \\
&\le (L+l)\tau^2(4\bar{\beta}\|x^t - x^{t-1}\| + \|x^{t+1} - x^t\|)^2 + (L+l)\bar{\beta}^2\|x^t - x^{t-1}\|^2 \\
&\le C(\|x^t - x^{t-1}\|^2 + \|x^{t+1} - x^t\|^2),
\end{aligned}$$

for some positive constant C, the third inequality in the above formulation follows from the definition of y^t in (3.1) and the definition of $\bar{\beta}$. Using this fact and the definition of $H_{t,\alpha}$, we see further that

$$0 \le H_{t+1,\alpha} - \zeta \le \eta(\|x^t - x^{t-1}\|^2 + \|x^{t+1} - x^t\|^2), \tag{3.33}$$

where $\eta = C + \alpha$, and the nonnegativity is a consequence of (3.32). On the other hand, let $\delta = \min\left\{\frac{L}{2} - \alpha, \alpha - \frac{L+l}{2}\bar{\beta}^2\right\}$. Then $\delta > 0$ and we see from (3.5) that

$$(H_{t+1,\alpha} - \zeta) - (H_{t,\alpha} - \zeta) \le -\delta(\|x^{t+1} - x^t\|^2 + \|x^t - x^{t-1}\|^2). \tag{3.34}$$

Combining (3.34) and (3.33), we obtain further that for sufficiently large t

$$(H_{t+1,\alpha} - \zeta) - (H_{t,\alpha} - \zeta) \leq -\frac{\delta}{\eta}(H_{t+1,\alpha} - \zeta). \tag{3.35}$$

Reorganizing (3.35), we see that for all sufficiently large t,

$$0 \leq H_{t+1,\alpha} - \zeta \leq \frac{1}{1+\frac{\delta}{\eta}}(H_{t,\alpha} - \zeta),$$

which implies that the sequence $\{H_{t,\alpha}\}$ is Q-linearly convergent. This completes the proof. □

We are now ready to prove the local linear convergence of the sequences $\{x^t\}$ and $\{F(x^t)\}$, using the Q-linear convergence of $\{H_{t,\alpha}\}$.

Theorem 3.2.7. *Suppose that $\bar{\beta} < \sqrt{\frac{L}{L+l}}$ and that Assumption 1 holds. Let $\{x^t\}$ be a sequence generated by PG*$_{\mathrm{e}}$*. Then we have the following results.*

(i) *The sequence $\{x^t\}$ is R-linearly convergent to a stationary point of F.*

(ii) *The sequence $\{F(x^t)\}$ is R-linearly convergent.*

Proof. Since $\bar{\beta} < \sqrt{\frac{L}{L+l}}$, we choose a fixed $\alpha \in (\frac{L+l}{2}\bar{\beta}^2, \frac{L}{2})$. Then, according to Lemma 3.2.6, the sequence $\{H_{t,\alpha}\}$ is Q-linearly convergent. For notational simplicity, we denote its limit by ζ. Take $\delta = \min\left\{\frac{L}{2} - \alpha, \alpha - \frac{L+l}{2}\bar{\beta}^2\right\}$. Then $\delta > 0$ and we obtain from (3.5) that

$$\|x^{t+1} - x^t\|^2 \leq \frac{1}{\delta}(H_{t,\alpha} - \zeta) - \frac{1}{\delta}(H_{t+1,\alpha} - \zeta) \leq \frac{1}{\delta}(H_{t,\alpha} - \zeta), \tag{3.36}$$

where the last inequality follows from the fact that the sequence $\{H_{t,\alpha}\}$ is nonincreasing and convergent to ζ, thanks to Lemmas 3.2.1 and 3.2.3. Combining the above inequality with the fact that the sequence $\{H_{t,\alpha}\}$ is Q-linearly convergent, we see that there exist $0 < c < 1$ and $M > 0$ such that

$$\|x^{t+1} - x^t\| \leq Mc^t \tag{3.37}$$

for all t. Consequently, for any $m_2 > m_1 \geq 1$, we have

$$\|x^{m_2} - x^{m_1}\| \leq \sum_{k=m_1}^{m_2-1} \|x^{t+1} - x^t\| \leq \frac{Mc^{m_1}}{1-c}.$$

From which, we see that $\{x^t\}$ is a Cauchy sequence and hence convergent. Denoting its limit by $\hat{x}$ and letting $m_2 \to \infty$ in the above relation, we see further that

$$\|x^{m_1} - \hat{x}\| \leq \frac{Mc^{m_1}}{1-c}.$$

From this relation, we have that $\{x^t\}$ is R-linearly convergent to its limit, which is a stationary point of F according to Lemma 3.2.4. This proves (i).

Next, we prove (ii). Notice that for any $t \geq 1$, we have from the definition of $H_{t,\alpha}$ that

$$\begin{aligned} |F(x^t) - \zeta| &= |H_{t,\alpha} - \zeta - \alpha\|x^t - x^{t-1}\|^2| \\ &\leq H_{t,\alpha} - \zeta + \alpha\|x^t - x^{t-1}\|^2 \\ &\leq H_{t,\alpha} - \zeta + \frac{\alpha}{\delta}(H_{t-1,\alpha} - \zeta), \end{aligned}$$

where the first inequality follows from the triangle inequality and the fact that the sequence $\{H_{t,\alpha}\}$ is nonincreasing and convergent to ζ according to Lemmas 3.2.1 and 3.2.3, and the second inequality follows from (3.36). This together with the Q-linear convergence of $\{H_{t,\alpha}\}$ and Lemma 2.2.1 implies the R-linear convergence of $\{F(x^t)\}$. This completes the proof. $\square$

3.2.3 FISTA with restart: a special case of $\mathrm{PG_e}$

In this subsection, we introduce FISTA with restart schemes and compare it with the our algorithm $\mathrm{PG_e}$.

Recently, O'Donoghue and Candès [26] proposed adaptive restart schemes for FISTA, which attract many people's attention. They mainly discussed FISTA with fixed and adaptive restarts. Moreover, they proved the global linear convergence of the objective value, when using FISTA with restart schemes for solving (1.1) with f being strongly convex and $g = 0$. Similar restart techniques can also be adopted in many other algorithms and applications. For example, in the popular software, TFOCS [11], the authors also applied the restart technique. However, for the convex nonsmooth problems such as the LASSO, they did not establish any linear convergence results. For the LASSO, they stressed that "after a certain number of iterations adaptive restarting can provide linear convergence"; see [26, Page 728]. Next, we will show that FISTA with the aforementioned restart techniques is a special case of our algorithm $\mathrm{PG_e}$. Hence, by our theories in the previous subsection, when applying FISTA with both of their restart schemes to solving LASSO, we obtain that both the sequence $\{x^t\}$ and $\{F(x^t)\}$ generated by FISTA with both the restart schemes are R-linearly convergent,

We first present the framework of FISTA [3, 45] for solving problem (1.1) with an additionally convex f.

```
FISTA Input: $x^0 \in \text{dom}\ g$, $\theta_{-1} = \theta_0 = 1$. Set $x^{-1} = x^0$.
    for $t = 0, 1, 2 \cdots$ do
        $\beta_t = \frac{\theta_{t-1} - 1}{\theta_t}$,
        $y^t = x^t + \beta_t(x^t - x^{t-1})$,
        $x^{t+1} = \text{Prox}_{\frac{1}{L}g}\left(y^t - \frac{1}{L}\nabla f(y^t)\right)$,
        $\theta_{t+1} = \frac{1 + \sqrt{1 + 4\theta_t^2}}{2}$.
    end for
```

In Subsection 1.1.1, we have introduced that FISTA is one of the variants of Nesterov's accelerated proximal gradient algorithms and the extrapolation parameter in FISTA takes a specific choice of $\{\beta_t\}$. From the description of β_t in FISTA above, one can easily deduce that $0 \leq \beta_t < 1$ for all t.[1] While since f is convex, then we can choose $l = 0$ and hence the threshold $\sqrt{\frac{L}{L+l}} = 1$ in PG_e. According to the above discussion, we see that FISTA is a special case of PG_e.

FISTA with restart schemes (see, for example, [11, 26]) is a new algorithm based on FISTA, which presents faster convergence property. In our thesis, we consider the same restart schemes used in [26], which are the fixed restart scheme and the adaptive restart scheme. For the fixed restart scheme, we choose a fixed positive integer T, then we reset $\theta_{t-1} = \theta_t = 1$ every T iterations. While for the adaptive restart scheme, here we use the gradient scheme,[2] we reset $\theta_t = \theta_{t+1} = 1$ whenever $\langle y^t - x^{t+1}, x^{t+1} - x^t \rangle > 0$; see [26, Eq. 13]. Obviously, if the fixed restart scheme is

[1] See the proof in Corollary 3.2.11 in the next subsection.

[2] There are two adaptive restart schemes considering in [26, Section 3.2]. One is the gradient scheme, the other is the function value scheme. It was shown in [26, Section 3.2] that the above mentioned two restart schemes perform similarly empirically and that the gradient scheme is advantageous over the function value scheme. Thus, in this thesis, we mainly consider the gradient scheme.

invoked in the FISTA with restart schemes, we will have $\bar{\beta} < 1$. Thus, we have the following immediate corollary of Theorem 3.2.7.

Corollary 3.2.8. *Suppose that f in* (1.1) *is convex and Assumption 1 holds. Suppose further that $\{x^t\}$ is a sequence generated by FISTA with the fixed restart scheme or both the fixed and adaptive restart schemes. Then*

(i) *$\{x^t\}$ is R-linearly convergent to a globally optimal solution of* (1.1)*.*

(ii) *$\{F(x^t)\}$ is R-linearly convergent to the globally optimal value of* (1.1)*.*

In view of the discussions and some examples given following Assumption 1, we obtain that the objective in the LASSO satisfies Assumption 1. Hence, when applying FISTA with the fixed restart scheme or both the fixed restart scheme and the adaptive restart scheme to solving the LASSO, we see from Corollary 3.2.8 that both the sequences $\{x^t\}$ and $\{F(x^t)\}$ generated by the FISTA with restarts are R-linearly convergent.

Before ending this subsection, we will give a remark to stress that there are two main differences between our Corollary 3.2.8 and the convergence results in [26].

Remark 1. (i) *The authors in [26] established the* global *linear convergence of function values for* (1.1) *with a strongly convex f and a void g, while we prove the* local *linear convergence of both $\{x^t\}$ and $\{F(x^t)\}$ for* (1.1) *with f being convex.*

(ii) *Their global linear convergence is only guaranteed when T is chosen sufficiently large; see [26, Eq. 6]. On the other hand, we do not have any restrictions on the number T, the width of the restart interval.*

3.2.4 A further study of $\mathrm{PG_e}$ when f is convex

In this subsection, we further study $\mathrm{PG_e}$ when f is convex and F is level-bounded. Under these assumptions, we can take $l = 0$ in our algorithm and hence we have $\sqrt{\frac{L}{L+l}} = 1$. Our Theorem 3.2.7 can then be applied to establishing R-linear convergence of both $\{x^t\}$ and $\{F(x^t)\}$ generated by $\mathrm{PG_e}$ when Assumption 1 holds and $\bar{\beta} < 1$. However, in many widely used accelerated proximal gradient algorithms, such as the FISTA, we have $\bar{\beta} = 1$. For these kinds of algorithms, nothing is known concerning the convergence behavior of $\{x^t\}$ except when $\beta_t = \frac{t-1}{t+\alpha-1}$ with $\alpha > 3$. In this subsection, we will show that $\|x^{t+1} - x^t\| \to 0$ under a very general choice of $\{\beta_t\}$; we will also demonstrate that this condition is satisfied by the choice of $\{\beta_t\}$ used in the FISTA. We would like to point out that using $\|x^{t+1} - x^t\| \to 0$, one can obtain as an immediate corollary that any accumulation point of $\{x^t\}$ is a global minimizer of (1.1). Thus, establishing $\|x^{t+1} - x^t\| \to 0$ is an important step towards understanding the convergence behavior of $\{x^t\}$. Moreover, this fact can be used for designing termination criterion.

We now present our analysis. We first give a simple auxiliary lemma.

Lemma 3.2.9. *Let $\{a_t\}$ be a nondecreasing nonnegative sequence with $\lim\limits_{t\to\infty} a_t = 1$. Then*

$$\sum_{t=1}^{\infty} |a_{t+1}(1 - a_{t+2}) - (1 + a_t)(1 - a_{t+1}) + 1 - a_t| < \infty. \tag{3.38}$$

Proof. Write $c_t = 1 - a_t$ for notational simplicity. Then

$$\begin{aligned}
&|a_{t+1}(1-a_{t+2}) - (1+a_t)(1-a_{t+1}) + 1 - a_t| \\
&= |a_{t+1}c_{t+2} - (1+a_t)c_{t+1} + c_t| \\
&= |a_{t+1}c_{t+2} - c_{t+1} - a_tc_{t+1} + c_t| \\
&= |a_{t+1}c_{t+2} - c_{t+1} - a_tc_{t+1} + c_t - a_{t+1}c_{t+1} + a_{t+1}c_{t+1}| \\
&\le |c_t - c_{t+1}| + a_{t+1}|c_{t+2} - c_{t+1}| + c_{t+1}|a_{t+1} - a_t| \\
&= |a_t - a_{t+1}| + a_{t+1}|a_{t+2} - a_{t+1}| + (1 - a_{t+1})|a_{t+1} - a_t|.
\end{aligned} \tag{3.39}$$

Next, we observe that

$$\begin{aligned}
&\sum_{t=1}^{N}(|a_t - a_{t+1}| + a_{t+1}|a_{t+2} - a_{t+1}| + (1 - a_{t+1})|a_{t+1} - a_t|) \\
&\le \sum_{t=1}^{N}(a_{t+1} - a_t) + \sum_{t=1}^{N}(a_{t+2} - a_{t+1}) + \sum_{t=1}^{N}(a_{t+1} - a_t) \\
&= 2a_{N+1} + a_{N+2} - 2a_1 - a_2,
\end{aligned} \tag{3.40}$$

where the inequality follows from the nondecreasing property of $\{a_t\}$ and the fact that $\lim_{t\to\infty} a_t = 1$. Letting $N \to \infty$ in (3.40) and invoking $\lim_{t\to\infty} a_t = 1$ and (3.39), we obtain (3.38) as desired. □

Proposition 3.2.10. *Suppose that f in* (1.1) *is in addition convex and F is a level-bounded function. Suppose further that the sequence $\{\beta_t\}$ is nondecreasing with $\sum_{t=1}^{\infty}(1-\beta_t) = \infty$ and $\bar{\beta} = 1$. Let $\{x^t\}$ be a sequence generated by PG*$_{\rm e}$*. Then $\lim_{k\to\infty} \|x^t - x^{t-1}\| = 0$.*

Proof. Without loss of generality, let the optimal value of (1.1) be 0. Hence, for all $z \in \mathcal{X}$, we have $F(z) = f(z) + g(z) = 0$. We will subsequently show that $\lim_{t\to\infty} H_{t,\frac{L}{2}} = 0$. Granting this, we conclude from the definition of $H_{t,\frac{L}{2}}$ that

$$\limsup_{t\to\infty} \frac{L}{2}\|x^t - x^{t-1}\|^2 \le \limsup_{t\to\infty}\left(F(x^t) + \frac{L}{2}\|x^t - x^{t-1}\|^2\right) = \lim_{t\to\infty} H_{t,\frac{L}{2}} = 0,$$

from which the desired conclusion follows immediately.

It now remains to show $\lim\limits_{t\to\infty} H_{t,\frac{L}{2}} = 0$. To this end, fix a $z \in \mathcal{X}$ and define an auxiliary sequence $h_t = \frac{1}{2}\|x^t - z\|^2$. Then it is not hard to show that

$$h_{t+1} - h_t = \langle x^{t+1} - x^t, x^{t+1} - z\rangle - \frac{1}{2}\|x^{t+1} - x^t\|^2. \tag{3.41}$$

Next, recall from (3.4) and the fact that $\ell = 0$ that

$$f(x^{t+1}) + g(x^{t+1}) \le f(z) + g(z) + \frac{L}{2}\|z - y^t\|^2 - \frac{L}{2}\|x^{t+1} - z\|^2.$$

This together with the definition of $H_{t,\frac{L}{2}}$ and the assumption that $F(z) = 0$ gives

$$0 \le H_{t+1,\frac{L}{2}} \le \frac{L}{2}\|z - y^t\|^2 - \frac{L}{2}\|x^{t+1} - z\|^2 + \frac{L}{2}\|x^{t+1} - x^t\|^2. \tag{3.42}$$

For the first term on the right hand side of (3.42), using the definition of y^t in PG$_\mathrm{e}$, we see that

$$\begin{aligned}\|z - y^t\|^2 &= \|x^t + \beta_t(x^t - x^{t-1}) - z\|^2\\ &= \|x^t - z\|^2 + 2\beta_t\langle x^t - z, x^t - x^{t-1}\rangle + \beta_t^2\|x^t - x^{t-1}\|^2\\ &= \|x^t - z\|^2 + 2\beta_t(h_t - h_{t-1}) + (\beta_t + \beta_t^2)\|x^t - x^{t-1}\|^2,\end{aligned} \tag{3.43}$$

where the last equality follows from (3.41). Combining (3.42) with (3.43), we obtain further that

$$\begin{aligned}H_{t+1,\frac{L}{2}} \le\ & \frac{L}{2}(\|x^t - z\|^2 + 2\beta_t(h_t - h_{t-1}) + (\beta_t + \beta_t^2)\|x^t - x^{t-1}\|^2)\\ &- \frac{L}{2}\|x^{t+1} - z\|^2 + \frac{L}{2}\|x^{t+1} - x^t\|^2\\ =\ & Lh_t + L\beta_t(h_t - h_{t-1}) + \frac{L}{2}(\beta_t^2 + \beta_t)\|x^t - x^{t-1}\|^2 - Lh_{t+1} + \frac{L}{2}\|x^{t+1} - x^t\|^2\\ =\ & \frac{L}{2}\|x^{t+1} - x^t\|^2 + \frac{L}{2}(\beta_t^2 + \beta_t)\|x^t - x^{t-1}\|^2 - L\xi_t\\ \le\ & \frac{L}{2}\|x^{t+1} - x^t\|^2 + L\|x^t - x^{t-1}\|^2 - L\xi_t,\end{aligned} \tag{3.44}$$

where $\xi_t = h_{t+1} - h_t - \beta_t(h_t - h_{t-1})$, and the last inequality is obtained by using the fact that $\{\beta_t\}$ is nondecreasing with $\bar{\beta} = 1$. Multiplying $1 - \beta_{t+1}$ to both sides of (3.44), we obtain

$$\begin{aligned}
&(1-\beta_{t+1})H_{t+1,\frac{L}{2}} \\
&\leq \frac{L}{2}(1-\beta_{t+1})\|x^{t+1}-x^t\|^2 + L(1-\beta_{t+1})\|x^t - x^{t-1}\|^2 - L(1-\beta_{t+1})\xi_t \qquad (3.45)\\
&\leq \frac{L}{2}(1-\beta_{t+1})\|x^{t+1}-x^t\|^2 + L(1-\beta_t)\|x^t - x^{t-1}\|^2 - L(1-\beta_{t+1})\xi_t,
\end{aligned}$$

where the last inequality is a consequence of the fact that $\{\beta_t\}$ is nondecreasing.

We next show that the sequence $\{\sum_{t=1}^N (1-\beta_{t+1})\xi_t\}$ is bounded. For notational simplicity, we define $\delta_{t+1} = 1 - \beta_{t+1}$. Then we have

$$\begin{aligned}
\sum_{t=1}^N \delta_{t+1}\xi_t &= \sum_{t=1}^N \delta_{t+1}(h_{t+1} - h_t - \beta_t(h_t - h_{t-1}))\\
&= \sum_{t=1}^N \delta_{t+1}h_{t+1} - \sum_{t=1}^N \delta_{t+1}h_t - \sum_{t=1}^N \delta_{t+1}\beta_t h_t + \sum_{t=1}^N \delta_{t+1}\beta_t h_{t-1}\\
&= \sum_{t=2}^{N+1} \delta_t h_t - \sum_{t=1}^N \delta_{t+1}h_t - \sum_{t=1}^N \delta_{t+1}\beta_t h_t + \sum_{t=0}^{N-1} \delta_{t+2}\beta_{t+1} h_t\\
&= \sum_{t=1}^{N+1} \delta_t h_t - \sum_{t=1}^N \delta_{t+1}h_t - \sum_{t=1}^N \delta_{t+1}\beta_t h_t + \sum_{t=0}^{N} \delta_{t+2}\beta_{t+1} h_t \qquad (3.46)\\
&\quad - \delta_1 h_1 - \delta_{N+2}\beta_{N+1}h_N\\
&= \sum_{t=1}^N [\beta_{t+1}\delta_{t+2} - (1+\beta_t)\delta_{t+1} + \delta_t]h_t\\
&\quad + \delta_{N+1}h_{N+1} + \delta_2\beta_1 h_0 - \delta_1 h_1 - \delta_{N+2}\beta_{N+1}h_N\\
&= \sum_{t=1}^N r_t h_t + \delta_{N+1}h_{N+1} + \delta_2\beta_1 h_0 - \delta_1 h_1 - \delta_{N+2}\beta_{N+1}h_N,
\end{aligned}$$

where $r_t = \beta_{t+1}\delta_{t+2} - (1+\beta_t)\delta_{t+1} + \delta_t$. Since $\{x^t\}$ is bounded from Corollary 3.2.2, we see immediately that $\{h_t\}$ is bounded from its definition. Using this fact and applying Lemma 3.2.9 with $a_t = \beta_t$, we obtain further that

$$\limsup_{N\to\infty} \left| \sum_{t=1}^{N} r_t h_t \right| \leq \sum_{t=1}^{\infty} |\beta_{t+1}\delta_{t+2} - (1+\beta_t)\delta_{t+1} + \delta_t| \cdot |h_t| < \infty,$$

where we recall that $\delta_t = 1 - \beta_t$ for all t. From this, (3.45), (3.46), the boundedness of $\{h_t\}$ and $\{\beta_t\}$, and Lemma 3.2.3, we have

$$\begin{aligned} &\sum_{t=1}^{\infty} (1-\beta_{t+1}) H_{t+1,\frac{L}{2}} \\ &\leq \frac{L}{2} \sum_{t=1}^{\infty} \delta_{t+1} \|x^{t+1} - x^t\|^2 + L \sum_{t=1}^{\infty} \delta_t \|x^t - x^{t-1}\|^2 + L \cdot \limsup_{N\to\infty} \left| \sum_{t=1}^{N} \delta_{t+1} \xi_t \right| \\ &\leq \frac{L}{2} \sum_{t=1}^{\infty} (1-\beta_{t+1}^2) \|x^{t+1} - x^t\|^2 + L \sum_{t=1}^{\infty} (1-\beta_t^2) \|x^t - x^{t-1}\|^2 \\ &\quad + L \cdot \limsup_{N\to\infty} \left| \sum_{t=1}^{N} (1-\beta_{t+1}) \xi_t \right| < \infty, \end{aligned} \tag{3.47}$$

where the second inequality holds because $0 \leq \beta_t \leq 1$ for all t.

We are now ready to show that $\lim_{t\to\infty} H_{t,\frac{L}{2}} = 0$. From Lemma 3.2.1 and Lemma 3.2.3, we know that $\{H_{t,\frac{L}{2}}\}$ is convergent and nonincreasing. Suppose to the contrary that $\lim_{t\to\infty} H_{t,\frac{L}{2}} = \inf_t H_{t,\frac{L}{2}} = H_\infty > 0$ for some H_∞. Then from this and (3.47), we have

$$H_\infty \sum_{t=1}^{\infty} (1-\beta_t) \leq \sum_{t=1}^{\infty} (1-\beta_t) H_{t,\frac{L}{2}} < \infty,$$

which is a contradiction to our assumption that $\sum_{t=1}^{\infty}(1-\beta_t) = \infty$. Thus, it must hold true that $\lim_{t\to\infty} H_{t,\frac{L}{2}} = 0$. This completes the proof. □

In the following, we will show that Proposition 3.2.10 can be used to analyze the convergence behavior of the sequence generated by the FISTA.

Corollary 3.2.11. *Suppose that f in* (1.1) *is convex and F is level-bounded. Then the sequence $\{x^t\}$ generated by the FISTA satisfies* $\lim_{t\to\infty} \|x^{t+1} - x^t\| = 0$.

Proof. According to Proposition 3.2.10, we only need to show that the sequence $\{\beta_t\}$ in the FISTA is nonnegative and nondecreasing with $\sup_t \beta_t = 1$ and $\sum_{t=1}^{\infty}(1-\beta_t) = \infty$.

First, using the definition that $\beta_t = \frac{\theta_{t-1}-1}{\theta_t}$ and $\theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2}$ with $\theta_{-1} = \theta_0 = 1$ in the FISTA, we have for any $t \geq 0$ that

$$\begin{aligned}\beta_{t+1} - \beta_t = \frac{\theta_t - 1}{\theta_{t+1}} - \frac{\theta_{t-1} - 1}{\theta_t} &= \frac{\theta_t^2 - \theta_t - \theta_{t-1}\theta_{t+1} + \theta_{t+1}}{\theta_t\theta_{t+1}} \\ &= \frac{\theta_t^2 - \theta_{t-1}\theta_{t+1} + \theta_{t+1} - \theta_t}{\theta_t\theta_{t+1}} \\ &= \frac{\theta_t\theta_{t-1}(\frac{\theta_t}{\theta_{t-1}} - \frac{\theta_{t+1}}{\theta_t}) + \theta_{t+1} - \theta_t}{\theta_t\theta_{t+1}},\end{aligned} \tag{3.48}$$

and

$$\theta_{t+1} - \theta_t = \frac{1 + \sqrt{1 + 4\theta_t^2}}{2} - \theta_t \geq 0. \tag{3.49}$$

Moreover, for any $t \geq 0$,

$$\frac{\theta_{t+1}}{\theta_t} = \frac{1 + \sqrt{1 + 4\theta_t^2}}{2\theta_t} = \frac{1}{2\theta_t} + \sqrt{1 + \frac{1}{4\theta_t^2}}. \tag{3.50}$$

Since $\{\theta_t\}$ is positive and nondecreasing from the definition of θ_t and (3.49), we see from (3.50) that $\left\{\frac{\theta_{t+1}}{\theta_t}\right\}$ is nonincreasing. Combining these facts with (3.48) and the fact that $\beta_0 = \beta_1 = 0$, we obtain further that $\{\beta_t\}$ is nondecreasing and nonnegative.

Next, using the fact $\sqrt{a^2 + b^2} \leq a + b$ for $a \geq 0$, $b \geq 0$, we see that for $t \geq 1$,

$$\theta_t = \frac{1 + \sqrt{1 + 4\theta_{t-1}^2}}{2} \leq \frac{1 + 1 + 2\theta_{t-1}}{2} = 1 + \theta_{t-1}. \tag{3.51}$$

Since $\theta_0 = 1$, by induction, we obtain further that $\theta_t \le t+1$ for any $t \ge 1$. Hence, from the nondecreasing property of $\{\theta_t\}$, we have for any $t \ge 1$ that

$$1-\beta_t = 1-\frac{\theta_{t-1}-1}{\theta_t} = \frac{\theta_t-\theta_{t-1}+1}{\theta_t} \ge \frac{1}{\theta_t} \ge \frac{1}{t+1},$$

which implies that $\sum_{t=1}^{\infty}(1-\beta_t) = \infty$. Finally, by induction, we obtain that $\theta_t \ge \frac{t+2}{2}$ for $t \ge 0$. This together with (3.51) implies that

$$1-\beta_t = \frac{\theta_t-\theta_{t-1}+1}{\theta_t} \le \frac{2}{\theta_t} \le \frac{4}{t+2}$$

for any $t \ge 1$. From this last relation, we see that $\beta_t \ge \frac{t-2}{t+2}$. This together with the nondecreasing property of $\{\beta_t\}$ and the fact that $0 \le \beta_t \le 1$ for all t implies that $\sup_t \beta_t = 1$. □

3.3 Numerical experiments

In this section, some numerical experiments are performed to show that both the sequence $\{x^t\}$ generated by $\mathrm{PG_e}$ and the corresponding objective values sequence $\{F(x^t)\}$ are R-linearly convergent.

Three different classes of problems are considered in this section, which are the ℓ_1 regularized logistic regression problem, the LASSO, and the nonconvex quadratic problem over a simplex. Among them, the first two examples are convex optimization problems, while the third one is possibly nonconvex optimization problem. We use different choices of the extrapolation coefficients β_t in $\mathrm{PG_e}$ for these problems. Concretely, we apply $\mathrm{PG_e}$ with β_t chosen as in FISTA with both the fixed and the adaptive restart schemes, β_t chosen as in FISTA, and $\beta_t \equiv 0$ (proximal gradient algorithm) to solving the first two convex optimization problems. On the other hand, we apply $\mathrm{PG_e}$ with $\beta_t \equiv 0.98\sqrt{\frac{L}{L+l}}$ and $\beta_t \equiv 0$ (proximal gradient algorithm) for the nonconvex optimization problems. We also use FISTA to solve the nonconvex problems as a heuristic.

All these numerical experiments are performed by Matlab 2014b on a 64-bit PC with a 3.60GHz Inter Core i7-4790 processor and 32GB of RAM.

3.3.1 ℓ_1 regularized logistic regression

We first consider the ℓ_1 regularized logistic regression problem:

$$v_{\log} := \min_{\tilde{x}\in\mathbb{R}^n, x_0\in\mathbb{R}} \sum_{i=1}^{m} \log(1+\exp(-b_i(a_i^T\tilde{x}+x_0))) + \lambda\|\tilde{x}\|_1, \tag{3.52}$$

where a_i is a vecor in $\mathbb{R}^n$ space, b_i is a integer in $\{-1,1\}$, $i=1,2,\cdots,m$, with b_i not all the same, $m<n$ and $\lambda>0$ is the regularization parameter. It is easy to see that (3.52) is in the form of (1.1) with

$$f(x)=\sum_{i=1}^{m}\log(1+\exp(-b_i(Dx)_i)), \qquad g(x)=\lambda\|\tilde{x}\|_1, \tag{3.53}$$

where $x := (\tilde{x}, x_0)\in\mathbb{R}^{n+1}$, and D is a matrix with the ith row given by $(a_i^T \ \ 1)$. Moreover, one can show that ∇f is Lipschitz continuous with modulus $0.25\lambda_{\max}(D^\top D)$. Hence, in our testing algorithms below, we take $L=0.25\lambda_{\max}(D^TD)$. Since f in (3.53) is convex, we take $l=0$ in our algorithms.

Next we will show that $v_{\log} > -\infty$ and the solution set $\mathcal{X}$ of (3.52) is nonempty. Recalling that the dual problem of (3.52) is given by

$$\begin{array}{rl}\displaystyle\max_{u\in\mathbb{R}^m} & d_{\log}(u) := -\sum_{i=1}^m[-b_iu_i\log(-b_iu_i)+(1+b_iu_i)\log(1+b_iu_i)]\\ \text{s.t.} & \|A^Tu\|_\infty\le\lambda, \quad e^Tu=0,\end{array} \tag{3.54}$$

where A is the matrix whose ith row is given by a_i^T. From [15, Theorem 3.3.5], we conclude that the optimal values of (3.52) and (3.54) are the same, and that an optimal solution of (3.54) exists. In addition, due to the facts that $\lambda>0$ and b_i are not all the same, the generalized Slater condition is satisfied for (3.54), i.e., there exists $\tilde{u}$ satisfying $\|A^T\tilde{u}\|_\infty < \lambda$, $e^T\tilde{u} = 0$ and $-1 < b_i\tilde{u}_i < 0$ for $i = 1, \ldots, m$. Hence, by [52, Corollary 28.2.2], an optimal solution of (3.52) exists. Consequently, we see that $v_{\text{log}} > -\infty$ and the solution set $\mathcal{X}$ of (3.52) is nonempty.

Thus, we can apply PG_e to solving problem (3.52). Moreover, in view of the discussion following Assumption 1, Assumption 1 is satisfied for (3.53). Hence, from Corollary 3.2.8, we can expect the R-linear convergence of the sequences $\{x^t\}$ and $\{F(x^t)\}$ generated by FISTA with both fixed restart and adaptive restart schemes.

Now we perform numerical experiments to study PG_e. We choose different extrapolation parameters $\{\beta_t\}$ in PG_e: β_t chosen as in FISTA with both the fixed and the adaptive restart schemes, where we perform a fixed restart every 500 iterations (FISTA-R500), β_t chosen as in FISTA and $\beta_t \equiv 0$ as in the proximal gradient algorithm (PG). The regularization parameter λ in (3.52) is always taken as $\lambda = 5$. We initialize all the above algorithms at the origin. For the termination, in view of [52, Theorem 31.3], we see that for any $\bar{x} \in \mathcal{X}$, $\nabla p(D\bar{x})$ is an optimal solution of (3.54). Specifically, we define

$$u^t = \min\left\{1, \frac{\lambda}{\|A^T\nabla p(Dx^t)\|_\infty}\right\}\nabla p(Dx^t),$$

and terminate the algorithms if the duality gap and the dual feasibility violation are small, i.e.,

$$\max\left\{\frac{|f(x^t) + g(x^t) - d_{\text{log}}(u^t)|}{\max\{f(x^t) + g(x^t), 1\}}, \frac{50|e^Tu^t|}{\max\{\|u^t\|, 1\}}\right\} \leq 10^{-6}.$$

The algorithms are also terminated when the number of iterations reaches 5000.

We consider random instances for our experiments. For each $(m, n, s) = (300, 3000, 30)$, $(500, 5000, 50)$ and $(800, 8000, 80)$, we generate an $m \times n$ matrix A with i.i.d. standard Gaussian entries. We then choose a support set T of size s uniformly at random, and generate an s-sparse vector $\hat{x}$ supported on T with i.i.d. standard Gaussian entries. The vector b is then generated as $b = \text{sign}(A\hat{x} + ce)$, where c is chosen uniformly at random from $[0, 1]$.

The computational results are presented in the following figures, i.e., Figures 3.1, 3.2 and 3.3. We plot $\|x^t - x^*\|$ against the number of iterations t in part (a) of each figure, where x^* denotes the approximate solution obtained at termination of the respective algorithm. While the part (b) of each figure plots $|F(x^t) - F_{\min}|$ versus the number of iterations t, where $F_{\min}$ denotes the minimum of three objective values obtained from the three respective algorithms. According to these figures, we see that both the sequence $\{x^t\}$ and $\{F(x^t)\}$ generated by FISTA with both fixed and adaptive restart schemes are R-linearly convergent, which are consistent with our theoretical results. Moreover, FISTA with both fixed and adaptive restart schemes always performs better than FISTA and the proximal gradient algorithm from the figures.

Figure 3.1: $l_1 - logistic: \ n = 3000, m = 300, s = 30$

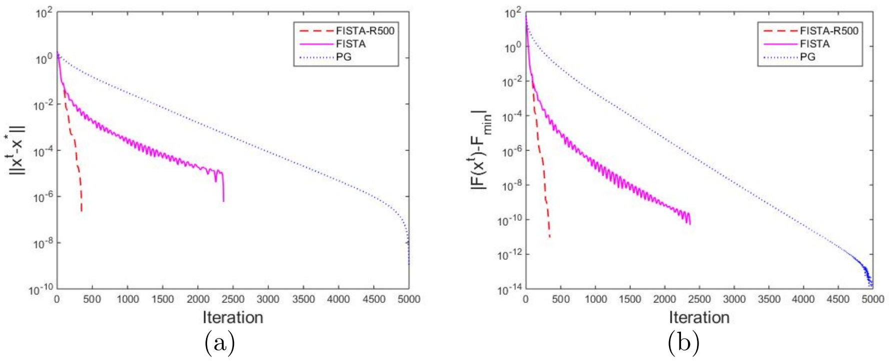
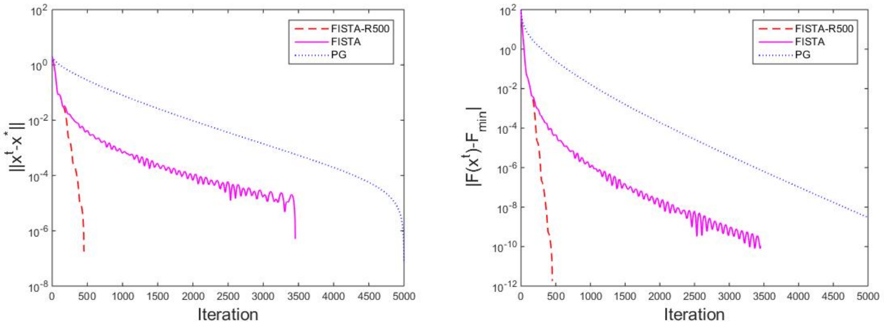
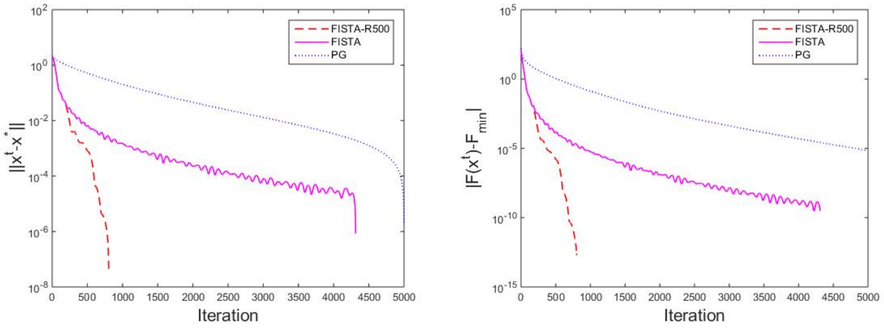


Figure 3.2: $l_1 - logistic: \ n = 5000, m = 500, s = 50$



(a) (b)

Figure 3.3: $l_1 - logistic: \ n = 8000, m = 800, s = 80$



(a) (b)

3.3.2 LASSO

In this subsection, we study the second convex optimization problem, i.e., the LASSO problem, which is in the following form:

$$v_{\rm ls} := \min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1, \tag{3.55}$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ are given. It can be seen from the objective function in (3.55) that (3.55) is in the form of (1.1) with

$$f(x) = \frac{1}{2}\|Ax - b\|^2, \qquad g(x) = \lambda\|x\|_1. \tag{3.56}$$

From the above relation, one can easily show that ∇f is Lipschitz continuous and $f + g$ has compact lower level sets. Thus, $\mathrm{PG_e}$ can be applied to solving (3.55).

Moreover, Assumption 1 is satisfied for (3.56) according to the discussion following Assumption 1. Hence, from Corollary 3.2.8, we can expect the R-linear convergence of both the sequences $\{x^t\}$ and $\{F(x^t)\}$ generated by FISTA with fixed and adaptive restart schemes. Finally, it is routine to show that a Lipschitz continuity modulus of ∇f can be chosen as $\lambda_{\max}(A^TA)$. From this result, we take $L = \lambda_{\max}(A^TA)$ in the algorithms below. Since f is convex, we set $l = 0$ in the algorithms.

Note that $f(x)$ in (3.56) can be formulated as $f(x) = h(Ax) = \frac{1}{2}\|Ax - b\|^2$, where $h(v) = \frac{1}{2}\|v - b\|^2$. Then the conjugate function of h can be easily computed as $h^*(u) := \sup_{v\in\mathbb{R}^m}\{u^Tv - h(v)\} = \frac{1}{2}\|u\|^2 + b^Tu$. As a consequence, the dual problem of (3.55) is given as follows:

$$\begin{array}{rl}\max\limits_{u\in\mathbb{R}^m} & d_{\text{ls}}(u) := -\frac{1}{2}\|u\|^2 - b^Tu \\ \text{s.t.} & \|A^Tu\|_\infty \le \lambda.\end{array} \tag{3.57}$$

From [15, Theorem 3.3.5], we can show that the optimal values of (3.55) and (3.57) are the same, and moreover, an optimal solution of (3.57) exists. We will use the dual problem to develop the termination criterion for our algorithms below.

Now we perform numerical experiments to study PG$_{\text{e}}$. We choose the same extrapolation parameters as in the previous subsection. The regularization parameter λ is taken as $\lambda = 5$. We initialize all the above algorithms at the origin and we use the duality gap to terminate the algorithms. As the previous subsection, for any optimal solution $\bar{x}$ of (3.55), we obtain $\nabla h(A\bar{x})$ is an optimal solution of (3.57) from [52, Theorem 31.3]. Then we define

$$u^t = \min\left\{1, \frac{\lambda}{\|A^T\nabla h(Ax^t)\|_\infty}\right\}\nabla h(Ax^t),$$

and terminate the algorithms if the duality gap is small, i.e.,

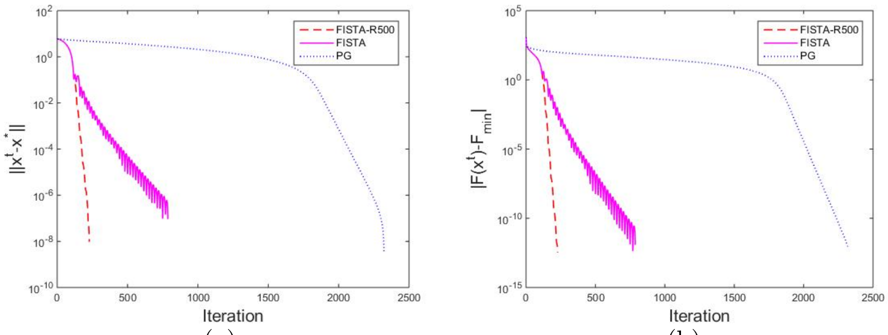
$$\frac{|f(x^t) + g(x^t) - d_{\text{ls}}(u^t)|}{\max\{f(x^t) + g(x^t), 1\}} \le 10^{-6}.$$

We also terminate them when the number of iterations reaches 5000.

The problems used in our experiments are generated as follows. For each $(m, n, s) = (300, 3000, 30)$, $(500, 5000, 50)$ and $(800, 8000, 80)$, we generate an $m \times n$ matrix A with i.i.d. standard Gaussian entries. We then choose a support set T of size s uniformly at random, and generate an s-sparse vector $\hat{x}$ supported on T with i.i.d. standard Gaussian entries. The vector b is then generated as $b = A\hat{x} + 0.01\tilde{e}$, where $\tilde{e}$ has standard i.i.d. Gaussian entries.

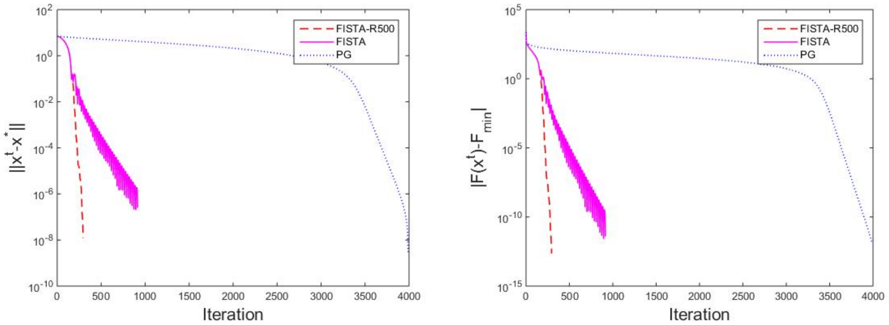
The computational results are presented in the following figures, i.e., Figures 3.4, 3.5 and 3.6. We plot $\|x^t - x^*\|$ against the number of iterations t in part (a) of each figure, where x^* denotes the approximate solution obtained at termination of the respective algorithm. Additionally, in the part (b) of each figure, we plot $|F(x^t) - F_{\min}|$ versus the number of iterations t, where $F_{\min}$ denotes the minimum of the three objective values obtained from the three respective algorithms. Similar as in the previous subsection, we see from these figures that both the sequence $\{x^t\}$ and $\{F(x^t)\}$ generated by FISTA with both fixed and adaptive restart schemes are R-linearly convergent, which are consistent with our theoretical results. Additionally, the algorithm with restart performs better than the others.

Figure 3.4: $l_1 - ls: \ n = 3000, m = 300, s = 30$



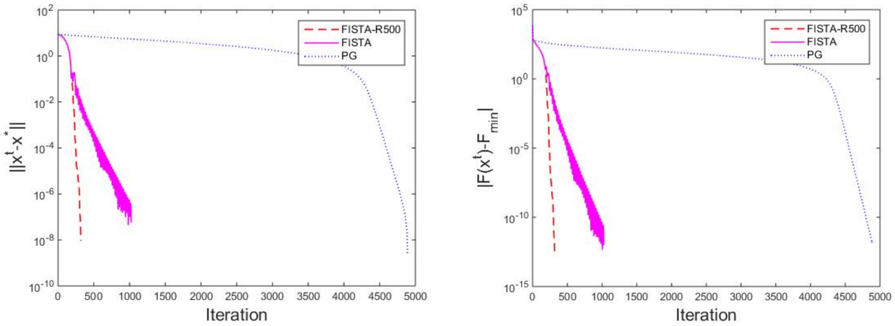
(a) (b)

Figure 3.5: $l_1 - ls: \ n = 5000, m = 500, s = 50$



(a) (b)

Figure 3.6: $l_1 - ls: \ n = 8000, m = 800, s = 80$



(a) (b)

3.3.3 Nonconvex quadratic programming with simplex constraints

In this subsection, we look at problems of the following form, which are possibly nonconvex:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \ & \frac{1}{2} x^T A x - b^T x \\ \text{s.t.} \ & e^T x = s, \quad x \geq 0, \end{aligned} \tag{3.58}$$

where $A \in \mathbb{R}^{n \times n}$ is a symmetric matrix that is not necessarily positive semidefinite, $b \in \mathbb{R}^n$ and s is a positive number. This is an example of nonconvex quadratic programming problems, which is an important class of problems in global optimization [23, 29, 32, 40]. From the objective and constrain condition in (3.58), one can easily reformulate (3.58) in the form of (1.1) with

$$f(x) = \frac{1}{2}x^T Ax - b^T x, \qquad g(x) = \delta_{\mathcal{S}}(x), \tag{3.59}$$

where $\mathcal{S} = \left\{x \in \mathbb{R}^n : \ e^T x = s, \ x \geq 0\right\}$. Moreover, it is clear that f has a Lipschitz continuous gradient and $f + g$ is level bounded. Hence, we can apply $\mathrm{PG_e}$ to solving (3.58). Additionally, Assumption 1 is satisfied for (3.59), according to the discussion following Assumption 1. As a consequence, from Theorem 3.2.7, we can expect the R-linear convergence of both the sequences $\{x^t\}$ and $\{F(x^t)\}$ generated by $\mathrm{PG_e}$ when $\bar{\beta} < \sqrt{\frac{L}{L+l}}$. Finally, we can decompose matrix A by $A = A_1 - A_2$, where A_1 and $-A_2$ are the projections of A onto the cone of positive semidefinite matrices and the cone of negative semidefinite matrices, respectively. Hence, we can rewrite f as $f = f_1 - f_2$, where $f_1(x) = \frac{1}{2}x^T A_1 x - b^T x$ and $f_2(x) = \frac{1}{2}x^T A_2 x$. From the above discussions, in our following numerical experiments, we take $L = \max\{\lambda_{\max}(A), |\lambda_{\min}(A)|\}$ and $l = |\lambda_{\min}(A)|$ so that L and l are the Lipschitz continuity moduli of ∇f_1 and ∇f_2, respectively, and $L \geq l$ by the definition of L.

Now some numerical experiments are performed to study $\mathrm{PG_e}$. We choose different extrapolation parameters $\{\beta_t\}$ in $\mathrm{PG_e}$: $\beta_t \equiv 0.98\sqrt{\frac{L}{L+l}}$ ($\mathrm{PG_e}$), β_t chosen as in FISTA and $\beta_t \equiv 0$ (PG). Here we want to point out that FISTA applied to the nonconvex problem (3.58) is not known to converge, unlike the other two algorithms which have convergence guarantee by our theory. We initialize all the above algorithms at the origin. Unlike the previous convex examples, we terminate these algorithms when the successive changes of the iterates are small, i.e.,

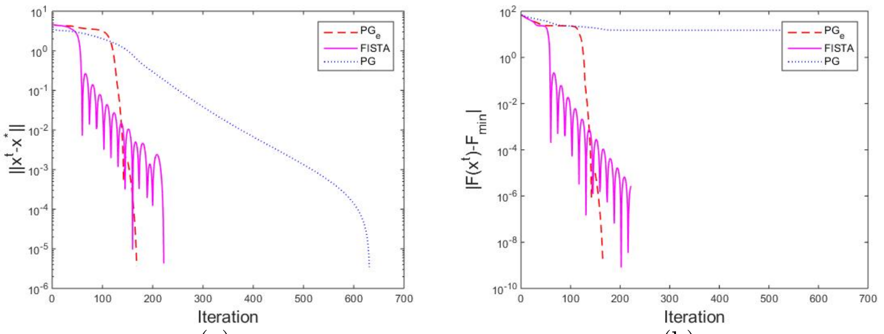
$$\frac{\|x^t - x^{t-1}\|}{\max\{\|x^t\|, 1\}} \leq 10^{-6}.$$

The algorithms are also terminated when the number of iterations reaches 5000.

Our test problem is generated as follows. We generate a 2000×2000 matrix D with i.i.d. standard Gaussian entries. We then generate a symmetric matrix $A = D + D^{\top}$. Finally, the vector b is generated with i.i.d. standard Gaussian entries, and s is generated as $\max\{1, 10t\}$, with t chosen uniformly at random from $[0, 1]$.

Our computational results are presented in Figure 3.7. Figure 3.7 (a) plots $\|x^t - x^*\|$ versus the number of iterations t, where x^* denotes the approximate solution obtained at termination of the respective algorithm. Additionally, Figure 3.7 (b) plots $|F(x^t) - F_{\min}|$ versus the number of iterations t, where $F_{\min}$ denotes the minimum of three objective values obtained from the three respective algorithms. From Figure 3.7 (a), we see that the sequence $\{x^t\}$ generated by $\mathrm{PG_e}$ with $\beta_t \equiv 0.98\sqrt{\frac{L}{L+l}}$ is R-linearly convergent, which is consistent with our theoretical results. However, Figure 3.7 (b) shows that not all the algorithms are approaching $F_{\min}$. This case is possible, which is because the iterates generated by the algorithm may get stuck at local minimizers.

Figure 3.7: *Nonconvex Quadratic Problem*



(a) (b)

We next perform another numerical experiment to test the quality (which means to see the function values at termination) of the approximate solution obtained from the above three algorithms. In this second experiment, we generate random instances as follows: we generate an $n \times n$ matrix D with i.i.d. standard Gaussian entries and symmetrize it to form $A = D + D^\top$; moreover, we generate a vector b with i.i.d. standard Gaussian entries, and an $s = \max\{1, 10t\}$, where t is chosen uniformly at random from $[0, 1]$.

In this test, we generate 50 random instances as the description above for each $n = 500$, 1000, 1500, 2000 and 2500. The following Table 3.1 presents the number of iterations averaged over the 50 instances for each n (iter), and the function value at termination (fval), also averaged over the 50 instances. From the reports in Table 3.1, we see that while $\mathrm{PG_e}$ with $\beta_t \equiv 0.98\sqrt{\frac{L}{L+l}}$ (i.e., $\mathrm{PG_e}$) is always the fastest algorithm, the function values obtained can be slightly compromised for some instances.

Table 3.1: Comparing $\mathrm{PG_e}$, FISTA and PG on random instances.

	$\mathrm{PG_e}$		FISTA		PG	
n	iter	fval	iter	fval	iter	fval
500	120	-56.02	175	-56.90	322	-57.96
1000	171	-69.77	274	-66.79	636	-66.93
1500	166	-66.29	270	-63.71	560	-65.29
2000	215	-80.72	271	-80.43	635	-81.21
2500	284	-81.70	359	-80.13	813	-83.81

3.4 Conclusions of this chapter

This chapter mainly studies the algorithm $\mathrm{PG_e}$ (3.1) for solving a class of nonconvex nonsmooth optimization problems. Assuming the error bound condition holds for the objective, we establish the R-linear convergence of both the sequence $\{x^t\}$ generated by the algorithm and the corresponding objective values sequence $\{F(x^t)\}$ if the extrapolation coefficients are below the threshold $\sqrt{\frac{L}{L+l}}$. If f in problem (1.1) is convex, the threshold reduced to 1. We further show that FISTA with fixed restart is a special case of $\mathrm{PG_e}$, hence our theory can be used to establish the R-linear convergence of the sequences $\{x^t\}$ and $\{F(x^t)\}$ generated by FISTA with fixed restart for solving (1.1) with a convex f, when the objective satisfies the error bound condition. In addition, for convex problems, we prove the successive changes of $\{x^t\}$ generated by $\mathrm{PG_e}$ go to 0 for a fairly choices of extrapolation coefficients whose threshold approach 1 under the assumption that the objective is level bounded. We show that the choices of extrapolation are general enough to cover the choices in FISTA. Finally, some numerical experiments are performed to verify our theoretical results.

Chapter 4

Proximal difference-of-convex algorithm with extrapolation

This chapter mainly deals with the difference-of-convex(DC) problems. These problems arise in many practical applications, and many nonconvex problems can be rewritten as the DC problems. In this chapter, we consider a proximal difference-of-convex algorithm with extrapolation(pDCA$_e$) for solving a kind of DC problems. The DC models we studied are given first, and then we present the algorithm pDCA$_e$. Next, we establish the global subsequential convergence of pDCA$_e$. Moreover, by assuming the Kurdyka-Łojasiewicz (KL) property holds for an auxiliary function, we establish the global convergence of $\{x^t\}$ generated by pDCA$_e$ and then analyze the convergence rate of $\{x^t\}$ under suitable conditions. Finally ,we perform numerical experiments on two DC problems. Our numerical experiments show that the pDCA$_e$ usually outperforms the proximal DCA with nonmonotone linesearch.

4.1 DC problem description and the pDCA$_e$

This section mainly gives a description of the DC problems we considered in this thesis, and then presents the proximal difference-of-convex algorithm with extrapolation (pDCA$_\mathrm{e}$).

The problems we considered in this chapter have the following form,

$$v := \min_{x\in\mathbb{R}^n} F(x) := \ f(x) + P(x), \tag{4.1}$$

where f is differentiable and convex, ∇f is Lipschitz continuous with a Lipschitz continuity modulus $L > 0$, and

$$P(x) = P_1(x) - P_2(x),$$

with P_1 being a proper closed convex function and P_2 being a *continuous* convex function. We assume in addition that F is level-bounded. This latter assumption implies that $v > -\infty$ and that the set of global minimizers of (4.1) is nonempty. Recently problem (4.1) becomes a hot topic in the optimization society and it can be found in many practical applications such as compressed sensing, where f can take the least squares loss function as the data fitting term, and P can be chosen as a nonsmooth regularizer for inducing some desirable structures in the solution. For more examples, we refer the readers to [1, 12, 28, 64, 65, 66] and the reference therein.

From the assumptions on f and P, problem (4.1) is a standard DC problem and thus we can apply the renowned DCA to solving it. However, as discussed in Subsection 1.1.2, directly using DCA may lead to difficult subproblems. Concretely, the subproblems of DCA for solving (4.1) take the following form:

$$x^{t+1} \in \operatorname*{Argmin}_{x\in\mathbb{R}^n} \left\{ f(x) + P_1(x) - \langle \xi^t, x\rangle \right\}, \tag{4.2}$$

where $\xi^t \in \partial P_2(x^t)$. It can be seen from (4.2) that these problems are convex, but they do not necessarily have closed form/simple solutions. In order to overcome this difficulty, recently, Gotoh, Takeda, and Tono [31] proposed a proximal DCA based on

DCA. When applied proximal DCA to solving (4.1), the subproblems are as follows,

$$
\begin{aligned}
x^{t+1} &= \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{\langle \nabla f(x^t)-\xi^t, x\rangle + \frac{L}{2}\|x-x^t\|^2 + P_1(x)\right\} \\
&= \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{\frac{L}{2}\left\|x-\left(x^t-\frac{1}{L}[\nabla f(x^t)-\xi^t]\right)\right\|^2 + P_1(x)\right\},
\end{aligned}
\tag{4.3}
$$

where $\xi^t \in \partial P_2(x^t)$. Compared with (4.2), by the definition of proximal operator given in Subsection 2.1.1, solving the subproblem (4.3) is equivalent to evaluating the proximal operator of $\frac{1}{L}P_1$, which is easy to compute for a large class of P_1; see, for example, [25, Tables 10.1 and 10.2].

Although the subproblems of proximal DCA are simple for many commonly used function P_1, using this algorithm for solving the practical problems maybe slow. The reason is that the proximal DCA is the same as the proximal gradient algorithm when $P_2 = 0$ and the proximal gradient algorithm can take a lot of iterations in practice [26, Section 5]. However, as we discussed in Subsection 1.1.1, performing various extrapolation techniques on the proximal gradient algorithm for convex optimization problems can successfully accelerate the original proximal gradient algorithm, see more details in [43, 44, 45, 46]. Stimulated by these facts, we attempt to incorporate extrapolation techniques into the proximal DCA to possibly accelerate this algorithm. Specifically, we consider the following algorithm for solving the DC optimization problem (4.1):

```
Proximal difference-of-convex algorithm with extrapolation (pDCA$_e$):
Input: $x^0 \in \operatorname{dom} P_1$, $\{\beta_t\} \subseteq [0,1)$ with $\sup_t \beta_t < 1$. Set $x^{-1} = x^0$.
    for $t = 0, 1, 2, \cdots$
        Take any $\xi^t \in \partial P_2(x^t)$ and set
            $y^t = x^t + \beta_t(x^t - x^{t-1}),$
            $x^{t+1} = \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ \langle \nabla f(y^t) - \xi^t, y \rangle + \frac{L}{2}\|y - y^t\|^2 + P_1(y) \right\}.$    (4.4)
    end for
```

Compared with the subproblem (4.3) in the proximal DCA, pDCA$_e$ is exactly the same as the proximal DCA when $\beta_t \equiv 0$. Thus, we can view the proximal DCA as a special case of pDCA$_e$. Moreover, from the framework of pDCA$_e$, we see that the extrapolation coefficients $\{\beta_t\}$ are general enough to comprise many commonly used extrapolation coefficients such as the extrapolation parameters used in FISTA with fixed restart schemes or FISTA with both fixed and adaptive restart schemes for solving (4.1) with a void P_2 [26]. We can see the concrete introduction of FISTA and the restart schemes for FISTA in Subsection 3.2.3.

Here we just give an overview of the restart schemes used in FISTA [26]. In these restart schemes, one first sets $\theta_{-1} = \theta_0 = 1$, then recursively takes for $t \geq 0$ that

$$\beta_t = \frac{\theta_{t-1} - 1}{\theta_t} \quad \text{with} \quad \theta_{t+1} = \frac{1 + \sqrt{1 + 4\theta_t^2}}{2}. \tag{4.5}$$

The restart schemes used in [26] focus on the update θ_t. Concretely, under some suitable conditions, one resets $\theta_{t-1} = \theta_t = 1$ for some $t > 0$. In the fixed restart scheme, one takes a fixed positive integer T and then resets $\theta_{t-1} = \theta_t = 1$ every T iterations. While in the adaptive restart scheme, one resets $\theta_{t-1} = \theta_t = 1$ if

$\langle y^{t-1} - x^t, x^t - x^{t-1}\rangle > 0$. In view of the above definitions and discussions, we can easily deduce by induction that the extrapolation coefficients $\{\beta_t\}$ chosen as in FISTA with fixed restart scheme or FISTA with both fixed and adaptive restart schemes satisfy $\{\beta_t\} \subseteq [0, 1)$ and $\sup_t \beta_t < 1$, which accord with the assumptions on $\{\beta_t\}$ in pDCA$_e$.[1] Moreover, the extrapolation parameters $\{\beta_t\}$ chosen as in FISTA with both fixed and adaptive restart schemes will be used in Section 4.3.

4.2 Convergence analysis of pDCA$_e$

In this section, we consider the convergence property pDCA$_e$ for solving (4.1). We first establish the global subsequential convergence of pDCA$_e$. Then, by making an additional differentiability assumption on P_2 and assuming that the Kurdyka-Łojasiewicz property holds for an auxiliary function, we prove the global convergence of the whole sequence generated by pDCA$_e$ and analyze the rate of convergence.

4.2.1 Global subsequential convergence of pDCA$_e$

We start with the following definition of stationary points; see, for example, [63, Equation (23)] and [30, Remark 1].

Definition 4.2.1. *Suppose that F is the objective function in* (4.1). *We say that $\bar{x}$ is a stationary point of F if*

$$0 \in \nabla f(\bar{x}) + \partial P_1(\bar{x}) - \partial P_2(\bar{x}).$$

We use $\mathcal{X}$ to denote the set of all stationary points of F.

Then it is not hard to show that any local minimizer of F is a stationary point of F. Concretely, let $\tilde{x}$ be a local minimizer of F, then we obtain from [53, Theorem 10.1]

[1] Indeed, FISTA with fixed restart scheme and FISTA with both fixed and adaptive restart schemes are special cases of pDCA$_e$ for (4.1) with $P_2 = 0$.

that $0 \in \partial F(\tilde{x})$. From this result, locally Lipschitz continuity of the continuous convex function P_2, the smoothness of f, [53, Exercise 8.8] and [53, Exercise 10.10], we see further that

$$\begin{aligned} 0 &\in \nabla f(\tilde{x}) + \partial P_1(\tilde{x}) + \partial(-P_2)(\tilde{x}) \\ &\subseteq \nabla f(\tilde{x}) + \partial P_1(\tilde{x}) + \bar{\partial}(-P_2)(\tilde{x}) \\ &= \nabla f(\tilde{x}) + \partial P_1(\tilde{x}) - \bar{\partial} P_2(\tilde{x}) \\ &= \nabla f(\tilde{x}) + \partial P_1(\tilde{x}) - \partial P_2(\tilde{x}), \end{aligned}$$

where the set inclusion follows from [16, Theorem 5.2.22] with $\bar{\partial}(-P_2)$ denoting the Clarke subdifferential of the locally Lipschitz function $-P_2$ (see [24, Page 27] and [16, Definition 5.2.3]), the first equality follows from [24, Proposition 2.3.1], while the last equality holds because both the Clarke subdifferential and the limiting subdifferential coincide with the classical subdifferential in convex analysis when the function is convex and continuous; see [24, Proposition 2.3.6] and [53, Proposition 8.12].

Next, we start to analyze the convergence properties of pDCA$_e$ applied to solving (4.1). Before giving the convergence results, we recall that the objective function F in (4.1) is level-bounded, the extrapolation coefficients $\{\beta_t\}$ in pDCA$_e$ satisfy $\sup\limits_t \beta_t < 1$ and $\{\beta_t\} \subseteq [0, 1)$.

Theorem 4.2.1. *Suppose that* $\{x^t\}$ *is a sequence generated by* pDCA$_e$ *for solving* (4.1)*. Then we have the following results.*

(i) *The sequence* $\{x^t\}$ *is bounded.*

(ii) $\lim_{t\to\infty} \|x^{t+1} - x^t\| = 0$.

(iii) *Any accumulation point of* $\{x^t\}$ *is a stationary point of* F.

Proof. Noting that x^{t+1} is the global minimizer of a strongly convex function from (4.4) in pDCA$_e$, by comparing the objective values of this strongly convex function at x^{t+1} and x^t, we obtain that

$$\begin{aligned}
&\langle \nabla f(y^t) - \xi^t, x^{t+1}\rangle + \frac{L}{2}\|x^{t+1} - y^t\|^2 + P_1(x^{t+1}) \\
&\le \langle \nabla f(y^t) - \xi^t, x^t\rangle + \frac{L}{2}\|x^t - y^t\|^2 + P_1(x^t) - \frac{L}{2}\|x^{t+1} - x^t\|^2.
\end{aligned} \tag{4.6}$$

From the Lipschitz continuity of ∇f with a modulus of $L > 0$, we see that

$$f(x^{t+1}) \le f(y^t) + \langle \nabla f(y^t), x^{t+1} - y^t\rangle + \frac{L}{2}\|x^{t+1} - y^t\|^2. \tag{4.7}$$

Adding $P(x^{t+1})$ on both sides of (4.7), then by the definition of P, we see further that

$$\begin{aligned}
f(x^{t+1}) + P(x^{t+1}) &\le f(y^t) + \langle \nabla f(y^t), x^{t+1} - y^t\rangle + \frac{L}{2}\|x^{t+1} - y^t\|^2 + P(x^{t+1}) \\
&= f(y^t) + \langle \nabla f(y^t), x^{t+1} - y^t\rangle + \frac{L}{2}\|x^{t+1} - y^t\|^2 + P_1(x^{t+1}) - P_2(x^{t+1}).
\end{aligned} \tag{4.8}$$

In view of the convexity of P_2, we immediately have

$$P_2(x^t) - P_2(x^{t+1}) \le \langle \xi^t, x^t - x^{t+1}\rangle, \tag{4.9}$$

where $\xi^t \in \partial P_2(x^t)$. Combining (4.9) and (4.8), we see that

$$\begin{aligned}
f(x^{t+1}) + P(x^{t+1}) &\le f(y^t) + \langle \nabla f(y^t), x^{t+1} - y^t\rangle + \frac{L}{2}\|x^{t+1} - y^t\|^2 \\
&\quad + P_1(x^{t+1}) - P_2(x^t) + \langle \xi^t, x^t - x^{t+1}\rangle \\
&= f(y^t) + \langle \nabla f(y^t), x^{t+1} - x^t\rangle + \langle \nabla f(y^t), x^t - y^t\rangle + \frac{L}{2}\|x^{t+1} - y^t\|^2 \\
&\quad + P_1(x^{t+1}) - P_2(x^t) + \langle \xi^t, x^t - x^{t+1}\rangle \\
&= f(y^t) + \langle \nabla f(y^t) - \xi^t, x^{t+1} - x^t\rangle + + \frac{L}{2}\|x^t - y^t\|^2 + P_1(x^{t+1}) \\
&\quad + \langle \nabla f(y^t), x^t - y^t\rangle - P_2(x^t)
\end{aligned}$$

$$
\begin{aligned}
&\leq f(y^t) + \frac{L}{2}\|x^t - y^t\|^2 + P_1(x^t) - \frac{L}{2}\|x^{t+1} - x^t\|^2 + \langle \nabla f(y^t), x^t - y^t\rangle - P_2(x^t) \\
&= f(y^t) + \langle \nabla f(y^t), x^t - y^t\rangle + \frac{L}{2}\|x^t - y^t\|^2 + P_1(x^t) - P_2(x^t) - \frac{L}{2}\|x^{t+1} - x^t\|^2 \\
&\leq f(x^t) + P(x^t) + \frac{L}{2}\|x^t - y^t\|^2 - \frac{L}{2}\|x^{t+1} - x^t\|^2,
\end{aligned}
\tag{4.10}
$$

where the second inequality is a consequence of (4.6) and the last inequality holds because f is convex and $P = P_1 - P_2$. Now, recalling the definition of y^t from (4.4), we see further from (4.10) that

$$f(x^{t+1}) + P(x^{t+1}) \leq f(x^t) + P(x^t) + \frac{L}{2}\beta_t^2\|x^t - x^{t-1}\|^2 - \frac{L}{2}\|x^{t+1} - x^t\|^2.$$

Consequently, we have upon rearranging terms that

$$\frac{L}{2}(1-\beta_t^2)\|x^t-x^{t-1}\|^2 \leq \left[f(x^t) + P(x^t) + \frac{L}{2}\|x^t - x^{t-1}\|^2\right]-\left[f(x^{t+1}) + P(x^{t+1}) + \frac{L}{2}\|x^{t+1} - x^t\|^2\right]. \tag{4.11}$$

Using the fact $\{\beta_t\} \subset [0,1)$ and (4.11) , we deduce that the sequence $\{f(x^t) + P(x^t) + \frac{L}{2}\|x^t - x^{t-1}\|^2\}$ is nonincreasing. From this and the assumption in the algorithm that $x^0 = x^{-1}$, we see further that for all $t \geq 0$

$$f(x^t) + P(x^t) \leq f(x^t) + P(x^t) + \frac{L}{2}\|x^t - x^{t-1}\|^2 \leq f(x^0) + P(x^0).$$

Since the objective $f + P$ is level-bounded from our assumption, we obtain from the above inequality that $\{x^t\}$ is bounded. This proves (i).

Next we prove (ii). Summing both sides of (4.11) from $t = 0$ to ∞, we obtain that

$$
\begin{aligned}
\frac{L}{2}\sum_{t=0}^{\infty}(1 - \beta_t^2)\|x^t - x^{t-1}\|^2 &\leq f(x^0) + P(x^0) - \liminf_{t\to\infty}\left[f(x^{t+1}) + P(x^{t+1}) + \frac{L}{2}\|x^{t+1} - x^t\|^2\right] \\
&\leq f(x^0) + P(x^0) - v < \infty.
\end{aligned}
$$

In view of the above relation and the fact $\sup_t \beta_t < 1$, we immediately deduce that $\lim_{t\to\infty} \|x^{t+1} - x^t\| = 0$. This proves (ii).

Finally, choose $\bar{x}$ to be an arbitrary accumulation point of $\{x^t\}$, thus there exists a subsequence $\{x^{t_i}\}$ such that $\lim_{i\to\infty} x^{t_i} = \bar{x}$. Then, in view of the first-order optimality condition of the subproblem (4.4) at point x^{t_i+1}, we have

$$-L(x^{t_i+1} - y^{t_i}) \in \partial P_1(x^{t_i+1}) + \nabla f(y^{t_i}) - \xi^{t_i}.$$

Substituting $y^{t_i} = x^{t_i} + \beta_{t_i}(x^{t_i} - x^{t_i-1})$ into the above inclusion, we obtain further that

$$-L[(x^{t_i+1} - x^{t_i}) - \beta_{t_i}(x^{t_i} - x^{t_i-1})] \in \partial P_1(x^{t_i+1}) + \nabla f(y^{t_i}) - \xi^{t_i}. \tag{4.12}$$

In addition, since P_2 is continuous and convex, which together with the boundedness of $\{x^{t_i}\}$ from (i) imply that the sequence $\{\xi^{t_i}\}$ is bounded. Thus, without loss of generality, we may assume that $\lim_{i\to\infty} \xi^{t_i}$ exists by passing to a further subsequence if necessary, which belongs to $\partial P_2(\bar{x})$ due to the closedness of ∂P_2. Using this and invoking $\|x^{t_i+1} - x^{t_i}\| \to 0$ from (ii) together with the closedness of ∂P_1 and the continuity of ∇f, we have upon passing to the limit in (4.12) that

$$0 \in \partial P_1(\bar{x}) + \nabla f(\bar{x}) - \partial P_2(\bar{x}),$$

which completes the proof. □

The following proposition gives convergence results of $\{F(x^t)\}$ for a sequence $\{x^t\}$ generated by pDCA$_e$. We will apply the results to establishing the global convergence of $\{x^t\}$ under additional assumptions in the next subsection.

Proposition 4.2.2. *Suppose that $\{x^t\}$ is a sequence generated by* pDCA$_e$ *for solving* (4.1). *Then the following statements hold.*

(i) $\zeta := \lim_{t\to\infty} F(x^t)$ *exists.*

(ii) $F \equiv \zeta$ *on* Ω*, where* Ω *is the set of accumulation points of* $\{x^t\}$*.*

Proof. In view of (4.11) and the fact $\{\beta_t\} \subseteq [0,1)$, the sequence $\{F(x^t) + \frac{L}{2}\|x^t - x^{t-1}\|^2\}$ is nonincreasing. Using this and the lower boundedness of $\{F(x^t) + \frac{L}{2}\|x^t - x^{t-1}\|^2\}$ together with the result $\|x^{t+1} - x^t\| \to 0$ from Theorem 4.2.1(ii), we immediately deduce that $\zeta := \lim_{t\to\infty} F(x^t)$ exists, which proves (i).

Now we prove (ii). We first note from Theorem 4.2.1(i) and (iii) that $\emptyset \neq \Omega \subseteq \mathcal{X}$. Take any $\hat{x} \in \Omega$. Hence, there exists a convergent subsequence $\{x^{t_i}\}$ such that $\lim_{i\to\infty} x^{t_i} = \hat{x}$. Since x^{t_i} is the minimizer of the subproblem (4.4), we see that

$$P_1(x^{t_i}) + \langle \nabla f(y^{t_i-1}) - \xi^{t_i-1}, x^{t_i}\rangle + \frac{L}{2}\|x^{t_i} - y^{t_i-1}\|^2 \leq P_1(\hat{x}) + \langle \nabla f(y^{t_i-1}) - \xi^{t_i-1}, \hat{x}\rangle + \frac{L}{2}\|\hat{x} - y^{t_i-1}\|^2.$$

Rearranging terms, we obtain further that

$$P_1(x^{t_i}) + \langle \nabla f(y^{t_i-1}) - \xi^{t_i-1}, x^{t_i} - \hat{x}\rangle + \frac{L}{2}\|x^{t_i} - y^{t_i-1}\|^2 \leq P_1(\hat{x}) + \frac{L}{2}\|\hat{x} - y^{t_i-1}\|^2. \tag{4.13}$$

On the other hand, observe that

$$\|\hat{x} - y^{t_i-1}\| = \|\hat{x} - x^{t_i} + x^{t_i} - y^{t_i-1}\| \leq \|\hat{x} - x^{t_i}\| + \|x^{t_i} - y^{t_i-1}\| \tag{4.14}$$

and that

$$\begin{aligned} \|x^{t_i} - y^{t_i-1}\| &= \|x^{t_i} - x^{t_i-1} - \beta_{t_i-1}(x^{t_i-1} - x^{t_i-2})\| \\ &\leq \|x^{t_i} - x^{t_i-1}\| + \|x^{t_i-1} - x^{t_i-2}\|, \end{aligned} \tag{4.15}$$

where we made use of the fact that $y^{t_i-1} = x^{t_i-1} + \beta_{t_i-1}(x^{t_i-1} - x^{t_i-2})$ for the equality. Since $\|x^{t+1} - x^t\| \to 0$ from Theorem 4.2.1(ii) and $\lim_{i\to\infty} x^{t_i} = \hat{x}$, we have by passing to the limits in (4.14) and (4.15) that

$$\|\hat{x} - y^{t_i-1}\| \to 0 \quad \text{and} \quad \|x^{t_i} - y^{t_i-1}\| \to 0. \tag{4.16}$$

In addition, we deduce from the convexity and continuity of P_2 and the fact that $\lim_{i\to\infty} x^{t_i} = \hat{x}$ that the sequence $\{\xi^{t_i}\}$ is bounded. Using this and (4.16), we obtain further that

$$\begin{aligned}\zeta &= \lim_{i\to\infty} f(x^{t_i}) + P(x^{t_i})\\ &= \lim_{i\to\infty} f(x^{t_i}) + P(x^{t_i}) + \langle \nabla f(y^{t_i-1}) - \xi^{t_i-1}, x^{t_i} - \hat{x}\rangle + \frac{L}{2}\|x^{t_i} - y^{t_i-1}\|^2\\ &\le \limsup_{i\to\infty} f(x^{t_i}) + P_1(\hat{x}) - P_2(x^{t_i}) + \frac{L}{2}\|\hat{x} - y^{t_i-1}\|^2 = F(\hat{x}),\end{aligned}$$

where we made use of (4.13) and the definition of P for the inequality. Finally, since F is lower semicontinuous, we also have

$$F(\hat{x}) \le \liminf_{i\to\infty} F(x^{t_i}) = \lim_{i\to\infty} F(x^{t_i}) = \zeta.$$

Consequently, $F(\hat{x}) = \lim_{i\to\infty} F(x^{t_i}) = \zeta$ from the above discussion. Since $\hat{x} \in \Omega$ is arbitrary, we conclude that $F \equiv \zeta$ on Ω. This completes the proof. □

4.2.2 Global convergence of pDCA$_e$

This subsection focuses on the convergence analysis of $\{x^t\}$ generated by pDCA$_e$ for solving (4.1) by adding some suitable assumptions. Moreover, we also establish the convergence rate of $\{x^t\}$. We start by introducing the following assumption.

Assumption 2. *The function P_2 in* (4.1) *is continuously differentiable on an open set $\mathcal{N}_0$ that contains $\mathcal{X}$. Moreover, the gradient ∇P_2 is locally Lipschitz continuous on $\mathcal{N}_0$.*

At first glance, Assumption 2 seems to be restrictive. However, it can be satisfied by many DC regularizers $P(x)$ that arise in practical applications such as compressed sensing [64], statistical learning problems [28, 65] and so on. Next we give some concrete examples which satisfy Assumption 2.

Example 1. *The first example is the least squares problem with ℓ_{1-2} regularization [64], which is in following form:*

$$\min_{x\in\mathbb{R}^n} F_{\ell_{1-2}}(x) = \frac{1}{2}\|Ax-b\|^2 + \lambda\|x\|_1 - \lambda\|x\|, \tag{4.17}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $\lambda > 0$ are given. In addition, we assume that A does not have zero columns in order to guarantee that the objective $F_{\ell_{1-2}}$ is level-bounded, see the concrete proof in [64, Lemma 3.1] and [36, Example 4.1(b)]. It is easy to see that problem (4.17) *corresponds to* (4.1) *with $f(x) = \frac{1}{2}\|Ax-b\|^2$, $P_1(x) = \lambda\|x\|_1$ and $P_2(x) = \lambda\|x\|$.*

Next, we will prove that 0 *is not a stationary point of $F_{\ell_{1-2}}$ when $2\lambda < \|A^T b\|_\infty$. Suppose to the contrary that $0 \in \mathcal{X}$, then by the definition of stationary point of* (4.1)*, we have $A^T b \in \lambda\partial\|0\|_1 - \lambda\partial\|0\|$. Computing the subdifferentials of $\|\cdot\|_1$ and $\|\cdot\|$ at point 0, we obtain further that*

$$A^T b \in \lambda[-1,1]^n - \lambda B(0,1),$$

where $B(0,1) = \{x \in \mathbb{R}^n : \|x\| \leq 1\}$. In view of this, we have $\|A^T b\|_\infty \leq 2\lambda$, which is a contradiction.

Hence, we conclude that if $\lambda < \frac{1}{2}\|A^T b\|_\infty$, then 0 *is not a stationary point of $F_{\ell_{1-2}}$. Due to closedness of the stationary points set $\mathcal{X}$, we can easily construct an open set $\mathcal{N}_0$ which contains $\mathcal{X}$ to make that P_2 is continuously differentiable with locally Lipschitz gradient on $\mathcal{N}_0$. From the above discussions, we see that Assumption 2 is satisfied for* (4.17) *if we choose $\lambda < \frac{1}{2}\|A^T b\|_\infty$.*

Example 2. *We next present the minmax concave penalty (MCP) regularization [65], whose DC decomposition can be found in [1, 30]:*

$$P(x) = \lambda\sum_{i=1}^{n}\int_0^{|x_i|}\left[1-\frac{x}{\theta\lambda}\right]_+ dx = \lambda\|x\|_1 - \underbrace{\lambda\sum_{i=1}^{n}\int_0^{|x_i|}\min\left\{1,\frac{x}{\theta\lambda}\right\}dx}_{P_2(x)},$$

where θ is a fixed positive number, $\lambda > 0$ is the regularization parameter and $[x]_+ = \max\{0, x\}$. It is routine to show that P_2 is continuously differentiable and

$$\nabla_i P_2(x) = \lambda \operatorname{sign}(x_i) \min\{1, |x_i|/(\theta\lambda)\}.$$

In addition, from the above formulation, we can readily show that ∇P_2 is Lipschitz continuous with a modulus $\frac{1}{\theta}$.

Example 3. *We consider the smoothly clipped absolute deviation (SCAD) regularization [28], whose DC decomposition can be found in [1, 30]:*

$$P(x) = \lambda \sum_{i=1}^{n} \int_0^{|x_i|} \min\left\{1, \frac{[\theta\lambda - x]_+}{(\theta - 1)\lambda}\right\} dx = \lambda\|x\|_1 - \underbrace{\lambda \sum_{i=1}^{n} \int_0^{|x_i|} \frac{[\min\{\theta\lambda, x\} - \lambda]_+}{(\theta - 1)\lambda} dx}_{P_2(x)},$$

where $\theta > 2$ is a constant and the regularization parameter $\lambda > 0$ is given. By simply computing, we immediately obtain that P_2 is continuously differentiable and the partial gradient of P_2 is given by

$$\nabla_i P_2(x) = \operatorname{sign}(x_i) \frac{[\min\{\theta\lambda, |x_i|\} - \lambda]_+}{\theta - 1}.$$

Using this relation, it is easy to show that $\frac{1}{\theta-1}$ is a Lipschitz continuity modulus of ∇P_2.

Example 4. *We consider the transformed ℓ_1 regularization [66] in this example. The DC decomposition of this regularization function is given in [1]:*

$$P(x) = \sum_{i=1}^{n} \frac{(a+1)|x_i|}{a + |x_i|} = \frac{a+1}{a}\|x\|_1 - \underbrace{\sum_{i=1}^{n} \left[\frac{a+1}{a}|x_i| - \frac{(a+1)|x_i|}{a + |x_i|}\right]}_{P_2(x)},$$

where $a > 0$ is given . We can see from [1, Section 5.4] that $P_2(x)$ is Lipshcitz continuously differentiable, and a Lipschitz continuity modulus of $\nabla P_2(x)$ can be taken as $\frac{2(a+1)}{a^2}$.

Example 5. *In the last example, we consider the logarithmic penalty function [19], whose DC decomposition can be found in [1, 30]:*

$$P(x) = \sum_{i=1}^{n} [\lambda \log(|x_i| + \epsilon) - \lambda \log \epsilon] = \frac{\lambda}{\epsilon}\|x\|_1 - \underbrace{\sum_{i=1}^{n} \lambda \left[\frac{|x_i|}{\epsilon} - \log(|x_i| + \epsilon) + \log \epsilon\right]}_{P_2(x)},$$

where $\lambda > 0$ and $\epsilon > 0$ are fixed numbers. By the formulation of $P_2(x)$, it is routine to show that $P_2(x)$ is continuously differentiable with a Lipschitz continuous gradient whose Lipschitz continuity modulus can be chosen as $\frac{\lambda}{\epsilon^2}$.

We next present our global convergence analysis. We will show that the sequence $\{x^t\}$ generated by pDCA_e is convergent to a stationary point of F under suitable assumptions. The global convergence results for many algorithms based on the KL property are considered in [3, 4, 5, 7]. Our analysis mainly follows and applies the recent simple general methodology developed in [14], within the necessary adequate adaptations required in the analysis to handle and use an auxiliary function H, which is defined as follows:

$$H(x, y) = f(x) + P(x) + \frac{L}{2}\|x - y\|^2. \tag{4.18}$$

Next we will give the global convergence results in this chapter.

Theorem 4.2.3. (Global convergence of pDCA_e) *Suppose that Assumption 2 holds. Suppose further that H is a KL function and the sequence $\{x^t\}$ is a sequence generated by pDCA_e for solving (4.1). Then the following statements hold.*

(i) $\lim_{t\to\infty} \text{dist}((0, 0), \partial H(x^t, x^{t-1})) = 0.$

(ii) *The sequence $\{H(x^t, x^{t-1})\}$ is nonincreasing and $\lim_{t\to\infty} H(x^t, x^{t-1}) = \zeta$, where ζ is given in Proposition 4.2.2.*

(iii) *The set of accumulation points of* $\{(x^t, x^{t-1})\}$ *is* $\Upsilon := \{(x,x) : \ x \in \Omega\}$ *and* $H \equiv \zeta$ *on* Υ*, where* Ω *is the set of accumulation points of* $\{x^t\}$*.*

(iv) *The sequence* $\{x^t\}$ *is convergent to a stationary point of* F*; moreover,* $\sum_{t=1}^{\infty} \|x^t - x^{t-1}\| < \infty$*.*

Proof. Using the result $\{x^t\}$ is bounded from Theorem 4.2.1(i) and the definition of Ω, we immediately obtain that

$$\lim_{t\to\infty} \operatorname{dist}(x^t, \Omega) = 0.$$

Since $\Omega \subseteq \mathcal{X}$ by Theorem 4.2.1(iii), thus we see that for arbitrary $\nu > 0$, there must exist $T_0 > 0$ so that $\operatorname{dist}(x^t, \Omega) < \nu$ and $x^t \in \mathcal{N}_0$ whenever $t \geq T_0$, where $\mathcal{N}_0$ is the open set from Assumption 2. Moreover, noting that Ω is compact due to the boundedness of $\{x^t\}$, by shrinking ν if necessary, we may assume without loss of generality that ∇P_2 is globally Lipschitz continuous on the bounded set $\mathcal{N} := \{x \in \mathcal{N}_0 : \ \operatorname{dist}(x, \Omega) < \nu\}$.

Next, considering the subdifferential of the function H in (4.18) at the point (x^t, x^{t-1}) for $t \geq T_0$, we have

$$\partial_x H(x^t, x^{t-1}) = \nabla f(x^t) + \partial P_1(x^t) - \nabla P_2(x^t) + L(x^t - x^{t-1}),$$

$$\partial_y H(x^t, x^{t-1}) = -L(x^t - x^{t-1}),$$

the above relations follow from the definition of P, the facts that P_2 is continuously differentiable in $\mathcal{N}$ and that $x^t \in \mathcal{N}$ for $t \geq T_0$. Hence, the subdifferential of H at point (x^t, x^{t-1}) for $t \geq T_0$ can be written as the following form:

$$\partial H(x^t, x^{t-1}) = [\{\nabla f(x^t) - \nabla P_2(x^t) + L(x^t - x^{t-1})\} + \partial P_1(x^t)] \times \{-L(x^t - x^{t-1})\}. \tag{4.19}$$

On the other hand, using the first order optimality condition of the subproblem (4.4) in pDCA$_e$, we have for any $t \geq T_0 + 1$ that

$$-L(x^t - y^{t-1}) - \nabla f(y^{t-1}) + \nabla P_2(x^{t-1}) \in \partial P_1(x^t),$$

where we use the continuous differentiability of P_2 in $\mathcal{N}$ and $x^{t-1} \in \mathcal{N}$ whenever $t \geq T_0 + 1$. In view of this relation, we obtain further that

$$\begin{aligned}
&- L(x^{t-1} - y^{t-1}) + \nabla f(x^t) - \nabla f(y^{t-1}) + \nabla P_2(x^{t-1}) - \nabla P_2(x^t) \\
&= \nabla f(x^t) - \nabla P_2(x^t) + L(x^t - x^{t-1}) - L(x^t - y^{t-1}) - \nabla f(y^{t-1}) + \nabla P_2(x^{t-1}) \\
&\in \nabla f(x^t) - \nabla P_2(x^t) + L(x^t - x^{t-1}) + \partial P_1(x^t).
\end{aligned}$$

Combining this with (4.19), we have

$$(-L(x^{t-1}-y^{t-1})+\nabla f(x^t)-\nabla f(y^{t-1})+\nabla P_2(x^{t-1})-\nabla P_2(x^t), -L(x^t-x^{t-1})) \in \partial H(x^t, x^{t-1}).$$

Using this relation, the definition of y^t from (4.4) and the global Lipschitz continuity of ∇f and ∇P_2 on $\mathcal{N}$, we deduce that there exists $C > 0$, for $t \geq T_0 + 1$

$$\operatorname{dist}((0,0), \partial H(x^t, x^{t-1})) \leq C(\|x^t - x^{t-1}\| + \|x^{t-1} - x^{t-2}\|). \tag{4.20}$$

Since $\|x^{t+1} - x^t\| \to 0$ from Theorem 4.2.1(ii), we conclude from (4.20) that

$$\lim_{t\to\infty} \operatorname{dist}((0,0), \partial H(x^t, x^{t-1})) = 0,$$

which proves (i).

We now prove (ii) and (iii). Combining the definition of H and (4.11) together with the fact that $\sup\limits_t \beta_t < 1$ from pDCA$_e$, there must exist a positive number D such that

$$H(x^t, x^{t-1}) - H(x^{t+1}, x^t) \geq D\|x^t - x^{t-1}\|^2 \tag{4.21}$$

for all t. In particular, the sequence $\{H(x^t, x^{t-1})\}$ is nonincreasing. And this sequence is also bounded below by v from the definition of H, hence we see that the sequence $\{H(x^t, x^{t-1})\}$ is convergent. Next, using the result $\|x^t - x^{t-1}\| \to 0$ according to Theorem 4.2.1(ii), we obtain that Υ is the set of accumulation points of $\{(x^t, x^{t-1})\}_{t\geq 1}$. Moreover, in view of Proposition 4.2.2(i), we see further that

$$\lim_{t\to\infty} H(x^t, x^{t-1}) = \zeta.$$

Furthermore, for any $(\hat{x}, \hat{x}) \in \Upsilon$, by the definition, we see that $\hat{x} \in \Omega$. Hence, in view of Proposition 4.2.2(ii), we obtain $H(\hat{x}, \hat{x}) = F(\hat{x}) = \zeta$. As $\hat{x} \in \Omega$ is arbitrary, we conclude that $H \equiv \zeta$ on Υ. This proves (ii) and (iii).

Finally, we prove (iv). In view of Theorem 4.2.1(iii), we just need to show that $\{x^t\}$ is convergent. We carry out our proof in two cases. We first consider the case that there exists a $t > 0$ such that $H(x^t, x^{t-1}) = \zeta$. From (ii), $\{H(x^t, x^{t-1})\}$ is nonincreasing and convergent to ζ, thus we conclude that for arbitrary $\bar{t} \geq 0$, $H(x^{t+\bar{t}}, x^{t+\bar{t}-1}) = \zeta$. Using this and (4.21), we immediately have that $x^t = x^{t+\bar{t}}$ for any $\bar{t} \geq 0$, which means that $\{x^t\}$ converges finitely.

We next consider the second case, which is that $H(x^t, x^{t-1}) > \zeta$ for all t. Recalling that H is a KL function, Υ is a compact subset of $\operatorname{dom} \partial H$ and $H \equiv \zeta$ on Υ, we see from Lemma 2.3.1 that there exist an $\epsilon > 0$ and a continuous concave function $\phi \in \Xi_a$ with $a > 0$ such that for all $(x, y) \in U$

$$1 \leq \phi'(H(x, y) - \zeta)\mathrm{dist}((0, 0), \partial H(x, y)), \tag{4.22}$$

where

$$U = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : \mathrm{dist}((x, y), \Upsilon) < \epsilon\} \cap \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : \zeta < H(x, y) < \zeta + a\}.$$

Using the fact $\{x^t\}$ is bounded due to Theorem 4.2.1(i) and Υ is the set of accumulation points of $\{(x^t, x^{t-1})\}_{t\geq 1}$ from (iii), we obtain that

$$\lim_{t\to\infty} \mathrm{dist}((x^t, x^{t-1}), \Upsilon) = 0.$$

Hence, we see from the above relation that there exists $T_1 > 0$ such that $\mathrm{dist}((x^t, x^{t-1}), \Upsilon) < \epsilon$ for all $t \geq T_1$. In addition, since the sequence $\{H(x^t, x^{t-1})\}$ is nonincreasing and convergent to ζ by (ii), there exists $T_2 > 0$ such that $\xi < H(x^t, x^{t-1}) < \xi + a$ for all $t \geq T_2$. Taking $\bar{T} = \max\{T_0 + 1, T_1, T_2\}$, from the above discussion, we conclude that the sequence $\{(x^t, x^{t-1})\}_{t \geq \bar{T}}$ belongs to U. Combining this with (4.22), we obtain that

$$\phi'(H(x^t, x^{t-1}) - \zeta) \cdot \mathrm{dist}((0,0), \partial H(x^t, x^{t-1})) \geq 1, \quad \text{for all } t \geq \bar{T}. \tag{4.23}$$

Using the concavity of ϕ, we see further that for any $t \geq \bar{T}$,

$$\begin{aligned} &\big[\phi(H(x^t, x^{t-1}) - \zeta) - \phi(H(x^{t+1}, x^t) - \zeta)\big] \cdot \mathrm{dist}((0,0), \partial H(x^t, x^{t-1})) \\ &\geq \phi'(H(x^t, x^{t-1}) - \zeta)) \cdot \mathrm{dist}((0,0), \partial H(x^t, x^{t-1})) \cdot (H(x^t, x^{t-1}) - H(x^{t+1}, x^t)) \\ &\geq H(x^t, x^{t-1}) - H(x^{t+1}, x^t), \end{aligned}$$

where the last inequality is made use of (4.23) and the fact that $\{H(x^t, x^{t-1})\}$ is nonincreasing according to (ii). Combining this with (4.20) and (4.21), we have

$$\big[\phi(H(x^t, x^{t-1}) - \zeta) - \phi(H(x^{t+1}, x^t) - \zeta)\big] \cdot C(\|x^t - x^{t-1}\| + \|x^{t-1} - x^{t-2}\|) \geq D\|x^t - x^{t-1}\|^2$$

whenever $t \geq \bar{T}$. By rearranging terms, we obtain that for any $t \geq \bar{T}$,

$$\|x^t - x^{t-1}\|^2 \leq \frac{C}{D}\big(\phi(H(x^t, x^{t-1}) - \zeta) - \phi(H(x^{t+1}, x^t) - \zeta)\big) \cdot \big(\|x^t - x^{t-1}\| + \|x^{t-1} - x^{t-2}\|\big). \tag{4.24}$$

Taking square root on both sides of (4.24), then in view of the AM-GM inequality, we have

$$\begin{aligned} \|x^t - x^{t-1}\| &\leq \sqrt{\frac{2C}{D}\left(\phi(H(x^t, x^{t-1}) - \zeta) - \phi(H(x^{t+1}, x^t) - \zeta)\right)} \cdot \sqrt{\frac{\|x^t - x^{t-1}\| + \|x^{t-1} - x^{t-2}\|}{2}} \\ &\leq \frac{C}{D}\big(\phi(H(x^t, x^{t-1}) - \zeta) - \phi(H(x^{t+1}, x^t) - \zeta)\big) + \frac{1}{4}\|x^t - x^{t-1}\| + \frac{1}{4}\|x^{t-1} - x^{t-2}\|, \end{aligned}$$

which implies that

$$\frac{1}{2}\|x^t-x^{t-1}\| \leq \frac{C}{D}\left(\phi(H(x^t,x^{t-1})-\zeta)-\phi(H(x^{t+1},x^t)-\zeta)\right)+\frac{1}{4}(\|x^{t-1}-x^{t-2}\|-\|x^t-x^{t-1}\|). \tag{4.25}$$

Summing the above relation from $t=\bar{T}$ to ∞, we have

$$\sum_{t=\bar{T}}^{\infty}\|x^t-x^{t-1}\| \leq \frac{2C}{D}\phi(H(x^{\bar{T}},x^{\bar{T}-1})-\zeta)+\frac{1}{2}\|x^{\bar{T}-1}-x^{\bar{T}-2}\|<\infty.$$

We conclude from the above relation that the sequence $\{x^t\}$ is convergent and the sequence $\{\|x^{t+1}-x^t\|\}_{t\geq 0}$ is summable, which completes the proof. □

We next consider the convergence rate of the sequence $\{x^t\}$ under the assumption that the auxiliary function H is a KL function whose $\phi \in \Xi_a$ (see Definition 2.3.1) takes the form $\phi(s) = cs^{1-\theta}$ for some $\theta \in [0,1)$. This kind of convergence rate analysis has also been performed for other optimization algorithms; see, for example, [3]. Our analysis applies the technique which was first introduced in [3] but makes use of the auxiliary function H in (4.18).

Remark 2. *Indeed, both the error bound condition and KL inequality can be applied to establishing the convergence rates of many first order methods. And the error bound condition can imply the KL-inequality with an exponent $\frac{1}{2}$, when the objective is level bounded. We refer readers to the recent paper [34] for the details.*

Theorem 4.2.4. *Let $\{x^t\}$ be a sequence generated by* pDCA$_e$ *for solving* (4.1). *Suppose that Assumption 2 holds and that $\{x^t\}$ converges to some $\bar{x}$ and the auxiliary function H is a KL function with ϕ in the KL inequality* (2.3) *chosen as $\phi(s) = cs^{1-\theta}$ for some $c>0$ and $\theta \in [0,1)$. Then we have the following results.*

(i) *If $\theta = 0$, then $\{x^t\}$ converges finitely, i.e., there exists $t_0 > 0$ so that x^t is constant for all $t > t_0$;*

(ii) *If $\theta \in (0, \frac{1}{2}]$, then $\{x^t\}$ converges linearly, i.e., there exist $c_1 > 0$, $t_1 > 0$ and $\eta \in (0,1)$ such that $\|x^t - \bar{x}\| < c_1 \eta^t$ for all $t > t_1$;*

(iii) *If $\theta \in (\frac{1}{2}, 1)$, then $\{x^t\}$ converges sublinearly, i.e., there exist $c_2 > 0$ and $t_2 > 0$ such that $\|x^t - \bar{x}\| < c_2 t^{-\frac{1-\theta}{2\theta-1}}$ for all $t > t_2$.*

Proof. First, we prove (i). If $\theta = 0$, we first prove that there must exist $t_0 > 0$ such that $H(x^{t_0}, x^{t_0-1}) = \zeta$. Suppose to the contrary that $H(x^t, x^{t-1}) > \zeta$ for all $t > 0$. Using the assumption $\lim\limits_{t \to \infty} x^t = \bar{x}$ and the fact that the sequence $\{H(x^t, x^{t-1})\}$ is nonincreasing and convergent to ζ by Theorem 4.2.3(ii) together with the fact $\phi(s) = cs$ and the KL inequality (4.23), we see that for all sufficiently large t,

$$\text{dist}((0,0), \partial H(x^t, x^{t-1})) \geq \frac{1}{c},$$

which contradicts Theorem 4.2.3(i). Thus, there exists $t_0 > 0$ so that $H(x^{t_0}, x^{t_0-1}) = \zeta$. Again using the result $\{H(x^t, x^{t-1})\}$ is nonincreasing and convergent to ζ, it must then hold that $H(x^{t_0+\bar{t}}, x^{t_0+\bar{t}-1}) = \zeta$ for any $\bar{t} \geq 0$. Thus, we conclude from (4.21) that $x^{t_0} = x^{t_0+\bar{t}}$ for any $\bar{t} \geq 0$. This proves (i).

We next analyze the other two cases, which means that $\theta \in (0,1)$. From the discussion above, we see that if there exists $t_0 > 0$ such that $H(x^{t_0}, x^{t_0-1}) = \zeta$, then one can show that $\{x^t\}$ is finitely convergent, and the desired conclusions hold trivially. Hence, for $\theta \in (0,1)$, we just need to consider the case when $H(x^t, x^{t-1}) > \zeta$ for all $t > 0$.

In the following, we define $H_t = H(x^t, x^{t-1}) - \zeta$ and $S_t = \sum_{i=t}^{\infty} \|x^{i+1} - x^i\|$, where S_t is well defined due to the summability of the sequence $\{\|x^{t+1} - x^t\|\}$ by Theorem 4.2.3(iv). Then, according to (4.25), we obtain that for all $t \geq \bar{T}$ (where $\bar{T}$

is defined as in (4.23)) that

$$
\begin{aligned}
S_t &= 2\sum_{i=t}^{\infty}\frac{1}{2}\|x^{i+1}-x^i\| \leq 2\sum_{i=t}^{\infty}\frac{1}{2}\|x^i-x^{i-1}\|\\
&\leq 2\sum_{i=t}^{\infty}\left[\frac{C}{D}\left(\phi(H(x^i,x^{i-1})-\zeta)-\phi(H(x^{i+1},x^i)-\zeta)\right)+\frac{1}{4}(\|x^{i-1}-x^{i-2}\|-\|x^i-x^{i-1}\|)\right]\\
&\leq \frac{2C}{D}\phi(H(x^t,x^{t-1})-\zeta)+\frac{1}{2}\|x^{t-1}-x^{t-2}\|\\
&= \frac{2C}{D}\phi(H_t)+\frac{1}{2}(S_{t-2}-S_{t-1}).
\end{aligned}
$$

Combining this with the fact that $\{S_t\}$ is nonincreasing, we obtain further that

$$S_t \leq \frac{2C}{D}\phi(H_t)+\frac{1}{2}(S_{t-2}-S_t) \tag{4.26}$$

for all $t \geq \bar{T}$. On the other hand, since $\lim\limits_{t\to\infty} x^t = \bar{x}$ from our assumption and the sequence $\{H(x^t,x^{t-1})\}$ is nonincreasing and convergent to ζ by Theorem 4.2.3(ii), we deduce from (4.23) with $\phi(s)=cs^{1-\theta}$ that for all sufficiently large t,

$$c(1-\theta)(H_t)^{-\theta}\mathrm{dist}((0,0),\partial H(x^t,x^{t-1})) \geq 1. \tag{4.27}$$

In addition, in view of (4.20) and the definition of S_t, we see that for all sufficiently large t,

$$\mathrm{dist}((0,0),\partial H(x^t,x^{t-1})) \leq C(S_{t-2}-S_t). \tag{4.28}$$

Combining (4.27) and (4.28), we obtain by rearranging terms that for all sufficiently large t

$$(H_t)^{\theta} \leq C\cdot c(1-\theta)\cdot(S_{t-2}-S_t).$$

From the definition of H_t and S_t and the fact $\theta\in(0,1)$, we see that all the terms in the above inequality are nonnegative. Hence, raising to a power of $\frac{1-\theta}{\theta}$ to both sides of the above relation and then scaling by c, it can be shown that

$$c(H_t)^{1-\theta} \leq c\cdot\left(C\cdot c(1-\theta)\cdot(S_{t-2}-S_t)\right)^{\frac{1-\theta}{\theta}}.$$

Combining this with (4.26) and recalling that the definition of $\phi(H_t)$ (i.e., $\phi(H_t) = c(H_t)^{1-\theta}$), we see that for all sufficiently large t,

$$\begin{aligned} S_t &\leq C_1(S_{t-2} - S_t)^{\frac{1-\theta}{\theta}} + \frac{1}{2}(S_{t-2} - S_t) \\ &\leq C_1(S_{t-2} - S_t)^{\frac{1-\theta}{\theta}} + S_{t-2} - S_t, \end{aligned} \tag{4.29}$$

where $C_1 = \frac{2C}{D}c \cdot (C \cdot c(1-\theta))^{\frac{1-\theta}{\theta}}$ and the second inequality is made use of the nonincreasing property of the sequence $\{S_t\}$ by the definition.

Now we split our proof into two cases: $\theta \in (0, \frac{1}{2}]$ or $\theta \in (\frac{1}{2}, 1)$.

We first consider the case that $\theta \in (0, \frac{1}{2}]$. Then we can deduce that $\frac{1-\theta}{\theta} \geq 1$. Using the result $\|x^{t+1} - x^t\| \to 0$ from Theorem 4.2.1(ii), we immediately obtain that $S_{t-2} - S_t \to 0$. In view of this and (4.29), we conclude that there exists $t_1 > 0$ so that for all $t \geq t_1$, we have

$$S_t \leq (C_1 + 1)(S_{t-2} - S_t),$$

which implies that $S_t \leq \frac{C_1+1}{C_1+2} S_{t-2}$. Hence,

$$\|x^t - \bar{x}\| \leq \sum_{i=t}^{\infty} \|x^{i+1} - x^i\| = S_t \leq S_{t_1-2} \left(\sqrt{\frac{C_1+1}{C_1+2}} \right)^{t-t_1+1}$$

for all $t \geq t_1$. This proves (ii).

Finally, we consider the case that $\theta \in (\frac{1}{2}, 1)$. Thus, we have $\frac{1-\theta}{\theta} < 1$. Combining this with (4.29) and the fact that $S_{t-2} - S_t \to 0$ together, we see that there exists $t_2 > 0$ such that for all $t \geq t_2$, we have

$$\begin{aligned} S_t &\leq C_1(S_{t-2} - S_t)^{\frac{1-\theta}{\theta}} + S_{t-2} - S_t \\ &\leq C_1(S_{t-2} - S_t)^{\frac{1-\theta}{\theta}} + (S_{t-2} - S_t)^{\frac{1-\theta}{\theta}} \\ &= (C_1 + 1)(S_{t-2} - S_t)^{\frac{1-\theta}{\theta}}. \end{aligned}$$

Since the terms on both sides of the above inequality are all nonnegative, raising to a power of $\frac{\theta}{1-\theta}$ to both sides of the above inequality, we see further that for all $t \geq t_2$

$$S_t^{\frac{\theta}{1-\theta}} \leq C_2(S_{t-2} - S_t),$$

where $C_2 = (C_1 + 1)^{\frac{\theta}{1-\theta}}$. Consider the sequence $\Delta_t := S_{2t}$. Then for any $t \geq \lceil \frac{t_2}{2} \rceil$, we have

$$\Delta_t^{\frac{\theta}{1-\theta}} \leq C_2(\Delta_{t-1} - \Delta_t).$$

Proceeding as in the proof of [3, Theorem 2] starting from [3, Equation (13)], one can show similarly that there exists a positive $C_3 > 0$ such that for all sufficiently large t

$$\Delta_t \leq C_3 t^{-\frac{1-\theta}{2\theta-1}},$$

see the first equation on [3, Page 15]. This implies that

$$\|x^t - \bar{x}\| \leq S_t \begin{cases} = \Delta_{\frac{t}{2}} \leq 2^\rho C_3 t^{-\rho} & \text{if } t \text{ is even,} \\ \leq S_{t-1} = \Delta_{\frac{t-1}{2}} \leq 2^\rho C_3 (t-1)^{-\rho} \leq 4^\rho C_3 t^{-\rho} & \text{if } t \text{ is odd} \end{cases}$$

for all sufficiently large t (≥ 2), where $\rho := \frac{1-\theta}{2\theta-1}$. This completes the proof. □

Remark 3. *We recall that there are many concrete examples of functions f satisfying the KL property at all points in $\operatorname{dom}\partial f$ with $\phi(s) = cs^{1-\theta}$ for some $\theta \in [0,1)$ and $c > 0$. Indeed, all proper closed semialgebraic functions satisfy this property; see, for example, [13, section 2] and [4, section 4.3]. We refer the readers to [4, 34] for more examples. In particular, one can show that if $f(x) = \frac{1}{2}\|Ax - b\|^2$ for some matrix A and vector b, P is given as in any one of the five examples at the beginning of this subsection, then the function H in (4.18) is a KL function with $\phi(s) = cs^{1-\theta}$ for some $\theta \in [0,1)$ and $c > 0$.*

4.3 Numerical experiments

In this section, we perform numerical experiments on two classes of DC regularized least squares problems. All our numerical experiments are performed by Matlab 2015b on a 64-bit PC with a 3.60GHz Inter Core i7-4790 processor and 32GB of RAM.

In the numerical experiments below, we mainly deal with the DC regularized least squares problem, which takes the following form:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|^2 + P_1(x) - P_2(x), \tag{4.30}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ are given, P_1 is a proper closed convex function and P_2 is a continuous convex function. Two different regularizers are used in our experiments, they are the ℓ_{1-2} regularizer studied in Example 1 and the logarithmic regularizer discussed in Example 5. We also apply two different algorithms to solving (4.30) with the aforementioned two regularizers, one is our algorithm pDCA$_e$, and the other is the proximal DCA with nonmonotone linesearch, which is based on the proximal DCA [31] . The concrete details of these algorithms discussed above are presented as follows.

pDCA$_e$. In this algorithm, we take $L = \lambda_{\max}(A^TA)$.[2] The extrapolation coefficients $\{\beta_t\}$ in pDCA$_e$ are chosen as (4.5), and both the fixed (with $T = 200$) and the adaptive restart schemes as described in Section 4.1 or Subsection 3.2.3 are performed. We initialize the algorithm at the origin and terminate it when

$$\frac{\|x^t - x^{t-1}\|}{\max\{1, \|x^t\|\}} < 10^{-5}.$$

pDCA$_{ls}$. This algorithm is based on the proximal DCA. The main difference between pDCA$_{ls}$ and the proximal DCA is that pDCA$_{ls}$ incorporates a nonmonotone

[2] $\lambda_{\max}(A^TA)$ is computed via the MATLAB code lambda = norm(A*A'); when $m \leq 2000$, and by opts.issym = 1; lambda= eigs(A*A',1,'LM',opts); otherwise.

linesearch strategy into the proximal DCA. Concretely, the framework of pDCA$_{ls}$ is exactly the same as that of the proximal gradient algorithm with nonmonotone linesearch considered in [63] (see also [22, Appendix A, Algorithm 1]) with $f(x) = \frac{1}{2}\|Ax - b\|^2$ and $P(x) = P_1(x) - P_2(x)$ when the subproblem [22, Appendix A, A.4] is replaced by

$$\min_{x \in \mathbb{R}^n} \left\{ \langle A^T(Ax^t - b) - \xi^t, x - x^t \rangle + \frac{L_t}{2}\|x - x^t\|^2 + P_1(x) \right\},$$

where $\xi^t \in \partial P_2(x^t)$. We just use the same notation in [22, Appendix A, Algorithm 1]. In detail, we set $c = 10^{-4}$, $\tau = 2$, $M = 4$, $L_0^0 = 1$, and

$$L_t^0 = \min\left\{ \max\left\{ \frac{\|A(x^t - x^{t-1})\|^2}{\|x^t - x^{t-1}\|^2}, 10^{-8} \right\}, 10^8 \right\}$$

for $t \geq 1$. We initialize the algorithm at the origin and terminate it when

$$\frac{\|x^t - x^{t-1}\|}{\max\{1, \|x^t\|\}} < 10^{-5}.$$

In our numerical experiments below, we compare our algorithm pDCA$_\mathrm{e}$ with pDCA$_\mathrm{ls}$ for solving (4.30) on random instances generated as follows. We first generate an $m \times n$ matrix A with i.i.d. standard Gaussian entries, and then normalize this matrix so that the columns of A have unit norms. A subset T of size s is then chosen uniformly at random from $\{1, 2, 3, \ldots, n\}$ and an s-sparse vector y having i.i.d. standard Gaussian entries on T is generated. Finally, we set $b = Ay + 0.01 \cdot \hat{n}$, where $\hat{n} \in \mathbb{R}^m$ is a random vector with i.i.d. Gaussian entries.

Next, we will present the concrete DC problems we consider in the numerical experiments, and then analyze the numerical results.

4.3.1 Least squares problems with ℓ_{1-2} regularizer

This subsection mainly discusses the least squares problem with ℓ_{1-2} regularized,

$$\min_{x\in\mathbb{R}^n} F_{\ell_{1-2}}(x) = \frac{1}{2}\|Ax-b\|^2 + \lambda\|x\|_1 - \lambda\|x\|, \tag{4.31}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ are given, and $\lambda > 0$ is the regularization parameter. This problem can be reformulated as the form of (4.30) with $P_1(x) = \lambda\|x\|_1$ and $P_2(x) = \lambda\|x\|$. Moreover, we assume that the matrix A in (4.31) always does not have zero columns. In view of this assumption, the discussions in Example 1, Theorem 4.2.3 and Remark 3, we conclude that $F_{\ell_{1-2}}$ is a level-bounded function, and that the sequence $\{x^t\}$ generated by pDCA$_e$ is globally convergent when $\lambda < \frac{1}{2}\|A^Tb\|_\infty$.

Next, we will present the numerical results we obtain from our numerical experiments. In this test, we consider $(m, n, s) = (720i, 2560i, 80i)$ for $i = 1, 2, \ldots, 10$. For each triple (m, n, s), we generate 50 instances randomly as described above. The following Tables 4.1 and 4.2 present the computational results corresponding to problem (4.31) with $\lambda = 5 \times 10^{-4}$ and $\lambda = 1 \times 10^{-3}$ respectively.[3] In detail, Tables 4.1 and 4.2 report the time for computing $\lambda_{\max}(A^TA)$ ($\mathbf{t}_{\lambda_{\max}}$), the number of iterations (iter), CPU times in seconds (CPU time),[4] and the function values at termination (fval), averaged over the 50 random instances. We can see from these tables that pDCA$_e$ always outperforms pDCA$_{ls}$.

4.3.2 Least squares problems with logarithmic regularizer

This subsection focuses on the least squares problem with logarithmic regularization function,

$$\min_{x\in\mathbb{R}^n} F_{\log}(x) = \frac{1}{2}\|Ax-b\|^2 + \sum_{i=1}^{n}\left[\lambda\log(|x_i|+\epsilon) - \lambda\log\epsilon\right], \tag{4.32}$$

[3] These λ satisfy $\lambda < \frac{1}{2}\|A^Tb\|_\infty$ for all our random instances.

[4] The CPU time reported for pDCA$_e$ does not include the time for computing $\lambda_{\max}(A^TA)$.

Table 4.1: Solving (4.31) with $\lambda = 5 \times 10^{-4}$

problem size				iter		CPU time		fval	
n	m	s	$\mathbf{t}_{\lambda_{\max}}$	pDCA$_{ls}$	pDCA$_e$	pDCA$_{ls}$	pDCA$_e$	pDCA$_{ls}$	pDCA$_e$
2560	720	80	0.1	1709	882	3.29	1.20	2.9152e-02	2.9140e-02
5120	1440	160	0.8	1728	902	14.64	5.56	6.0489e-02	6.0465e-02
7680	2160	240	0.7	1757	926	32.63	12.62	9.3975e-02	9.3937e-02
10240	2880	320	1.5	1768	954	58.11	23.07	1.2712e-01	1.2706e-01
12800	3600	400	2.7	1775	970	91.01	36.59	1.6056e-01	1.6049e-01
15360	4320	480	4.2	1788	978	127.92	51.59	1.9320e-01	1.9312e-01
17920	5040	560	6.8	1773	982	171.04	70.18	2.2553e-01	2.2543e-01
20480	5760	640	9.0	1767	982	222.97	91.53	2.5625e-01	2.5614e-01
23040	6480	720	12.3	1793	982	286.26	115.87	2.9232e-01	2.9220e-01
25600	7200	800	16.4	1771	978	354.42	145.77	3.2321e-01	3.2306e-01

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ are given, ϵ is a fixed positive number, and $\lambda > 0$ is the regularization parameter. In view of Example 5, one can easily rewrite $F_{\log}$ as the form of (4.30) with $P_1(x) = \frac{\lambda}{\epsilon}\|x\|_1$ and $P_2(x) = \sum_{i=1}^n \lambda \left[\frac{|x_i|}{\epsilon} - \log(|x_i| + \epsilon) + \log \epsilon\right]$. Moreover, it is routine to show that $F_{\log}$ is level-bounded. Combining this result with Theorem 4.2.3 and Remark 3, we can conclude that $\{x^t\}$ generated by pDCA$_e$ is globally convergent to a stationary point of (4.32).

Next, we will analyze the numerical results we obtain from our numerical tests. We consider $(m, n, s) = (720i, 2560i, 80i)$ for $i = 1, 2, \ldots, 10$ in the numerical experiments below. For each triple (m, n, s), we generate 50 instances randomly as described above. The following Tables 4.3 and 4.4 present the computational results corresponding to problem (4.32) with $\lambda = 5 \times 10^{-4}$ and $\lambda = 1 \times 10^{-3}$ respectively.[5] Concretely, Tables 4.1 and 4.2 report the time for computing $\lambda_{\max}(A^TA)$ ($\mathbf{t}_{\lambda_{\max}}$), the number of iterations (iter), CPU times in seconds (CPU time),[6] and the function values at termination (fval), averaged over the 50 random instances. We can conclude

[5] We set $\epsilon = 0.5$ in (4.32).

[6] The CPU time reported for pDCA$_e$ does not include the time for computing $\lambda_{\max}(A^TA)$.

Table 4.2: Solving (4.31) with $\lambda = 1 \times 10^{-3}$

problem size				iter		CPU time		fval	
n	m	s	$\mathbf{t}_{\lambda_{\max}}$	pDCA$_{ls}$	pDCA$_e$	pDCA$_{ls}$	pDCA$_e$	pDCA$_{ls}$	pDCA$_e$
2560	720	80	0.1	918	596	1.75	0.82	5.9412e-02	5.9406e-02
5120	1440	160	0.8	926	602	7.84	3.74	1.2070e-01	1.2069e-01
7680	2160	240	0.7	950	602	17.58	8.30	1.8853e-01	1.8851e-01
10240	2880	320	1.5	947	602	30.60	14.42	2.5495e-01	2.5492e-01
12800	3600	400	2.7	937	602	47.39	22.72	3.1448e-01	3.1445e-01
15360	4320	480	4.2	939	602	66.58	31.79	3.7809e-01	3.7805e-01
17920	5040	560	6.8	959	602	92.42	43.43	4.5133e-01	4.5128e-01
20480	5760	640	8.8	954	602	118.99	55.95	5.1992e-01	5.1986e-01
23040	6480	720	12.4	953	602	149.64	70.96	5.7837e-01	5.7831e-01
25600	7200	800	16.6	954	602	188.81	89.25	6.4832e-01	6.4825e-01

from these tables that pDCA$_e$ always outperforms pDCA$_{ls}$.

Table 4.3: Solving (4.32) with $\lambda = 5 \times 10^{-4}$

problem size				iter		CPU time		fval	
n	m	s	$\mathbf{t}_{\lambda_{\max}}$	pDCA$_{ls}$	pDCA$_e$	pDCA$_{ls}$	pDCA$_e$	pDCA$_{ls}$	pDCA$_e$
2560	720	80	0.1	855	600	1.69	0.82	3.7906e-02	3.7899e-02
5120	1440	160	0.8	867	602	7.24	3.70	7.6135e-02	7.6122e-02
7680	2160	240	0.7	878	602	16.03	8.13	1.1437e-01	1.1435e-01
10240	2880	320	1.5	867	602	27.77	14.37	1.5118e-01	1.5115e-01
12800	3600	400	2.7	874	602	43.66	22.46	1.9070e-01	1.9067e-01
15360	4320	480	4.3	860	602	60.39	31.62	2.2817e-01	2.2813e-01
17920	5040	560	6.8	874	602	83.83	42.92	2.6709e-01	2.6704e-01
20480	5760	640	8.8	871	602	107.99	55.39	3.0447e-01	3.0442e-01
23040	6480	720	12.1	865	602	135.40	70.66	3.4205e-01	3.4199e-01
25600	7200	800	16.4	872	602	169.65	87.89	3.8134e-01	3.8127e-01

4.4 Conclusions of this chapter

In this chapter, we mainly consider the algorithm pDCA$_{\mathrm{e}}$ for solving a class of DC optimization problems (4.1) and further study the convergence behaviors of pDCA$_{\mathrm{e}}$.

Table 4.4: Solving (4.32) with $\lambda = 1 \times 10^{-3}$

problem size				iter		CPU time		fval	
n	m	s	$\mathbf{t}_{\lambda_{\max}}$	pDCA$_{ls}$	pDCA$_e$	pDCA$_{ls}$	pDCA$_e$	pDCA$_{ls}$	pDCA$_e$
2560	720	80	0.1	468	378	0.90	0.51	7.5333e-02	7.5330e-02
5120	1440	160	0.8	470	397	3.91	2.47	1.5081e-01	1.5080e-01
7680	2160	240	0.7	471	401	8.54	5.52	2.2800e-01	2.2799e-01
10240	2880	320	1.5	470	400	14.85	9.57	3.0344e-01	3.0343e-01
12800	3600	400	2.7	469	402	23.22	15.16	3.7838e-01	3.7837e-01
15360	4320	480	4.2	473	402	32.65	21.18	4.5567e-01	4.5565e-01
17920	5040	560	6.8	474	402	44.58	28.69	5.3133e-01	5.3131e-01
20480	5760	640	8.8	474	402	57.91	37.02	6.0635e-01	6.0632e-01
23040	6480	720	12.2	476	402	72.78	46.77	6.8363e-01	6.8360e-01
25600	7200	800	16.1	476	402	90.38	58.47	7.5855e-01	7.5853e-01

We first present the framework of pDCA$_\mathrm{e}$, and show that the extrapolation coefficients $\{\beta_t\}$ are general enough to cover those used in FISTA with fixed restart [26] and the proximal DCA [31]. Then we establish the global subsequential convergence of $\{x^t\}$ generated by pDCA$_\mathrm{e}$. Moreover, by assuming the Kurdyka-Łojasiewicz property of an auxiliary function and the differentiability of $P_2(x)$ in (4.1), we establish global convergence of pDCA$_\mathrm{e}$. In addition, we analyze the convergence rate of $\{x^t\}$. Finally, we perform numerical experiments to illustrate our theoretical results. The numerical results show that our algorithm usually outperforms the proximal DCA with nonmonotone linesearch for two classes of DC regularized least squares problems.

Chapter 5

Conclusions of the thesis and future work

In this chapter, we conclude the contents of this thesis, and point out some possible work which we will do in the future.

5.1 Conclusions of the thesis

Nonconvex nonsmooth optimization problems have always been hot issues in many fields. Recently, as the era of big data is coming, these problems play more and more important roles in a lot of application areas. Motivated by this, this thesis focuses on the proximal algorithms with extrapolation for solving nonconvex nonsmooth optimization problems. We first study the proximal gradient algorithm with extrapolation for the minimization of the sum of a Lipschitz differentiable function and a proper closed convex function. And then we propose a proximal difference-of-convex algorithm with extrapolation for minimizing the sum of a Lipschitz differentiable convex function, a proper closed convex function and a continuous concave function, which is (4.1).

In this thesis, we mainly consider the convergence behavior of the proximal gradient algorithm with extrapolation for solving (1.1) and the proximal difference-of-convex algorithm with extrapolation for solving (4.1). More precisely, using the error bound condition which was used in [38], we establish the R-linear convergence of the sequence $\{x^t\}$ generated by the proximal gradient algorithm with extrapolation under the assumption that the extrapolation coefficient is chosen below a certain threshold. Moreover, we also establish the R-linear convergence of the objective sequence $\{F(x^t)\}$. In addition, for solving problem (1.1) with a convex f, we show that the threshold of the extrapolation coefficients reduce to 1 and FISTA with fixed restart is a special case of our proximal gradient algorithm with extrapolation. As a consequence, we conclude R-linear convergence of the iterates generated by FISTA with fixed restart for solving (1.1) with a convex f, when the objective satisfies the error bound condition. Again for convex problems, we show that the successive changes of iterates $\{\|x^{t+1} - x^t\|\}$ go to 0 for many choices of the extrapolation coefficients that approach 1 which cover FISTA.

For the difference-of-convex model (4.1), we first present our proposed algorithm, and show that the extrapolation coefficients in our algorithm can cover the extrapolation coefficients used in FISTA with fixed restart and proximal difference-of-convex algorithm. Then we establish the global subsequential convergence of the sequence $\{x^t\}$ generated by the proximal difference-of-convex algorithm with extrapolation without any additional assumption on the objective function. We also establish the global convergence of $\{x^t\}$ generated by the proximal difference-of-convex algorithm with extrapolation under additional assumptions that P_2 in (4.1) is differentiable and an auxiliary function is a Kurdyka-Łojasiewicz function. These assumptions can be satisfied by plenty of functions. Moreover, we analyze the convergence rate of $\{x^t\}$, which depends on the Kurdyka-Łojasiewicz exponent of an auxiliary function.

Finally, some numerical experiments are performed to illustrate our theoretical results and the efficiency of proximal algorithms with extrapolation.

5.2 Future work

We mainly investigate the convergence behavior of proximal algorithms with extrapolation for nonconvex optimization problems in this thesis. Although we establish some convergence results, there are also some work we can do in the future.

When considering proximal gradient algorithm with extrapolation for solving (1.1), from the framework of our algorithm, we see that proximal grdadient algorithm is a special case of our algorithm. Hence, the linear convergence results we established in this thesis extend corresponding results obtained for the proximal gradient algorithm [37, 38, 39]. However, we have to point out that the local convergence rates of the sequences $\{x^t\}$ and $\{F(x^t)\}$ generated by FISTA for solving (1.1) with a convex f are still unknown, even under the error bound condition. In addition, while the global convergence rates in terms of objective values when applying FISTA and the proximal gradient algorithm to solving (1.1) with a convex f are both known (resp., $O(1/t^2)$ and $O(1/t)$), such a rate is still unknown for FISTA with restart. These are interesting questions for future research.

For the difference-of-convex problems (4.1), we proposed the proximal difference-of-convex algorithm with extrapolation for solving them. Our analysis is based on an assumption on the extrapolation coefficients, i.e., $\sup\limits_t \beta_t < 1$. When $\sup\limits_t \beta_t = 1$, whether the sequence $\{x^t\}$ converges or not is still unknown, not even for subsequential convergence. When we analyze the global convergence of the sequence $\{x^t\}$, we add an assumption that P_2 is a differentiable function, then we establish the global convergence results. However, for many difference-of-convex functions, P_2 may be not differentiable, what is the convergence behavior in this case? These questions are possible future research directions.

Bibliography

[1] M. Ahn, J.S. Pang, and J. Xin. Difference-of-convex learning I: directional stationarity, optimality, and sparsity. *Manuscript, University of Southern California, Los Angeles, California, U. S. A.*, 2016.

[2] A. Alvarado, G. Scutari, and J.S. Pang. A new decomposition method for multiuser DC-programming and its applications. *IEEE Transactions on Signal Processing*, 62: 2984–2998, 2014.

[3] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions invoving analytic features. *Mathematical Programming Series B*, 116: 5–16, 2009.

[4] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35: 438–457, 2010.

[5] H. Attouch, J. Bolte, and B.F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Progamming, Series A*, 137: 91–129, 2013.

[6] H. Attouch and Z. Chbani. Fast inertial dynamics and FISTA algorithms in convex optimization. *arXiv preprint arXiv*:1507.01367v1, 2015.

[7] S. Banert and R.I. Boţ. A general double-proximal gradient algorithm for d.c. programming. *arXiv preprint arXiv*:1610.06538v1, 2016.

[8] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2: 183–202, 2009.

[9] A. Beck and M. Teboulle. A linearly convergent dual-based gradient projection algorithm for quadratically constrained convex minimization. *Mathematics of Operations Research*, 31: 398–417, 2006.

[10] S. Becker, J. Bobin, and E.J. Candès. NESTA: a fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4: 1–39, 2011.

[11] S. Becker, E.J. Candès, and M.C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3: 165–218, 2011.

[12] W. Bian and X. Chen. Optimality and complexity for constrained optimization problems with nonconvex regularization. *To appear in Mathematics of Operations Research.*

[13] J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17: 1205–1223, 2007.

[14] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Progamming, Series A*, 146: 459–494, 2014.

[15] J.M. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization.* 2nd edition, Springer, 2006.

[16] J.M. Borwein and Q.J. Zhu. *Techniques of Variational Analysis*. Springer, 2005.

[17] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9: 717–772, 2009.

[18] E.J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51: 4203–4215, 2005.

[19] E.J. Candès, M. Wakin, and S. Boyd. Enhancing spasity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14: 877–905, 2008.

[20] A. Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20: 89–97, 2004.

[21] A. Chambolle and Ch. Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization Theory and Applications*, 166: 968–982, 2015.

[22] X. Chen, Z. Lu, and T.K. Pong. Penalty methods for a class of non-Lipschitz optimization problems. *SIAM Journal on Optimization*, 26: 1465–1492, 2016.

[23] X. Chen, J. Peng, and S. Zhang. Sparse solutions to random standard quadratic optimization problems. *Mathematical programming, Series A*, 141: 273–293, 2013.

[24] F.H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, Philadelphia, 1990.

[25] P.L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, 49: 185–212, 2011.

[26] B. O'Donoghue and E.J. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15: 715–732, 2015.

[27] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52: 1289–1306, 2006.

[28] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96: 1348–1360, 2001.

[29] L.E. Gibbons, D.W. Hearn, P.M. Pardalos, and M.V. Ramana. Continuous characterizations of the maximal clique problem. *Mathematics of Operations Research*, 22: 754–768, 1997.

[30] P. Gong, C. Zhang, Z. Lu, J.Z. Huang, and J. Ye. A general iterative shinkage and thresholding algorithm for non-convex regularized optimization problems. *ICML*, 2013.

[31] J. Gotoh, A. Takeda, and K. Tono. DC formulations and algorithms for s-parse optimization problems. Preprint, *METR 2015-27*, Department of Mathematical Informatics, University of Tokyo. Available at http://www.keisu.t.u-tokyo.ac.jp/research/techrep/index.html.

[32] T. Ibaraki and N. Katoh. *Resource Allocation Problems: Algorithmic Approaches.* MIT Press, Cambridge, 1988.

[33] P.R. Johnstone and P. Moulin. Local and global convergence of an inertial version of forward-backward splitting. *arXiv preprint arXiv*:1502.02281v4, 2015.

[34] G. Li and T.K. Pong. Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. *arXiv preprint arXiv*:1602.02915v2, 2016.

[35] P.L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16: 964–979, 1979.

[36] T. Liu and T.K. Pong. Further properties of the forward-backward envelope with applications to difference-of-convex programming. *To appear in Computational Optimization and Applications*, DOI: 10.1007/s10589-017-9900-2.

[37] Z.-Q. Luo and P. Tseng. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30: 408–425, 1992.

[38] Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46: 157–178, 1993.

[39] Z.-Q. Luo and P. Tseng. On the convergence rate of dual ascent methods for linearly constrained convex minimization. *Mathematics of Operations Research*, 18: 846–867, 1993.

[40] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7: 77–91, 1952.

[41] J.E. Maingé and A. Moudafi. On the convergence of an approximate proximal method for DC functions. *Journal of Computational Mathematics*, 24: 475–480, 2006.

[42] J.J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93: 273–299, 1965.

[43] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(\frac{1}{k^2})$. *Soviet Mathematics Doklady*, 27: 372–376, 1983.

[44] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Kluwer Academic Publishers, Boston, 2004.

[45] Y. Nesterov. Gradient methods for minimizing composite objective function. *CORE Discussion Paper*, 2007.

[46] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical programming, Series B*, 109: 319–344, 2007.

[47] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming, Series A*, 103: 127–152, 2005.

[48] P. Ochs, Y. Chen, T. Brox, and T. Pock. iPiano: Inertial Proximal Algorithm for Nonconvex Optimization *SIAM Journal on Imaging Sciences*, 7: 1388–1419, 2014.

[49] N. Parikh and S. Boyd. Proximal Algorithms. *Foundations and Trends in Optimization*, 1: 123–231, 2013.

[50] D.T. Pham and H.A. Le Thi. Convex analysis approach to DC programming: theory, algorithm and applications. *Acta Mathematica Vietnamica*, 22: 289–355, 1997.

[51] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4: 1–17, 1964.

[52] R.T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, 1970.

[53] R.T. Rockafellar and R.J.-B. Wets. *Variational Analysis.* Springer, 1998.

[54] M. Sanjabi, M. Razaviyayn, and Z.-Q. Luo. Optimal joint base station assignment and beamforming for heterogeneous networks. *IEEE Transactions on Signal Processing*, 62: 1950–1961,2014.

[55] S. Tao, D. Boley, and S. Zhang. Local linear convergence of ISTA and FISTA on the LASSO problem. *SIAM Journal on Optimization*, 26: 313-336, 2016.

[56] H.A. Le Thi and D.T. Pham. The DC programming and DCA revised with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133: 25–46, 2005.

[57] H.A. Le Thi, D.T. Pham, and V.N. Huynh. Exact penalty and error bounds in DC Programming. *Journal of Global Optimization*, 52: 509–535, 2012.

[58] H.A. Le Thi, D.T. Pham, and D.M. Le. Exact penalty in D.C. programming. *Vietnam Journal of Mathematics*, 27: 169–178, 1999.

[59] P. Tseng and S. Yun. A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. *Computational Optimization and Applications*, 47: 179–206, 2010.

[60] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical programming, Series B*, 117: 387-423, 2009.

[61] P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming, Series B*, 125: 263-295, 2010.

[62] H. Tuy. *Convex Analysis and Global Optimization, Second Edition.* Springer, 2016.

[63] S.J. Wright, R. Nowak, and M.A.T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57: 2479–2493, 2009.

[64] P. Yin, Y. Lou, Q. He, and J. Xin. Minimization of ℓ_{1-2} for compressed sensing. *SIAM Journal on Scientific Computing*, 37: A536–A563, 2015.

[65] C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38: 894–942, 2010.

[66] S. Zhang and J. Xin. Minimization of transformed L_1 penalty: theory, difference of convex function algorithm, and robust application in compressed sensing. *arXiv preprint arXiv*:1411.5735v3, 2014.

[67] Z. Zhou and A.M.-C. So. A unified approach to error bounds for structured convex optimization problems. *To appear in Mathematical Programming.*