ADAPTATION OF EXTRAPOLATION TECHNIQUES IN THE PROXIMAL GRADIENT ALGORITHM AND AN ITERATIVELY REWEIGHTED l_1 ALGORITHM

PEIRAN YU

M.Phil

The Hong Kong Polytechnic University

2018

The Hong Kong Polytechnic University

Department of Applied Mathematics

Adaptation of extrapolation techniques in the proximal gradient algorithm and an iteratively reweighted l_1 algorithm

Peiran YU

A thesis submitted in partial fulfilment of the requirements

for the degree of Master of Philosophy

March 2018

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)

YU Peiran (Name of student)

Abstract

Many areas like engineering, statistics and computing et.al involve solving optimization models. The proximal gradient algorithm is a popular kind of algorithm for solving these problems. This algorithm, though simple and widely used, can be slow in practice; see for example [64, 90, 95]. To accelerate the proximal gradient algorithm, various approaches such as extrapolation techniques or line-search have been adopted. In this thesis, we first present some existing convergence results of the proximal gradient algorithm such as the global complexity. Then, we focus on the adaptation of the extrapolation techniques which is usually easier to implement in practice and leads to provably better iteration complexity for a large class of convex problems. We review this latter fact concerning iteration complexity and also present some existing convergence properties of the proximal gradient algorithm with extrapolation.

Another popular kind of first-order algorithms is the iteratively reweighted algorithm. For this class of algorithms, the line-search techniques have also been adopted for acceleration while the extrapolation techniques have not. In view of the success in accelerating the proximal gradient algorithm empirically via extrapolations, in the last section of this thesis, we investigate how extrapolation techniques can be suitably incorporated into an iteratively reweighted ℓ_1 algorithm. We specifically consider extrapolation techniques motivated from three popular optimal first-order methods: the fast iterative soft-thresholding algorithm (FISTA) [10,68], the method

by Auslender and Teboulle [6] and the method by Lan, Lu and Monteiro [53]. For each algorithm, we exhibit an explicitly checkable condition on the extrapolation parameters so that the sequence generated provably clusters at a stationary point of the optimization problem. We also investigate global convergence under additional Kurdyka-Łojasiewicz assumptions on certain potential functions. Our numerical experiments show that our algorithms usually outperform the general iterative shrinkage and thresholding algorithm in [46] and an adaptation of the iteratively reweighted ℓ_1 algorithm in [56, Algorithm 7] with nonmonotone line-search for solving random instances of log penalty regularized least squares problems in terms of both CPU time and solution quality. The results concerning extrapolation techniques applied to iteratively reweighted algorithm are based on the manuscript [101] available on ArXiv.

Acknowledgements

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

First and foremost, I wish to express my deep thanks to my supervisor Dr. Ting Kei Pong for his enlightening guidance, invaluable discussions and insightful ideas throughout the years.

Furthermore, at the forefront of my M.Phil experience has been the guidance and kindness of my co-supervisor Prof. Xiaojun Chen who has been a constant source of inspiration and mentorship.

Finally, I would like to express my special thanks to my parents and my friends for their love, encouragement and support.

Contents

List of Notations

$\mathbb{R}$	the set of real numbers
$\mathbb{R}_+$	the set of nonnegative real numbers including 0
$\mathbb{R}_{++}$	the set of positive real numbers
$\bar{\mathbb{R}}$	$\mathbb{R} \bigcup \{\infty\}$
$\mathbb{N}$	the set of natural numbers
$\mathbb{R}^n$	the n-dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$
$\mathbb{R}^{m \times n}$	the set of $m \times n$ real matrices
x^T	the transpose of a vector x
$\arg\min f$	the minimizer of f that has a unique minimizer
$\operatorname{Arg\,min} f$	the set of minimizers of f
ri V	the relative interior of a convex set V
$\|x\|_1$	the ℓ_1 norm of a vector x
$\|x\|$	the ℓ_2 norm of a vector x
$\|x\|_\infty$	the ℓ_∞ norm of a vector x
$a \circ b$	the Hadamard (entrywise) product of $a, b \in \mathbb{R}^n$
$a \circ C$	the set $\{a \circ s \ : \ s \in C\}$ for a vector $a \in \mathbb{R}^n$ and a set $C \subseteq \mathbb{R}^n$
$\lvert x \rvert$	the vector whose ith entry is $\lvert x_i \rvert$.

Chapter 1

Introduction

1.1 Models and applications

1.1.1 Compressed sensing

As we are in the midst of digitization, data in various fields such as engineering, statistics and commerce has become so huge that it challenges not only the capacity of devices which sense it but also the speed and capability to compute it. Thus, people begin to transform data into simplified representations which capture the most important information. Usually, the simplified signal is a sparse representation, where the "sparse" here means a data set that has a high percentage of zero. For example, pictures that have small amount of sharp edges have this sparsity property under a proper multiscale wavelet transform; see [61]. Taking advantage of the sparsity, we can reduce the size of the database, saving both the time of storage and computation. Sparse signals are "compressed" and then transmitted. The process of recovering the original sparse signal is called *compressed sensing*. This technique has been widely used in signal processing, image processing, error correcting, parameter estimation and so on; see [7,8,22,24,25,59,60,82,87]. In 2006, E. Candes, J. Romberg and T. Tao proposed that the sparse database can be exactly recovered under suitable conditions, see [23]. Their main approach is to solve a convex optimization problem. More precisely, suppose there is a signal $b \in \mathbb{R}^m$ which comes from a sparse signal

$x \in \mathbb{R}^n$ under the relation $Ax = b$, where $m \ll n$ and $A \in \mathbb{R}^{m\times n}$ is the sensing matrix (or the measurement matrix). E. Candes, J. Romberg and T. Tao [23] proved that, under suitable conditions on A such as the restricted isometry property, one can recover the original sparse signal by solving

$$\min_{\{x:\ Ax=b\}} \|x\|_0,$$

where $\|\cdot\|_0$ represents the number of nonzero coefficients of x. However, this problem is NP-hard in general; see [42]. Thus, they considered a convex relaxation of it,

$$\min_{\{x:\ Ax=b\}} \|x\|_1, \tag{1.1}$$

which gurantees an exact recovery under suitable assumptions; see [23, Theorem 1.3]. This model is convex and thus multiple schemes can be used for solving it such as interior-point methods [15, 18, 28, 39, 52, 91, 100]. Intuitively, $\|\cdot\|_1$ induces sparsity by its nonsmoothness. Figure 1.1 shows that by adjusting the ℓ_1 norm of (x_1, x_2),

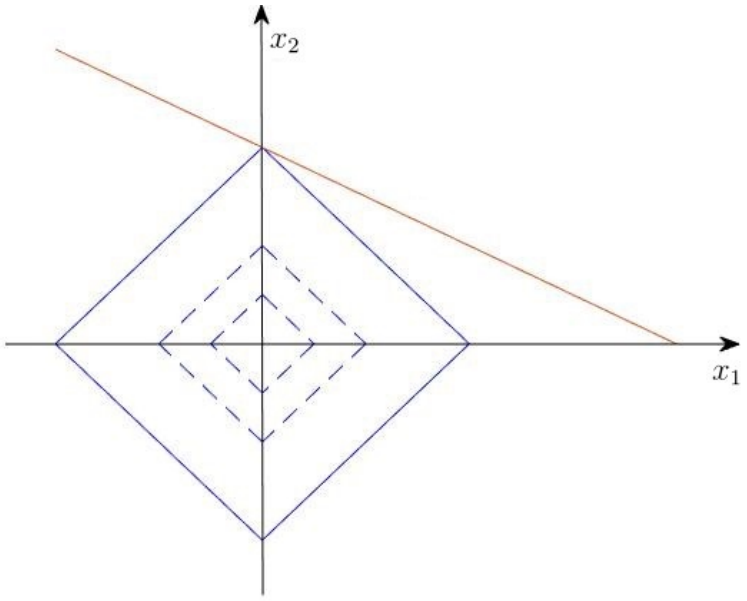

Figure 1.1: This picture shows how l_1 norm induces sparsity. The original solution (x_1, x_2) obeys the relation that $x_1 + x_2 = 1$ and the blue one is $\|(x_1, x_2)\|_1 = 1$.

we can find a sparse solution that satisfies $x_1 + x_2 = 1$.

In many cases, the probed signal b is impacted by noise. Thus, we need to consider the following noisy version of the compressed sensing problem:

$$\min_{\{x:\ \|Ax-b\|\le\delta\}} \|x\|_0, \tag{1.2}$$

where $\delta \in \mathbb{R}^+$ is related to the level of the noise. Still, due to the nonsmooth nonconvex structure of $\|\cdot\|_0$, this model remains difficult to analyze and, as proved in [63, Theorem 1], this problem is still NP-hard. Thus, we turn to solving the $\|\cdot\|_1$ relaxation:

$$\min_{\{x:\ \|Ax-b\|\le\delta\}} \|x\|_1. \tag{1.3}$$

This is a convex optimization problem and can be solved by various schemes [15, 18, 28, 30, 39, 52, 91, 100]. Moreover, if $\delta > 0$ and A has full row rank, then the above problem is equivalent to the following unconstrained optimization problem for some $\lambda > 0$:

$$\min_x \|Ax - b\|_2^2 + \lambda\|x\|_1, \tag{1.4}$$

if the maximizer of

$$\max_{\mu}\min_{x} L(x;\mu) := \|x\|_1 + \mu(\|Ax - b\|^2 - \delta)$$

is strictly positive for some $\mu > 0$.

Intuitively, in (1.4), when λ decreases, the residual $\|Ax - b\|^2$ decreases. Thus, λ controls the residual; see [28]. Many efficient algorithms have been developed in the literature [54, 68, 72, 80, 90] to solve this model and the convergence properties of some of these algorithms will be reviewed in subsequent literature.

Compressed sensing has also been extended to the matrix case to recover "low rank" matrices. Low rank matrix completion is to reconstruct the whole (low rank) matrix from partial entries. This problem can be modeled as an optimization problem of finding the minimum rank or nuclear norm (as a convex relaxation) of some

matrix; see [75]. The matrix completion is motivated from the Netflix problem, which estimates someone's ratings for contemporary movies based on previous rating records; see [20, 55, 60, 75, 76, 86] for a concrete description and the analysis of respective algorithms.

1.1.2 Sparsity in statistical inferences

In the statistical field, the estimation of parameters of a linear system involves solving the following model known as LASSO, which stands for "least absolute shrinkage and selection operator":

$$\min_{\|x\|_1 \leqslant \delta} \|Ax - b\|^2, \tag{1.5}$$

where $\delta > 0$ represents a tuning parameter that indicates sparsity and $x_j \in \mathbb{R}$ for $j = 1, \ldots, n$ is one predictor variable while $b_i \in \mathbb{R}$ for $i = 1, \ldots, m$ are representative responses. LASSO surpasses the ordinary least squares estimation in both prediction accuracy and interpretation by shrinking some coefficients and setting others to 0; see [84]. One way to solve LASSO is to solve (1.4) for some suitable $\lambda > 0$ instead. LASSO and its variants are widely used for variable selection; see [84, 85, 102, 106]. For example, when partitioning $\{1, \ldots, N\}$ into p groups denoted as $K_1, \ldots, K_p$, we get the grouped LASSO

$$\min_{\beta \in \mathbb{R}^N} \left\{ \|y - \sum_{j=1}^{p} X_j \beta_j\|^2 + \lambda \sum_{j=1}^{p} \|\beta_j\|_{R_j} \right\},$$

where X_j with $j \in \{1, 2, \ldots, p\}$ is a submatrix of a given matrix $X \in \mathbb{R}^{M \times N}$ with columns corresponding to the predictors in group K_j; β_j stands for the coordinates that are indexed by K_j and $\|z\|_{R_j} = (z^T R_j z)^{\frac{1}{2}}$ with a series of symmetric $d \times d$ positive definite matrices $\{R_j\}_{j=1}^{p}$. This grouped LASSO enables variables that have correlations to be selected or ruled out together; see [41, 48, 81, 102].

In fact, the key motivation to introduce LASSO is the $\|\cdot\|_1$ part, which induces sparsity. This idea can also be adopted in other regression models. For example, one variant of the LASSO is the logistic loss model, which gives:

$$\min_{\alpha\in\mathbb{R}^p} \ell_n(\alpha) + \lambda\|\alpha\|_1, \tag{1.6}$$

where $\ell_n(\alpha)$ is a negative log-likelihood function

$$\ell_n(\alpha) = -\sum_{i=1}^{n}\left\{Y_i \log\frac{e^{\langle x_i,\alpha\rangle}}{1+e^{\langle x_i,\alpha\rangle}} + (1-Y_i)\log\left[1-\frac{e^{\langle x_i,\alpha\rangle}}{1+e^{\langle x_i,\alpha\rangle}}\right]\right\},$$

where $Y \in \{0,1\}^n$ and $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})^T$ for each $i = 1, 2, \ldots, n$. This optimization model can be further applied to the detection of the splice sites in DNA, the estimation of German credit data and breast cancer data; see [47, 51, 62] for concrete models, analysis of consistency and relevant algorithms to solve (1.6).

However, in some cases, LASSO may be biased when applied to high dimensional variable selection. One way to deal with the bias of LASSO model is to replace the ℓ_1 in (1.4) with the smoothly clipped absolute deviation (SCAD) penalty function proposed by J. Fan and R. Li; see [38]. The SCAD function $P(x)$ is given by

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

where $\theta > 2$ and $\lambda > 0$. This function can be rewritten in the form $P(x) = \sum_{i=1}^{n}\phi(|x_i|)$, where

$$\phi(t) = \begin{cases} \lambda t, & t \leqslant \lambda, \\ \frac{-t^2+2\theta\lambda t-\lambda^2}{2(\theta-1)}, & \lambda < t \leqslant \theta\lambda, \\ \frac{(\theta+1)\lambda^2}{2}, & t > \theta\lambda. \end{cases}$$

According to [38], this penalty function possesses three good properties simultaneously: continuity, sparsity, and unbiasedness, which the ℓ_1 penalty function does

not. The SCAD penalty function has been widely used in high dimensional problems; see [38, 50, 94]. In [50], an algorithm is proposed to solve (1.4) with ℓ_1 replaced by SCAD.

Another way to tackle the bias is to solve (1.4) with ℓ_1 being replaced by the a minimax concave penalty (MCP) function proposed by C. Zhang; see [104]. The MCP function, which is denoted by $P(x)$, is defined as follow:

$$P(x) = \lambda \sum_{i=1}^{n} \int_0^{|x_i|} [1 - \frac{t}{\theta\lambda}]_+ dt,$$

where θ and λ are some fixed strict positive real numbers. This function can be rewritten in the form $P(x) = \sum_{i=1}^{n} \phi(|x_i|)$, where

$$\phi(t) = \begin{cases} \lambda t - \frac{t^2}{2\theta\lambda}, & t \leqslant \theta\lambda \\ \frac{\theta\lambda^2}{2}, & t > \theta\lambda. \end{cases}$$

In [104], C. Zhang showed that the MCP function has a high probability of getting correct selection without some conditions required by the LASSO in [105].

Both the MCP and the SCAD penalty functions are nonconvex; see Figure 1.2. In addition, they can be rewritten as the difference of two convex (DC) functions and P. Gong et al. proposed a general iterative shrinkage and thresholding algorithm to solve these problems and analyzed the convergence properties of their algorithm in [45]. One can also apply the proximal difference-of-convex algorithm with extrapolation proposed in [96] to (1.4) with ℓ_1 being replaced by either the MCP function or the SCAD function. In [96], the sequence generated by their algorithm was proved to have global subsequential convergence under mild conditions.

Alternatively, one may also use other penalty functions to induce sparsity instead of the l_1 norm in (1.4). One such example is the l_p quasi-norm, where $p \in (0, 1)$:

$$\min \|Ax - b\|_2^2 + \lambda \|x\|_p^p. \tag{1.7}$$

Intuitively, when $0 < p < 1$, due to its "sharpness" around the origin, $\|\cdot\|_p$ may be more effective than $\|\cdot\|_1$ for inducing sparsity; see Figure 1.2.

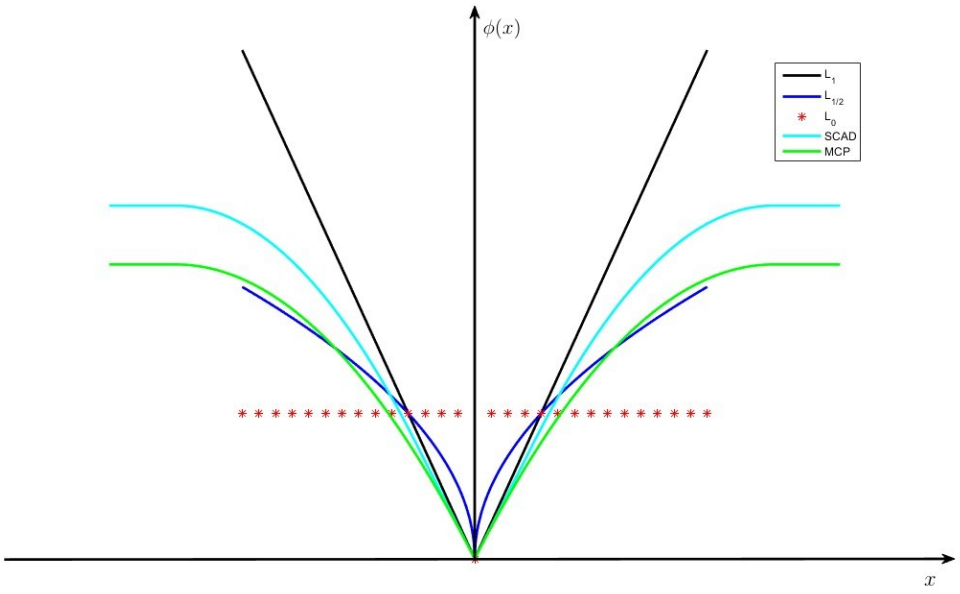

Figure 1.2: the l_1 regularization function intuitively induces sparsity. The black one is $\phi(x) = |x|$; the deep blue one is $\phi(x) = |x|^{\frac{1}{2}}$; the light blue one is the SCAD penalty function while the green one is the MCP function; the red one corresponds to the l_0 norm.

The problem nonconvex programming problem (1.7) is NP-hard when $0 < p < 1$; see [31]. However, algorithms have been developed to find *stationary points* of this problem and the convergence properties are also investigated in [26, 32, 41, 48, 81, 84, 104].

1.2 Aim and structure of this thesis

Motivated by the aforementioned applications, in this thesis, we discuss algorithms for solving those models. In particular, we focus on solving (1.4) and its variants with ℓ_1 being replaced by MCP, SCAD and other regularizers. These models can be solved by popular first-order methods such as the proximal gradient method (PG), the iteratively reweighted algorithms and their variants; see [12,30,54,68,72,80,90,95]. More precisely, the proximal gradient method is designed to solve the following class of problems:

$$\min \ F(x) := f(x) + P(x), \tag{1.8}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ has Lipschitz continuous gradient with modulus $L > 0$ and $P : \mathbb{R}^n \to (-\infty, \infty]$ is proper closed. The models (1.4) and the Lagrangian function of (1.5) are in this form.

We will discuss some existing convergence properties of the sequence generated by the proximal gradient method in both convex and nonconvex cases. In the convex case, i.e. when f and P are convex, we will show under some suitable assumptions on the stepsize that the sequence of function values generated by PG decreases to the optimal function value at a rate of $O(\frac{1}{k})$, in which case we say PG has a global complexity of $O(\frac{1}{k})$. In the nonconvex case, i.e. when either f or P is not convex, we will prove under another assumption on the stepsize that PG is a descent algorithm and the sequence generated by this algorithm accumulates at a stationary point of problem (1.8).

In some applications, the proximal gradient method can be slow; see for example [5, 10, 64–66, 68]. Therefore, we will discuss some accelerated variants of PG. In [10,68,89], extrapolation techniques such as Nesterov's acceleration schemes were incorporated into PG for solving (1.8) when f and P are convex, resulting in a global complexity of $O\left(\frac{1}{k^2}\right)$. In [70, 96], PG with Nesterov's extrapolation techniques [64]

coupled with a restart scheme was shown to exhibit fast convergence in their numerical tests. In this thesis, we will discuss how in [89] PG with Nesterov's extrapolation techniques has the complexity of $O(\frac{1}{k^2})$ when applied to solving convex problems, following the discussions in [90]. We will also discuss some convergence results of PG with suitable extrapolation schemes when applied to solving nonconvex problem, following the discussions in [95].

We will then discuss another kind of first-order algorithms, the iteratively reweighted algorithms which include the iteratively reweighted ℓ_2 algorithm and the iteratively reweighted ℓ_1 algorithm. The iteratively reweighted ℓ_2 algorithm is introduced first and then we focus on the iteratively reweighted ℓ_1 algorithm which is related to our proposed algorithms in later part of this thesis. The iteratively reweighted ℓ_1 algorithm was originally designed for the following problem

$$\min_{Ax=b} P(x)$$

with

$$P(x) = \sum_{i=1}^{n} \phi(x_i), \tag{1.9}$$

where $\phi : \mathbb{R} \to \mathbb{R}$. This model includes model (1.1) and its variants with ℓ_1 in (1.1) replaced by ℓ_p, the logistic function and other regularizers; see [26, 27, 33]. More precisely, the iteratively reweighted ℓ_1 algorithm solves a subproblem that has the following form in each iteration:

$$x^{k+1} = \min_{\{x:\ Ax=b\}} \sum_{i=1}^{n} w_i^k |x_i|,$$

where $\{w_i^k\} \subseteq \mathbb{R}_+$ are called the weights. It is easy to see that when $w_i^k = 1$ for all $i = 1, 2, \ldots, n$, the subproblem above reduces to (1.1). Intuitively, by properly choosing the weights, the weighted l_1 may be more efficient in inducing sparsity than l_1; see [26, Figure. 1].

After discussing some popular first-order methods such as PG, its extrapolated variants, and the iteratively reweighted algorithms, we will investigate how extrapolation techniques can be suitably incorporated into the iteratively reweighted ℓ_1 algorithm for solving special instances of (1.8). These results are joint work with Ting Kei Pong, my supervisor. Assuming P in (1.8) has the form (1.9), under suitable assumptions on f and ϕ, we specifically consider extrapolation techniques motivated from three popular optimal first-order methods: the fast iterative soft-thresholding algorithm (FISTA) [10, 68], the method by Auslender and Teboulle [6] and the method by Lan, Lu and Monteiro [53]. We call the corresponding iteratively reweighted ℓ_1 algorithm with extrapolation IRL$_1e_1$, IRL$_1e_2$ and IRL$_1e_3$, respectively. For each algorithm, we show that the sequence generated clusters at a stationary point of (1.8) under certain condition on the extrapolation parameters. These conditions are satisfied by many choices of extrapolation parameters: for instance, one can pick the parameters as in FISTA with fixed restart [70] in IRL$_1e_1$. Furthermore, under some additional assumptions such as the Kurdyka-Łojasiewicz property (see for example, [3, 4]) on some suitable potential functions, we show that the sequences generated by IRL$_1e_1$ and IRL$_1e_3$ are indeed convergent. We then perform numerical experiments comparing our algorithms (with our proposed choices of extrapolation parameters) against the general iterative shrinkage and thresholding algorithm (GIST) [46] and an adaptation of the iteratively reweighted ℓ_1 algorithm [56, Algorithm 7] with nonmonotone line-search (IRL$_1ls$) for solving log penalty regularized least squares problems on random instances. In our experiments, our iteratively reweighted ℓ_1 algorithms with extrapolation usually outperform GIST and IRL$_1ls$ in both CPU time and solution quality. Moreover, IRL$_1e_1$ and IRL$_1e_3$ usually perform better than IRL$_1e_2$.

The rest of the thesis is organized as follows: we discuss necessary preliminaries in Chapter 2. Some existing convergence results concerning PG and PG with extrapo-

lation are discussed in Chapter 3. We also briefly describe the iteratively reweighted algorithms. In Chapter 4, we incorporate extrapolation techniques into the iteratively reweighted ℓ_1 algorithms. We analyze convergence properties of the sequences generated by the resulting algorithms and some numerical results are presented.

Chapter 2

Preliminaries

We call the set $\{x:\ f(x)<\infty\}$ the domain of f and denote it by $\mathrm{dom} f$. We say that a function $f:\mathbb{R}^n\to\bar{\mathbb{R}}$ is proper if $\mathrm{dom} f\neq\emptyset$. Such a function is said to be lower semicontinuous at x if

$$\liminf_{x^i\to x} f(x)\geqslant f(x)$$

for all sequences $x^i\to x$, and is said to be closed if it is lower semicontinuous at every point in $\mathbb{R}^n$.

For a proper function f, the regular subdifferential of f at $x\in\mathrm{dom} f$ is defined and denoted in [79, Definition 8.3] by

$$\widehat{\partial} f(x):=\left\{\zeta:\ \liminf_{z\to x,z\neq x}\frac{f(z)-f(x)-\langle\zeta,z-x\rangle}{\|z-x\|}\geqslant 0\right\}.$$

The subdifferential of f at $x\in\mathrm{dom} f$ (which is also called the limiting subdifferential) is defined and denoted in [79, Definition 8.3] by

$$\partial f(x):=\left\{\zeta:\ \exists x^v\xrightarrow{f}x,\zeta^v\to\zeta\text{ with }\xi^v\in\widehat{\partial} f(x^v)\text{ for each }v\right\},\tag{2.1}$$

where $x^v\xrightarrow{f}x$ means both $x^v\to x$ and $f(x^v)\to f(x)$. We denote $\{x:\ \partial f(x)\neq\emptyset\}=:\mathrm{dom}\,\partial f$. We set $\partial f(x)=\widehat{\partial} f(x)=\emptyset$ for $x\notin\mathrm{dom}\, f$ by convention. Moreover, $\partial f(x)=\{\nabla f(x)\}$ if f is continuously differentiable at x; [79, Exercise 8.8].

If f is proper convex, from the definition of the subdifferential of f at $x \in \mathrm{dom} f$ and the convexity of f, the subdifferential of f degenerates to

$$\{\zeta : \langle \zeta, y - x \rangle \leqslant f(y) - f(x) \text{ for any } y\}; \tag{2.2}$$

see [16, Theorem 6.2.2].

The limiting subdifferential enjoys several basic calculus rules:

Proposition 2.1. *The following statements hold:*

(i) ***(Sum rule)*** *Suppose that $f = f_1 + f_2$ with f_1 Lipschitz continuous at some $\bar{x} \in \mathrm{dom} f$ and f_2 is proper closed. Then*

$$\partial f(\bar{x}) \subseteq \partial f_1(\bar{x}) + \partial f_2(\bar{x}).$$

In addition, if f_1 is continuously differentiable at $\bar{x}$, the above inclusion holds as an equality.

(ii) ***(Separable function)*** *Let $f(x) = \sum_{i=1}^{n} f_i(x_i)$, where each $f_i : \mathbb{R}^{n_i} \to \bar{\mathbb{R}}$ is a proper closed function, $x = (x_1, x_2, \ldots, x_n)$ with $x_i \in \mathbb{R}^{n_i}$. Then at any $x \in \mathrm{dom}\, f$, one has*

$$\partial f(x) = \partial f_1(x_1) \times \partial f_2(x_2) \times \cdots \times \partial f_n(x_n).$$

Proof. The first statement comes from [79, Exercise 10.10] and [79, Exercise 8.8]; the second statement follows from [79, Proposition 10.5]. □

Next we introduce a repeatedly used function in this thesis, the indicator function. The indicator function δ_C of C is defined as

$$\delta_C(x) := \begin{cases} 0 & x \in C, \\ \infty & x \notin C. \end{cases}$$

The normal cone of C at an $x \in C$ is defined as $N_C(x) := \partial\delta_C(x)$, and the distance from a point $x \in \mathbb{R}^n$ to C is denoted by $\mathrm{dist}(x, C)$.

Now we introduce the *proximal mapping* which is widely used in designing first-order methods.

The proximal mapping of a proper closed function $h : \mathbb{R}^n \to \bar{\mathbb{R}}$ is defined as follows:

$$\mathrm{Prox}_h(x) = \operatorname*{Arg\,min}_u \left\{ h(u) + \frac{1}{2}\|u - x\|^2 \right\}.$$

When h is the indicator function of a closed set C, $\mathrm{Prox}_h(x)$ reduces to $\mathrm{Proj}_C(x)$, the set of points in C that are closest to x.

One fact about the proximal mapping in optimization is that x^* is a global minimizer of a convex function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ if and only if it is the fixed point of $\mathrm{Prox}_f(x)$, which can be deduced from the optimality condition [16, Theorem 3.1.5]. Actually, many algorithms can be considered as finding a fixed point of some operators; see [9, 72, 78]. In many applications, $\mathrm{Prox}_h(x)$ usually has a closed form; see [28,34,35]. For example, when solving the models in Chapter 1, many algorithms use the proximal mapping of $h(x) = \lambda\|x\|_1$. Using the optimality condition, we have the following explicit formula for $h(x) = \lambda\|x\|_1$:

$$\mathrm{Prox}_h(x)_i = \begin{cases} x_i - \lambda, & x_i \geqslant \lambda; \\ 0, & |x_i| < \lambda; \\ x_i + \lambda, & x_i \leqslant -\lambda. \end{cases}$$

More closed forms of the proximal mapping of different functions can be found in [9, Table 10.2].

Definition 2.1 (***Strong convexity***). *We say that a proper function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is strongly convex with modulus m if for any $x, y \in \mathrm{dom}\, f$ and $\lambda \in [0, 1]$, the following*

inequality holds:

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda)f(y) - \frac{m}{2}\lambda(1-\lambda)\|x-y\|^2. \tag{2.3}$$

Subproblems of many commonly used optimization algorithms are strongly convex (even when the original problem is not convex), and the strong convexity is also crucial in the convergence analysis of these algorithms; see for examples [44, 90, 96]. Here we list two most useful equivalent definitions of strong convexity. Before that, we present a lemma which will be used in the proof of the equivalence of the definitions of strong convexity.

Lemma 2.1. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be proper, closed and convex, if $x \in \mathrm{ri}(\mathrm{dom} f)$, then for any $d \in \mathrm{span}(\mathrm{dom} f - x)$,*

$$f'(x;d) = \max_{\xi \in \partial f(x)} \langle \xi, d \rangle ,$$

and the maximum is attained, where $f'(x;d) := \lim\limits_{t \downarrow 0} \frac{f(x+td)-f(x)}{t}$ is the directional derivative in the direction d at x and $\mathrm{span} D$ of a set D means the smallest linear subspace containing D.

Proof. Denote $Z := \mathrm{span}(\mathrm{dom} f - x)$ and fix any $d_0 \in Z$.

Setting $\psi(d) := \alpha f'(x; d_0)$, where $d = \alpha d_0$ and $\alpha \in \mathbb{R}$, we have

$$\psi(d) = \begin{cases} \alpha f'(x;d_0) = f'(x;\alpha d_0), & \text{for } \alpha \geqslant 0; \\ \alpha f'(x;d_0) = -(-\alpha) f'(x;d_0) = -f'(x;-\alpha d_0) \leqslant f'(x;\alpha d_0), & \text{for } \alpha < 0, \end{cases} \tag{2.4}$$

where the first relation is due to the positive homogeneity of ψ on $\mathrm{span}\{d_0\}$ that can be easily obtained by the definition of $f'(x;d)$, and the inequality is because $f'(x;\cdot)$ is sublinear on Z by [16, Proposition 3.1.2], thanks to the assumption that $x \in \mathrm{ri}(\mathrm{dom} f)$.

Therefore, we get

$$\psi(d) \leqslant f'(x;d), \forall d \in \operatorname{span}\{d_0\}.$$

Using this together with the sublinearity of $f'(x;\cdot)$ on Z and the linearity of $\psi(\cdot)$ on $\operatorname{span}\{d_0\}$, we can apply the Hahn-Banach Theorem and conclude that there exists a vector ξ^* such that

$$\langle \xi^*, d\rangle \leqslant f'(x;d) \text{ for } d \in Z;\ \langle \xi^*, d\rangle = \psi(d) \text{ for } d \in \operatorname{span}\{d_0\}. \tag{2.5}$$

Thus, for all $d \in Z$,

$$\langle \xi^*, d\rangle \leqslant f'(x;d) \leqslant f(x+d) - f(x), \tag{2.6}$$

where the second inequality is easily a consequence of the convexity of f. The inequality (2.6) implies that $\xi^* \in \partial f(x)$. Using this fact and taking $\alpha = 1$ in the first equation in (2.4) and using (2.5), we have

$$f'(x;d_0) = \psi(d_0) = \langle \xi^*, d_0\rangle \leqslant \max_{\xi \in \partial f(x)} \langle \xi, d_0\rangle, \tag{2.7}$$

where the second equality is from (2.5).

On the other hand, by definition of the subgradient, it is easy to check that $f'(x;d_0) \geqslant \max_{\xi \in \partial f(x)} \langle \xi, d_0\rangle$. Combining this with (2.7) we have

$$f'(x;d_0) = \max_{\xi \in \partial f(x)} \langle \xi, d_0\rangle$$

and the maximum is attained at ξ^*. Since $d_0 \in Z$ is arbitrary, the proof is completed.

□

Now, we are ready to give some equivalent definitions of strong convexity.

Proposition 2.2. *Suppose that $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is proper closed. Then the following statements are equivalent:*

(i) *f is strongly convex;*

(ii) *There exists an $m > 0$ such that for any points $y \in \operatorname{dom} f$, $x \in \operatorname{dom} \partial f$ and any $\xi \in \partial f(x)$, we have*

$$f(y) \geqslant f(x) + \langle \xi, y - x \rangle + \frac{m}{2}\|y - x\|^2; \tag{2.8}$$

(iii) *There exists an $m > 0$ such that for any points $x \in \operatorname{dom} \partial f$ and $y \in \operatorname{dom} \partial f$ and any $\xi_x \in \partial f(x)$, $\xi_y \in \partial f(y)$, we have*

$$\langle \xi_x - \xi_y, x - y \rangle \geqslant m\|y - x\|^2. \tag{2.9}$$

Proof. First we prove $(i) \Rightarrow (ii)$. Fix any points $y \in \operatorname{dom} f$, $x \in \operatorname{dom} \partial f$ and choose any $\xi \in \partial f(x)$. Choosing a $\lambda \in [0, 1)$, rearranging (2.3), we have

$$\begin{aligned}
\frac{\lambda}{1-\lambda} f(x) + f(y) &\geqslant \frac{f(\lambda x + (1-\lambda)y)}{1-\lambda} + \frac{\lambda m}{2}\|x - y\|^2 \\
\Longrightarrow f(y) &\geqslant \frac{f(\lambda x + (1-\lambda)y) - f(x)}{1-\lambda} + f(x) + \frac{\lambda m}{2}\|x - y\|^2 \\
&\geqslant \frac{\langle \xi, (1-\lambda)(y - x) \rangle}{1-\lambda} + f(x) + \frac{\lambda m}{2}\|x - y\|^2 \\
&= \langle \xi, y - x \rangle + f(x) + \frac{\lambda m}{2}\|x - y\|^2,
\end{aligned}$$

where $\xi \in \partial f(x)$ and the last inequality is due to (2.2). Passing to the limit as $\lambda \uparrow 1$, we obtain (2.8).

For $(ii) \Rightarrow (i)$, we first prove that (2.3) holds for $y \in \operatorname{ri}(\operatorname{dom} f)$, $x \in \operatorname{dom} f, \lambda \in (0, 1)$. Fix any points $y \in \operatorname{ri}(\operatorname{dom} f)$, $x \in \operatorname{dom} f$. For any $\lambda \in (0, 1)$,

$$x_\lambda := \lambda x + (1-\lambda)y = y + \lambda(x - y) \in \operatorname{ri}(\operatorname{dom} f) \subseteq \operatorname{dom} \partial f,$$

where the inclusion is by [77, Theorem 23.4]. Thus, using (2.8), we have

$$f(y) \geqslant f(x_\lambda) + \langle \xi, y - x_\lambda \rangle + \frac{m}{2}\|y - x_\lambda\|^2,$$

$$f(x) \geqslant f(x_\lambda) + \langle \xi, x - x_\lambda \rangle + \frac{m}{2}\|x - x_\lambda\|^2,$$

where $\xi \in \partial f(x_\lambda)$. Summing the first inequality multiplied by $1-\lambda$ and the second multiplied by λ, we see that (2.3) holds for $y \in \mathrm{ri}(\mathrm{dom} f)$, $x \in \mathrm{dom} f, \lambda \in (0,1)$.

For $x, y \in \mathrm{dom} f$, $\lambda \in (0,1)$, taking a $y_0 \in \mathrm{ri}(\mathrm{dom} f)$ and for any $t \in (0,1)$ we have

$$y_t = y + t(y_0 - y) \in \mathrm{ri}(\mathrm{dom}\ f) \subseteq \mathrm{dom}\, \partial f,$$

where the inclusion is by [77, Theorem 23.4]. Thus, using the fact that (2.3) holds for $y_t \in \mathrm{ri}(\mathrm{dom} f)$ for $t \in (0,1)$, $x \in \mathrm{dom} f$, we have

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &\leqslant \liminf_{t \downarrow 0} f(\lambda x + (1-\lambda) y_t) \\ &\leqslant \liminf_{t \downarrow 0} \left(\lambda f(x) + (1-\lambda) f(y_t) - \frac{m}{2}\lambda(1-\lambda)\|x - y_t\|^2 \right) \\ &\leqslant \lambda f(x) + (1-\lambda) \limsup_{t \downarrow 0} f(y_t) - \frac{m}{2}\lambda(1-\lambda)\|x-y\|^2 \\ &\leqslant \lambda f(x) + (1-\lambda) f(y) - \frac{m}{2}\lambda(1-\lambda)\|x-y\|^2, \end{aligned}$$

where the first inequality is because f is lower semicontinuous while the last inequality is because $t \mapsto f(y_t)$ is upper semicontinuous in the closure of its domain by [103, Proposition 2.1.6]. Thus, we have shown that (2.3) holds for $x, y \in \mathrm{dom} f, \lambda \in (0,1)$. Noticing that when $\lambda = 0$ or $\lambda = 1$, (i) holds trivially, we obtain that $(ii) \Rightarrow (i)$.

For $(ii) \Rightarrow (iii)$, using (2.8), for any $x \in \mathrm{dom}\, \partial f$ and $y \in \mathrm{dom}\, \partial f$ and any $\xi_x \in \partial f(x)$, $\xi_y \in \partial f(y)$, we have

$$f(y) \geqslant f(x) + \langle \xi_x, y - x \rangle + \frac{m}{2}\|y - x\|^2;$$

$$f(x) \geqslant f(y) + \langle \xi_y, x - y \rangle + \frac{m}{2}\|y - x\|^2.$$

Summing these two inequalities, we obtain (2.9).

Finally, we prove $(iii) \Rightarrow (ii)$. We first consider the case that $y \in \mathrm{ri}(\mathrm{dom} f)$ and $x \in \mathrm{dom} \partial f$. Supposing that (iii) holds, then for $\xi \in \partial f(x)$, using the Riemann-

Lebesgue theorem (see for example [19]), we have

$$\begin{aligned} f(y) &= f(x) + \int_0^1 f'(x_\lambda; y-x) d\lambda \\ &= f(x) + \langle \xi, y-x \rangle + \int_0^1 f'(x_\lambda; y-x) - \langle \xi, y-x \rangle \, d\lambda, \end{aligned} \tag{2.10}$$

where x_λ denotes $x + \lambda(y-x)$. Since $x \in \mathrm{dom}\partial f \subseteq \mathrm{dom} f$, $y \in \mathrm{ri}(\mathrm{dom} f)$ and $\lambda \in (0,1)$ implies,

$$x_\lambda = x + \lambda(y-x) = (1-\lambda)x + \lambda y \in \mathrm{ri}(\mathrm{dom} f).$$

This property enables us to apply Lemma 2.1 at the point x_λ at any $\lambda \in (0,1)$ to obtain

$$\begin{aligned} & f'(x_\lambda; y-x) - \langle \xi, y-x \rangle = \max_{\xi_\lambda \in \partial f(x_\lambda)} \langle \xi_\lambda, y-x \rangle - \langle \xi, y-x \rangle \\ & = \langle \xi_\lambda^*, y-x \rangle - \langle \xi, y-x \rangle = \langle \xi_\lambda^* - \xi, y-x \rangle \\ & = \frac{1}{\lambda} \langle \xi_\lambda^* - \xi, x_\lambda - x \rangle \geqslant \frac{1}{\lambda} m \|x_\lambda - x\|^2, \end{aligned} \tag{2.11}$$

where we denote a maximizer that attains $\max_{\xi_\lambda \in \partial f(x_\lambda)} \langle \xi_\lambda, y-x \rangle$ by ξ_λ^*, which exists due to Lemma 2.1. Applying the inequality (2.11) to the right hand side of (2.10), we further have

$$\begin{aligned} f(y) &\geqslant f(x) + \langle \xi, y-x \rangle + \int_0^1 \frac{1}{\lambda} m \|x_\lambda - x\|^2 d\lambda \\ &= f(x) + \langle \xi, y-x \rangle + \int_0^1 \lambda m \|x-y\|^2 d\lambda \\ &= f(x) + \langle \xi, y-x \rangle + \frac{m}{2} \|x-y\|^2. \end{aligned}$$

Thus, for $y \in \mathrm{ri}(\mathrm{dom} f), x \in \mathrm{dom}\partial f$, $(iii) \Rightarrow (ii)$ holds.

For $y \in \mathrm{dom} f, x \in \mathrm{dom}\partial f$, taking a $y_0 \in \mathrm{ri}(\mathrm{dom} f)$ and for any $t \in (0,1)$ we have

$$y_t = y + t(y_0 - y) \in \mathrm{ri}(\mathrm{dom} f).$$

Thus, with the fact that (2.8) holds for $y_t \in \mathrm{ri}(\mathrm{dom} f), x \in \mathrm{dom}\partial f$, we have

$$\begin{aligned} f(y) &\geqslant \limsup_{t \downarrow 0} f(y_t) \geqslant f(x) + \lim_{t \downarrow 0} \left[\langle \xi, y_t - x \rangle + \frac{m}{2} \|y_t - x\|^2 \right] \\ &= f(x) + \langle \xi, y - x \rangle + \frac{m}{2} \|y - x\|^2 \end{aligned}$$

where the first inequality is because $t \mapsto f(y_t)$ is upper semicontinuous on the closure of its domain by [103, Proposition 2.1.6]. Thus, for $y \in \mathrm{dom} f, x \in \mathrm{dom}\partial f$, (2.8) holds. □

Corollary 2.1. *If a proper closed function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is strongly convex, then the set of minimizers is nonempty. Moreover, if x^* is a minimizer of f, then for any $y \in \mathbb{R}^n$*

$$f(y) \geqslant f(x^*) + \frac{m}{2} \|y - x^*\|^2 \tag{2.12}$$

for some $m > 0$. In this case, the set of minimizers of such f is a singleton set.

Proof. We first show that the set of minimizers is nonempty. To this end, let $x^0 \in \mathrm{ri}\,\mathrm{dom} f$ and recall that $\partial f(x^0) \neq \emptyset$. Let $\xi^0 \in \partial f(x^0)$. Then by (2.8), we have for any $x \in \{x \ : \ f(x) \leqslant f(x^0)\}$,

$$\begin{aligned} 0 \geqslant f(x) - f(x^0) &\geqslant \left\langle \xi^0, x - x^0 \right\rangle + \frac{m}{2} \|x^0 - x\|^2 \\ &\geqslant -\|\xi^0\| \|x - x^0\| + \frac{m}{2} \|x^0 - x\|^2, \end{aligned}$$

i.e., we have

$$\|x - x^0\| \leqslant \frac{2}{m} \|\xi^0\|.$$

Therefore, the set $\{x \ : \ f(x) \leqslant f(x^0)\}$ is bounded.

If $f(x^0) = \inf_x f(x)$, then we conclude immediately that the set of minimizers is nonempty because it contains x^0. Otherwise, choose a sufficiently large positive integer $n_0 \geqslant 1$ so that $\inf_x f(x) + \frac{1}{n_0} < f(x^0)$. Then there exists a sequence $\{x^n\}$ so that for any $n \geqslant n_0$, we have

$$f(x^n) \leqslant \inf_x f(x) + \frac{1}{n} < f(x^0).$$

Then $\{x^n\}$ is bounded. Let $\{x^{n_j}\}$ be a convergent subsequence of $\{x^n\}$ and let $\lim_{j\to\infty} x^{n_j} = x^*$. Then, passing to the limit in the inequality above along the subsequence $\{x^{n_j}\}$ and using the lower semicontinuity of f, we have

$$\inf_x f(x) \leqslant f(x^*) \leqslant \liminf_{j\to\infty} f(x^{n_j}) \leqslant \liminf_{j\to\infty} \left[\inf_x f(x) + \frac{1}{n_j}\right] = \inf_x f(x),$$

where the first inequality follows from the definition of infimum. Thus, we conclude that x^* is a minimizer of f and hence the set of minimizers of f is nonempty.

We now prove (2.12). Following the definition of subgradient and the fact that that x^* is the minimizer, we have $0 \in \partial f(x^*)$. Setting $\xi = 0$ in (2.8), we obtain (2.12). Suppose there are two different minimizers of f, and we denote them as x_1 and x_2. Then using (2.12) and the assumption that $x_1 \neq x_2$, we get a contradiction that $f(x_2) > f(x_1)$ and $f(x_1) > f(x_2)$. Thus, the set of minimizers of such f is a singleton set. □

Next we introduce a property which is widely used in the optimization literature for establishing the convergence of algorithms; see [2–4].

Definition 2.2 (***the Kurdyka-Łojasiewicz property***)**.** *We say that a proper closed function h satisfies the Kurdyka-Łojasiewicz (KL) property at $\widehat{x} \in \mathrm{dom}\partial h$ if there are $a \in (0, \infty]$, a neighborhood V of $\widehat{x}$ and a continuous concave function $\varphi : [0, a) \to [0, \infty)$ with $\varphi(0) = 0$ such that*

1. *φ is continuously differentiable on $(0, a)$ with $\varphi' > 0$;*

2. *For any $x \in V$ with $h(\widehat{x}) < h(x) < h(\widehat{x}) + a$, it holds that*

$$\phi'(h(x) - h(\widehat{x}))\mathrm{dist}(0, \partial h(x)) \geqslant 1.$$

If a proper closed function h satisfies the KL property at every point in $\mathrm{dom}\partial h$*, we say that it is a KL function.*

This property is satisfied by a bunch of proper closed semi-algebraic functions; see [4, Section 2].

The next lemma concerns the uniformized KL property and was proved in [14, Lemma 6].

Lemma 2.2 (Uniformized KL property)**.** *Suppose that h is a proper closed function and let Γ be a compact set. If h is a constant on Γ and satisfies the KL property at each point of Γ, then there exist $\epsilon, a > 0$ and a concave continuous function $\varphi : [0, a) \to [0, \infty)$ that is continuously differentiable on $(0, a)$ with $\varphi(0) = 0$ and $\varphi' > 0$ on $(0, a)$ such that*

$$\varphi'(h(x) - h(\widetilde{x}))\mathrm{dist}\,(0, \partial h(x)) \geqslant 1$$

for any $\widetilde{x} \in \Gamma$ and any x satisfying $h(\widetilde{x}) < h(x) < h(\widetilde{x}) + a$ and $\mathrm{dist}(x, \Gamma) < \epsilon$*.*

Next, we show a simple fact that will be used to deduce the global complexity of the proximal gradient method with extrapolation.

Lemma 2.3. *Suppose the sequence $\{\theta_N\}$ satisfies $\theta_0 = 1$ and $\theta_{N+1} = \frac{\sqrt{\theta_N^4 + 4\theta_N^2} - \theta_N^2}{2}$ for $N \geqslant 0$. Then $\theta_N \leqslant \frac{2}{N+2}$ for any $N \in \mathbb{N}$.*

Proof. We show this by induction. For $N = 0$, the conclusion obviously holds. Suppose that for some $N = k \geqslant 0$, we have $\theta_k \leqslant \frac{2}{k+2}$. Next we prove $\theta_{k+1} \leqslant \frac{2}{k+3}$ by

a contradiction argument. Since

$$\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$$

$$\Longrightarrow \theta_{k+1}^2 + \theta_{k+1}\theta_k^2 = \theta_k^2,$$

if $\theta_{k+1} > \frac{2}{k+3}$, then

$$\left(\frac{2}{k+3}\right)^2 + \left(\frac{2}{k+3}\right)\theta_k^2 < \theta_{k+1}^2 + \theta_{k+1}\theta_k^2 = \theta_k^2$$

$$\Longrightarrow \frac{4}{(k+3)(k+1)} < \theta_k^2.$$

Since $\frac{4}{(k+3)(k+1)} > (\frac{2}{k+2})^2$, we deduce that $\theta_k > \frac{2}{k+2}$, which contradicts the induction assumption that $\theta_k \leqslant \frac{2}{k+2}$. Thus, for any $N \in \mathbb{N}$, $\theta_N \leqslant \frac{2}{N+2}$. □

Before ending this chapter, we prove an auxiliary lemma that will be used in our convergence analysis in Chapter 4. This lemma concerns properties of the regularizer ϕ used in the objective function to be studied in Chapter 4.

Lemma 2.4. *Let $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a continuous concave function with $\phi(0) = 0$ that is continuously differentiable on $(0, \infty)$. Moreover, suppose that $\ell := \lim_{t\downarrow 0} \phi'(t)$ exists. Then the following statements hold:*

(i) *$\phi'(t)$ is nonincreasing and nonnegative when $t > 0$, and $\ell = \phi'_+(0) \geqslant 0$.*[1]

(ii) *$\partial\phi(|\cdot|)(t) = \phi'_+(|t|)\partial|t|$ for all $t \in \mathbb{R}$.*

Proof. First we proof $\phi'(t)$ is nonincreasing. For any $0 < b < a$, since ϕ is concave,

[1] Here and throughout, $\phi'_+(t)$ denotes the right-hand derivative, i.e., $\phi'_+(t) := \lim_{h\downarrow 0} \frac{\phi(t+h)-\phi(t)}{h}$.

for any $\lambda \in (0, 1]$, we have

$$\begin{aligned} \phi(\lambda a + (1-\lambda)b) &\geqslant \lambda\phi(a) + (1-\lambda)\phi(b) \\ \Longrightarrow \frac{\phi(b + \lambda(a-b)) - \phi(b)}{\lambda} &\geqslant \phi(a) - \phi(b) \\ \Longrightarrow \phi'(b)(a-b) &\geqslant \phi(a) - \phi(b) \end{aligned}$$

and similarly

$$\phi'(a)(b-a) \geqslant \phi(b) - \phi(a).$$

Summing the last two inequalities and dividing $a - b$ on both sides we get

$$\phi'(b) - \phi'(a) \geqslant 0.$$

This show that $\phi'(t)$ is nonincreasing when $t > 0$.

Now, for any $\epsilon > 0$, there exist $t_\epsilon \in (0, \epsilon)$ such that

$$0 \leqslant \phi(\epsilon) - \phi(0) = \phi'(t_\epsilon)\epsilon,$$

from which we have $\phi'(t_\epsilon) \geqslant 0$. Thus, $\sup_{0<t\leqslant\epsilon} \phi'(t) \geqslant 0$.

$$\ell = \lim_{t\downarrow 0} \phi'(t) = \lim_{\epsilon\downarrow 0} \sup_{0<t\leqslant\epsilon} \phi'(t) \geqslant 0.$$

Finally, we establish the nonnegativity of ϕ'. Suppose to the contrary that there is a $t_0 > 0$ such that $\phi'(t_0) < 0$. Then for any $t > t_0$, due to the concavity of ϕ, we have

$$\phi(t) \leqslant \phi(t_0) + \phi'(t_0)(t - t_0),$$

where the second inequality is because ϕ' is nonincreasing. Since $\phi(t_0) < 0$, we can find a large t such that the right side of the inequality above becomes negative and thus leads $\phi(t) < 0$, which is contradictive with the assumption that $\phi(t) \geqslant 0$. Thus, for any $t > 0$, $\phi'(t) \geqslant 0$.

To prove (ii), write $g(t) = \phi(|t|)$. Then g is differentiable at any $t \neq 0$ with $\partial g(t) = \phi'(|t|)\partial|t|$.

We now consider the case $t = 0$. In this case, we note first from the definition of regular subdifferential that

$$\begin{aligned}
\widehat{\partial} g(0) &:= \left\{ \mu : \liminf_{y \to 0, y \neq 0} \frac{g(y) - g(0) - \mu y}{|y|} \geqslant 0 \right\} \\
&= \left\{ \mu : \min \left\{ \liminf_{y \to 0, y > 0} \frac{g(y) - g(0) - \mu y}{|y|}, \liminf_{y \to 0, y < 0} \frac{g(y) - g(0) - \mu y}{|y|} \right\} \geqslant 0 \right\} \qquad (2.13) \\
&= \left\{ \mu : \min \left\{ \phi'_+(0) - \mu, \phi'_+(0) + \mu \right\} \geqslant 0 \right\} = [-\phi'_+(0), \phi'_+(0)].
\end{aligned}$$

In addition, from [16, Theorem 6.2.5] and the formula of $\nabla g(t)$ when $t \neq 0$, we have

$$\partial_\circ g(0) = \text{conv } \left\{ \lim_i \nabla g(t_i) : t_i \to 0, t_i \neq 0 \right\} = [-\ell, \ell].$$

Since $\ell = \phi'_+(0)$ according to (i), the above equality together with (2.13) gives

$$[-\phi'_+(0), \phi'_+(0)] = \widehat{\partial} g(0) \subseteq \partial g(0) \subseteq \partial_\circ g(0) = [-\phi'_+(0), \phi'_+(0)],$$

where the first inclusion follows from the definition of the subdifferentials and the second inclusion follows from [17, Theorem 5.2.22]. □

Chapter 3

A brief survey on proximal gradient methods and iteratively reweighted algorithms

3.1 The proximal gradient method

The model (1.4) can be reformulated into semidefinite programming problems and can be solved by interior point methods; see [15, 18, 28, 39, 52, 91, 100] for example. However, these methods can be inefficient when the problem size is large. For large-scale problems, first-order methods such as the proximal gradient method and its variants have been adapted for solving the model (1.4); see [12, 30, 54, 68, 72, 80, 90, 95]. More about the proximal gradient method, its variants and the convergence properties of them can be found in [57, 58, 92]. Specifically, the proximal gradient method is designed to solve the following class of problems:

$$\min \ F(x) := f(x) + P(x), \tag{3.1}$$

where $f : \mathbb{R}^n \to (-\infty, +\infty]$ has Lipschitz continuous gradient with modulus $L > 0$ and $P : \mathbb{R}^n \to (-\infty, \infty]$ is proper closed. The model (1.4) is in this form.

The proximal gradient method can be described as follows:

Proximal Gradient Method (PG)

Step 0. Input: an initial point $x^0 \in \mathrm{dom} P$ and a step size $\gamma > 0$.

Step 1. For $k = 0, 1, 2, ...$

$$x^{k+1} \in \operatorname*{Arg\,min}_x \left\{ f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2\gamma}\|x - x^k\|^2 + P(x) \right\}. \tag{3.2}$$

Step 2. If a termination criterion is not met, go to Step 1.

We next discuss the convergence properties of the sequence generated by this method in both convex and nonconvex cases. We say that x is a stationary point of Problem (3.1) if

$$0 \in \partial F(x).$$

Since f is continuously differentiable, by [79, Exercise 8.8], we know that x is a stationary point of problem (3.1) if and only if

$$0 \in \nabla f(x) + \partial P(x).$$

From [79, Theorem 10.1], we know that a local minimizer of Problem (3.1) is a stationary point.

Theorem 3.1. *Suppose that P in (3.1) is convex and $\inf F > -\infty$. Then the following statements hold.*

(i) *[88, Proposition 1] Suppose that $\gamma \in (0, \frac{2}{L})$ and let $\{x^k\}$ be the sequence generated by PG. Then the sequence $\{F(x^k)\}$ is nonincreasing and any accumulation point of $\{x^k\}$ is a stationary point of problem (3.1).*

(ii) **(Global Complexity)** *[90, Theorem 1] Suppose in addition that f is convex and $\gamma = \frac{1}{L}$. Let $\{x^k\}$ be the sequence generated by PG, then for any x,*

$$F(x^k) \leqslant F(x) + \frac{L}{2k}\|x - x^0\|^2. \tag{3.3}$$

Proof. We start by proving the first conclusion. Using the fact that ∇f is Lipschitz continuous with modulus L, we see that for any x,

$$\begin{aligned}
&f(x^{k+1}) + P(x^{k+1}) \leqslant f(x^k) + \left\langle \nabla f(x^k), x^{k+1} - x^k \right\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 + P(x^{k+1}) \\
&= f(x^k) + \left\langle \nabla f(x^k), x^{k+1} - x^k \right\rangle + \frac{1}{2\gamma}\|x^{k+1} - x^k\|^2 + P(x^{k+1}) \\
&+ \left(\frac{L}{2} - \frac{1}{2\gamma}\right)\|x^{k+1} - x^k\|^2 \\
&\leqslant f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2\gamma}\|x - x^k\|^2 + P(x) - \frac{1}{2\gamma}\|x - x^{k+1}\|^2 \\
&+ \left(\frac{L}{2} - \frac{1}{2\gamma}\right)\|x^{k+1} - x^k\|^2,
\end{aligned} \tag{3.4}$$

where the second inequality follows from Corollary 2.1 and the definition of x^{k+1} as a minimizer of the strongly convex objective in the subproblem (3.2).

Letting $x = x^k$ in the above inequality, we obtain further that

$$\begin{aligned}
&f(x^{k+1}) + P(x^{k+1}) \\
&\leqslant f(x^k) + P(x^k) - \frac{1}{2\gamma}\|x^k - x^{k+1}\|^2 + \left(\frac{L}{2} - \frac{1}{2\gamma}\right)\|x^{k+1} - x^k\|^2 \\
&= f(x^k) + P(x^k) + \left(\frac{L}{2} - \frac{1}{\gamma}\right)\|x^{k+1} - x^k\|^2.
\end{aligned} \tag{3.5}$$

Since $\gamma \in (0, \frac{2}{L})$, we have $(\frac{L}{2} - \frac{1}{\gamma})\|x^{k+1} - x^k\|^2 \leqslant 0$, showing that $\{F(x^k)\}$ is nonincreasing.

Summing (3.5) from $k = 0$ to ∞, we have

$$-\infty < \liminf_k f(x^k) + P(x^k) \leqslant f(x^0) + P(x^0) + \left(\frac{L}{2} - \frac{1}{\gamma}\right)\sum_{k=0}^{\infty}\|x^{k+1} - x^k\|^2,$$

where the first inequality follows from the assumption that $\inf(f + P) > -\infty$. Since

$\gamma \in (0, \frac{2}{L})$, which means $\frac{L}{2} - \frac{1}{\gamma} < 0$, we conclude that

$$\sum_{k=1}^{\infty} \|x^{k+1} - x^k\|^2 < \infty,$$

which implies

$$\lim_k \|x^{k+1} - x^k\| = 0. \tag{3.6}$$

Let $\bar{x}$ be an accumulation point of $\{x^k\}$ and let $\{x^{k_j}\}$ be a subsequence of $\{x^k\}$ such that $x^{k_j} \to \bar{x}$. By the first-order optimality condition of the subproblem in PG and using [79, Exercise 8.8], we have

$$0 \in \nabla f(x^{k_j-1}) + \frac{1}{2\gamma}(x^{k_j} - x^{k_j-1}) + \partial P(x^{k_j}).$$

Using the closedness of the subgradient operator together with (3.6) and the continuity of ∇f, we have upon passing to the limit that

$$0 \in \nabla f(\bar{x}) + \partial P(\bar{x}).$$

Thus, $\bar{x}$ is a stationary point of (3.1).

Next, supposing in addition that f is convex, we prove the second conclusion. Letting $\gamma = \frac{1}{L}$ in (3.4) and using the convexity of f, we obtain that

$$\begin{aligned}
&f(x^{k+1}) + P(x^{k+1}) \\
&\leqslant f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2}\|x - x^k\|^2 + P(x) - \frac{L}{2}\|x - x^{k+1}\|^2 \\
&\leqslant f(x) + P(x) + \frac{L}{2}\|x - x^k\|^2 - \frac{L}{2}\|x - x^{k+1}\|^2.
\end{aligned}$$

Summing this inequality from $k = 0$ to $N - 1$, we have

$$\begin{aligned}
N(f(x^N) + P(x^N)) &\leqslant \sum_{k=0}^{N-1} \left[f(x^{k+1}) + P(x^{k+1})\right] \\
&\leqslant N(f(x) + P(x)) + \frac{L}{2}\|x - x^0\|^2 - \frac{L}{2}\|x - x^N\|^2 \\
&\leqslant N(f(x) + P(x)) + \frac{L}{2}\|x - x^0\|^2,
\end{aligned} \tag{3.7}$$

where the first inequality follows from the first conclusion that $\{F(x^k)\}$ is nonincreasing. Thus, by dividing N on both sides of (3.7), we obtain (3.3).

This completes the proof. □

From the above theorem, we see that when f and P are convex, if a global minimizer x^* of (3.1) exists and we let $x = x^*$ in (3.3), then we see that the sequence of function values generated by PG with $\gamma = \frac{1}{L}$ decreases to the optimal function value at a rate of $O(\frac{1}{k})$.

Next, we discuss the behavior of PG in nonconvex cases.

Theorem 3.2. *Suppose that* $\inf F > -\infty$. *Let* $\gamma \in (0, \frac{1}{L})$ *and* $\{x^k\}$ *be the sequence generated by PG. Then the sequence* $\{F(x^k)\}$ *is nonincreasing and any accumulation point of* $\{x^k\}$ *is a stationary point of problem* (3.1).

Proof. Using the Lipschitz continuity of ∇f, we have

$$\begin{aligned} f(x^{k+1}) + P(x^{k+1}) &\leqslant f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 + P(x^{k+1}) \\ &= f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{1}{2\gamma}\|x^{k+1} - x^k\|^2 + P(x^{k+1}) + \left(\frac{L}{2} - \frac{1}{2\gamma}\right)\|x^{k+1} - x^k\|^2 \\ &\leqslant f(x^k) + P(x^k) + \left(\frac{L}{2} - \frac{1}{2\gamma}\right)\|x^{k+1} - x^k\|^2, \end{aligned} \tag{3.8}$$

where the last inequality follows from the definition of x^{k+1} as a minimizer of the subproblem (3.2). Since $\gamma \in \left(0, \frac{1}{L}\right)$, we have $\left(\frac{L}{2} - \frac{1}{2\gamma}\right)\|x^{k+1} - x^k\|^2 \leqslant 0$, which means that $\{F(x^k)\}$ is nonincreasing.

The second conclusion can be obtained by a similar argument as in the proof of Theorem 3.1 (i). □

3.2 Proximal gradient method with extrapolation

In some applications, the proximal gradient method can be slow; see for example [5, 10, 64–66, 68]. Many variants have been proposed to accelerate PG. While performing line-search provides an empirical way for accelerating an optimization method, incorporating extrapolation is another classical technique for empirical acceleration; see [5, 10, 68, 83, 90, 93]. The application of extrapolation techniques has a long history, dating back to Polyak's heavy ball method [73]. During the past decade, Nesterov's extrapolation techniques [64–66, 68] have been widely adopted and so-called optimal first-order methods have been developed for convex composite optimization problems; see, for example, [6, 10, 11, 53, 64, 90]. As we will introduce in Section 3.2, these are first-order methods that exhibit a function value convergence rate of $O(1/k^2)$, where k is the number of iterations. Extrapolation techniques have also been applied to the proximal gradient algorithm for some classes of nonconvex problems and good empirical performance has been observed; see, for example, [43, 96, 99]. In particular, the proximal gradient method with extrapolation is as follows:

Proximal gradient method with extrapolation ($\mathbf{PG}_e$)

Step 0. Input initial $x^0 = x^{-1} \in \mathrm{dom} P$, $k = 0$, a nonnegative sequence $\{\beta_k\} \subseteq [0, 1]$.

step 1. $y^k = x^k + \beta_k(x^k - x^{k-1})$.

step 2. Set

$$x^{k+1} \in \operatorname*{Arg\,min}_x \left\{ f(y^k) + \left\langle \nabla f(y^k), x - y^k \right\rangle + \frac{L}{2}\|x - y^k\|^2 + P(x) \right\}. \quad (3.9)$$

Step 3. If a termination criterion is not met, set $k = k + 1$ and go to Step 1.

When f and P are convex, under the conditions $\limsup \beta_k < 1$, $\beta_k \geqslant 0$ and

β_k in addition satisfies $\sum_k \beta_k \|x^k - x^{k-1}\|^2 < \infty$, the sequence generated by this method is weakly convergent to a solution of problem (3.1); see [1]. From the first-order optimality condition [79, Theorem 10.1] and the analysis in [1], Problem (3.1) can be interpreted as a fixed point problem. The problem of finding an x such that $x = \text{Prox}_F(x)$ for F in (3.1), and can be solved by the algorithms proposed in [1]. In [1, Proposition 2.1], F. Alvarez and H. Attouch showed that when $\{\beta_k\}$ is nondecreasing and $\limsup \beta_k < \frac{1}{3}$, the sequence generated by this algorithm is globally convergent to a solution of problem (3.1). In [49], P. R. Johnstone and P. Moulin established a convergence results when $\limsup_k \beta_k < 1$ and $\beta_k \geqslant 0$, and they also showed that when $P(x) = \lambda\|x\|_1$ and under a strict complementarity condition, this method with a special choice of β_k generates $\{x^k\}$ and $\{F(x^k)\}$ that are linearly convergent.

Other popular choices of $\{\beta_k\}$ in this algorithm that surpass the proximal gradient method both theoretically and experimentally can be found in [21, 64, 68, 70, 96]. One choice proposed by Nesterov in [64] has a complexity of $O(\frac{1}{k^2})$, which is used extensively in subsequent work. Here we present a detailed and representative analysis of the sequence generated by PG_e with a general $\{\beta_k\}$. In [95], for the problem

$$\min_x f(x) + P(x), \tag{3.10}$$

where P is proper closed convex and f has a Lipschitz continuous gradient, it was shown that if $\{\beta_k\}$ is below some threshold value, then any accumulation point of the sequence generated by PG_e is a stationary point of (3.10).

Note that if f has a Lipschitz continuous gradient, it can be decomposed into the form $f = f_1 - f_2$, where f_1, f_2 are both convex and have Lipschitz continuous gradients. For example, we can always let $f_1(x) = f(x) + \frac{c}{2}\|x\|^2$; $f_2(x) = \frac{c}{2}\|x\|^2$ for some $c \geqslant L_f$, where L_f is the Lipschitz constant of ∇f. The function f_1 is convex

because for any $x, y \in \mathbb{R}^n$,

$$\begin{aligned}
&\langle \nabla f_1(x) - \nabla f_1(y), x - y \rangle \\
=& \langle \nabla f(x) + cx - \nabla f(y) - cy, x - y \rangle \\
=& \langle \nabla f(x) - \nabla f(y), x - y \rangle + c\|x - y\|^2 \\
\geqslant& -\|\nabla f(x) - \nabla f(y)\| \cdot \|x - y\| + c\|x - y\|^2 \\
\geqslant& -L_f\|x - y\|^2 + c\|x - y\|^2 \geqslant 0,
\end{aligned}$$

where the second inequality is because f is Lipschitz continuous and the assumption that $c \geqslant L_f$. By [65, Theorem 2.1.3], we conclude that f_1 is convex.

In what follows, without loss of generality, we suppose $f = f_1 - f_2$ for some f_1 and f_2 that are convex and have Lipschitz constant gradients whose moduli are L and l respectively. We further set $L \geqslant l$, which can always be satisfied by taking larger L. Under this setting, the Lipschitz continuity modulus of f is also L.

Lemma 3.1. *([95, Lemma 3.1]) Suppose the P in (3.10) is convex and $\beta_k \in [0, 1]$ and $\bar{\beta} := \sup_k \beta_k \leqslant \sqrt{\frac{L}{l+L}}$. Let $\{x^k\}$ be the sequence generated by PG_e and $\alpha > 0$. Denote*

$$H_{k,\alpha} := F(x^k) + \alpha\|x^k - x^{k-1}\|^2.$$

Then for all k,

$$H_{k+1,\alpha} - H_{k,\alpha} \leqslant \left(-\frac{L}{2} + \alpha\right)\|x^{k+1} - x^k\|^2 + \left(\frac{L+l}{2}\beta_k^2 - \alpha\right)\|x^k - x^{k-1}\|^2. \quad (3.11)$$

Proof. Note that the objective function of problem (3.9) is strongly convex and x^{k+1} is the minimizer of (3.9). By Corollary 2.1, for any $z \in \mathbb{R}^n$ we have

$$\begin{aligned}
&f(y^k) + \left\langle \nabla f(y^k), x^{k+1} - y^k \right\rangle + \frac{L}{2}\|x^{k+1} - y^k\|^2 + P(x^{k+1}) \\
\leqslant& f(y^k) + \left\langle \nabla f(y^k), z - y^k \right\rangle + \frac{L}{2}\|z - y^k\|^2 + P(z) - \frac{L}{2}\|z - x^{k+1}\|^2.
\end{aligned}$$

Rearrange this inequality, we have for any $z \in \mathbb{R}^n$,

$$P(x^{k+1}) \leqslant \left\langle \nabla f(y^k), z - x^{k+1} \right\rangle + \frac{L}{2}\|z - y^k\|^2 + P(z) - \frac{L}{2}\|z - x^{k+1}\|^2 - \frac{L}{2}\|x^{k+1} - y^k\|^2. \tag{3.12}$$

On the other hand, since ∇f is Lipschitz continuous with constant L, we have

$$f(x^{k+1}) \leqslant f(y^k) + \left\langle \nabla f(y^k), x^{k+1} - y^k \right\rangle + \frac{L}{2}\|x^{k+1} - y^k\|^2. \tag{3.13}$$

Summing (3.12) and (3.13), since $f = f_1 - f_2$ as we assume, we get for any $z \in \mathbb{R}^n$,

$$\begin{aligned}
F(x^{k+1}) &\leqslant f(y^k) + \left\langle \nabla f(y^k), z - y^k \right\rangle + P(z) + \frac{L}{2}\|z - y^k\|^2 - \frac{L}{2}\|z - x^{k+1}\|^2 \\
&= f_1(y^k) + \left\langle \nabla f_1(y^k), z - y^k \right\rangle - \left[f_2(y^k) + \left\langle \nabla f_2(y^k), z - y^k \right\rangle + \frac{l}{2}\|z - y^k\|^2 \right] + P(z) \\
&\quad + \frac{l+L}{2}\|z - y^k\|^2 - \frac{L}{2}\|x^{k+1} - z\|^2 \\
&\leqslant f_1(z) - \left[f_2(y^k) + \left\langle \nabla f_2(y^k), z - y^k \right\rangle + \frac{l}{2}\|z - y^k\|^2 \right] + P(z) \\
&\quad + \frac{l+L}{2}\|z - y^k\|^2 - \frac{L}{2}\|x^{k+1} - z\|^2 \\
&\leqslant f_1(z) - f_2(z) + P(z) + \frac{l+L}{2}\|z - y^k\|^2 - \frac{L}{2}\|x^{k+1} - z\|^2 \\
&= F(z) + \frac{l+L}{2}\|z - y^k\|^2 - \frac{L}{2}\|x^{k+1} - z\|^2,
\end{aligned}$$

where the second inequality follows from the convexity of f_1 and the third inequality is because f_2 has Lipschitz continuous gradient with constant l.

Setting $z = x^k$ in the inequality above, the inequality above becomes

$$F(x^{k+1}) \leqslant F(x^k) + \frac{l+L}{2}\beta_k^2\|x^k - x^{k-1}\|^2 - \frac{L}{2}\|x^{k+1} - x^k\|^2,$$

where we made use of the definition of y^k in (3.9) so that $z - y^k = x^k - y^k = -\beta_k(x^k - x^{k-1})$.

Rearrange this inequality, we get

$$
\begin{aligned}
&F(x^{k+1}) + \alpha\|x^{k+1} - x^k\|^2 - F(x^k) - \alpha\|x^k - x^{k-1}\|^2 \\
&\leqslant \frac{l+L}{2}\beta_k^2\|x^k - x^{k-1}\|^2 - \frac{L}{2}\|x^{k+1} - x^k\|^2 + \alpha\|x^{k+1} - x^k\|^2 - \alpha\|x^k - x^{k-1}\|^2 \\
&= \left(-\frac{L}{2} + \alpha\right)\|x^{k+1} - x^k\|^2 + \left(\frac{L+l}{2}\beta_k^2 - \alpha\right)\|x^k - x^{k-1}\|^2.
\end{aligned}
$$

Recalling the definition of $H_{k,\alpha}$ we have the desired conclusion. □

Lemma 3.2. *([95, Lemma 3.3]) Suppose the P in (3.10) is convex and suppose in addition that $\bar{\beta} := \sup_k \beta_k \leqslant \sqrt{\frac{L}{l+L}}$. Let $\{x^k\}$ be a sequence generated by PG_e and let $\alpha \in [\frac{L+l}{2}\bar{\beta}^2, \frac{L}{2}]$. If $\inf_{x\in\mathbb{R}^n} F(x) > -\infty$, then*

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

Proof. Since $\alpha \in [\frac{L+l}{2}\bar{\beta}^2, \frac{L}{2}]$, by (3.11), we know

$$
\begin{aligned}
H_{k+1,\alpha} - H_{k,\alpha} &\leqslant \left(-\frac{L}{2} + \alpha\right)\|x^{k+1} - x^k\|^2 + \left(\frac{L+l}{2}\beta_k^2 - \alpha\right)\|x^k - x^{k-1}\|^2 \\
&\leqslant \left(\frac{L+l}{2}\beta_k^2 - \alpha\right)\|x^k - x^{k-1}\|^2.
\end{aligned}
$$

Summing this inequality from $k = 0$ to $k = N-1$, we have

$$H_{N,\alpha} - H_{0,\alpha} \leqslant \sum_{k=0}^{N-1}\left(\frac{L+l}{2}\beta_k^2 - \alpha\right)\|x^k - x^{k-1}\|^2. \tag{3.14}$$

Since $\inf_{x\in\mathbb{R}^n} F(x) > -\infty$, by the definition of $H_{k,\alpha}$, we have

$$\inf_k H_{k,\alpha} \geqslant \inf_{x\in\mathbb{R}^n} F(x) > -\infty.$$

Thus, passing N in (3.14) to infinity, we have

$$-\infty < \liminf_N H_{N,\alpha} - H_{0,\alpha} \leqslant \sum_{k=0}^{\infty}\left(\frac{L+l}{2}\beta_k^2 - \alpha\right)\|x^k - x^{k-1}\|^2,$$

which gives the desired conclusion. □

Now we are ready to present the following result in [95].

Theorem 3.3. *([95, Lemma 3.4]) Suppose the P in* (3.10) *is convex and $\bar{\beta} < \sqrt{\frac{L}{l+L}}$ and $\inf_{x\in\mathbb{R}^n} F(x) > -\infty$. Let $\{x^k\}$ be a sequence generated by PG_e. Then any accumulation point of $\{x^k\}$ is a stationary point of problem* (3.10).

Proof. Let $\bar{x}$ be an accumulation point of $\{x^k\}$ generated by PG_e and let $\{x^{k_j}\}$ be a subsequence of $\{x^k\}$ such that $x^{k_j} \overset{j}{\to} \bar{x}$. Then by the first-order optimality condition of subproblem (3.9) in PG_e and the definition of y^{k_j}, using [79, Exercise 8.8], we have

$$-L\left[(x^{k_j+1} - x^{k_j}) - \beta_{k_j}(x^{k_j} - x^{k_j-1})\right] \in \nabla f(y^{k_j}) + \partial P(x^{k_j+1}). \tag{3.15}$$

Since $\bar{\beta} < \sqrt{\frac{L}{l+L}}$, setting $\alpha \in (\frac{L+l}{2}\bar{\beta}^2, \frac{L}{2})$, for all k, we have

$$\alpha - \frac{L+l}{2}\beta_k^2 \geqslant \alpha - \frac{L+l}{2}\bar{\beta}^2 > 0.$$

Thus by Lemma 3.2, we have

$$\begin{aligned}\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|^2 &\leqslant \sum_{k=0}^{\infty} \frac{\alpha - \frac{L+l}{2}\beta_k^2}{\alpha - \frac{L+l}{2}\bar{\beta}^2}\|x^{k+1} - x^k\|^2 \\ &= \frac{1}{\alpha - \frac{L+l}{2}\bar{\beta}^2}\sum_{k=0}^{\infty}\left(\alpha - \frac{L+l}{2}\beta_k^2\right)\|x^{k+1} - x^k\|^2 < \infty.\end{aligned}$$

This means $\lim_{k\to\infty} \|x^{k+1} - x^k\|^2 = 0$. Together with the definition of y^{k_j}, the continuity of ∇f and the closedness of ∂P, by passing k in (3.15) to infinity, we have

$$0 \in \nabla f(\bar{x}) + \partial P(\bar{x}).$$

Therefore the conclusion is proved. □

3.2.1 A fast iterative shrinkage-thresholding algorithm (FISTA)

In [89], Tseng presented various variants of the proximal gradient method that utilize Nesterov's acceleration scheme (see [64]) and established the complexity of $O\left(\frac{1}{k^2}\right)$. In [10,68], extrapolation techniques for accelerating PG were incorporated for solving (3.2) when f and P are convex, which results in a global complexity of $O\left(\frac{1}{k^2}\right)$. In [70, 96], the PG_e with β_k chosen as in Nesterov's extrapolation techniques [64] coupled with a restart scheme exhibits fast convergence in their numerical tests. Also, in [21], a global complexity of $O\left(\frac{1}{k^{1+d}}\right)$ was established under suitable assumptions on β_k, where $d \in (0, 1]$.

In this thesis, we study a fast iterative shrinkage-thresholding algorithm (FISTA) proposed in [68]. We follow the analysis in [89] in detail. The precise algorithm is as follows:

A fast iterative shrinkage-thresholding algorithm (FISTA)

Step 0. Input initial $x^0 = x^{-1} \in \mathrm{dom}P$, $\theta_0 = \theta_{-1} = 1$. Set $k = 0$.

step 1. $y^k = x^k + \theta_k(\theta_{k-1}^{-1} - 1)(x^k - x^{k-1})$;

step 2. Set

$$x^{k+1} = \arg\min_x \left\{ f(y^k) + \left\langle \nabla f(y^k), x - y^k \right\rangle + \frac{L}{2}\|x - y^k\|^2 + P(x) \right\}. \tag{3.16}$$

Step 3. Compute $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$.

Step 4. If a termination criterion is not met, let $k = k + 1$ and go to Step 1.

Theorem 3.4 (***Global complexity***). *[90, Theorem1] Suppose f and P in* (3.1) *are convex and* $\inf F > -\infty$. *Let $\left\{x^k\right\}$ be the sequence generated by FISTA. Then for any $x \in \mathrm{dom}\, P$,*

$$F(x^k) \leqslant F(x) + \frac{2L\|x - x^0\|^2}{(k+2)^2}, \text{ for } k \geqslant 1. \tag{3.17}$$

Proof. Since the objective function of the subproblem of FISTA (3.16) is a strongly convex function with modulus $\frac{1}{L}$ and x^{k+1} is the minimizer of the subproblem of FISTA (3.16), by Corollary 2.1, for any y, we have

$$\begin{aligned} & f(y^k) + \left\langle \nabla f(y^k), x^{k+1} - y^k \right\rangle + \frac{L}{2}\|x^{k+1} - y^k\|^2 + P(x^{k+1}) \\ & \leqslant f(y^k) + \left\langle \nabla f(y^k), y - y^k \right\rangle + \frac{L}{2}\|y - y^k\|^2 + P(y) - \frac{L}{2}\|y - x^{k+1}\|^2. \end{aligned} \tag{3.18}$$

Since ∇f is Lipschitz continuous with modulus L, we have for any y,

$$\begin{aligned} F(x^{k+1}) &= f(x^{k+1}) + P(x^{k+1}) \\ &\leqslant f(y^k) + \left\langle \nabla f(y^k), x^{k+1} - y^k \right\rangle + \frac{L}{2}\|x^{k+1} - y^k\|^2 + P(x^{k+1}) \\ &\leqslant f(y^k) + \left\langle \nabla f(y^k), y - y^k \right\rangle + \frac{L}{2}\|y - y^k\|^2 + P(y) - \frac{L}{2}\|y - x^{k+1}\|^2 \\ &\leqslant f(y) + P(y) + \frac{L}{2}\|y - y^k\|^2 - \frac{L}{2}\|y - x^{k+1}\|^2 \\ &= F(y) + \frac{L}{2}\|y - y^k\|^2 - \frac{L}{2}\|y - x^{k+1}\|^2, \end{aligned} \tag{3.19}$$

where the second inequality follows from (3.18) and the last inequality follows from the convexity of f.

Fixing any $x \in \mathrm{dom} f$ and letting $y = (1-\theta_k)x^k + \theta_k x$ with any fixed x, (3.19) becomes

$$\begin{aligned} & F(x^{k+1}) \\ & \leqslant F((1-\theta_k)x^k + \theta_k x) + \frac{L}{2}\|(1-\theta_k)x^k + \theta_k x - y^k\|^2 - \frac{L}{2}\|(1-\theta_k)x^k + \theta_k x - x^{k+1}\|^2 \\ & = F((1-\theta_k)x^k + \theta_k x) + \frac{L\theta_k^2}{2}\left\| x + \left(\frac{1}{\theta_k} - 1\right)x^k - \frac{1}{\theta_k}y^k \right\|^2 \\ & \quad - \frac{L\theta_k^2}{2}\left\| \left(\frac{1}{\theta_k} - 1\right)x^k + x - \frac{1}{\theta_k}x^{k+1} \right\|^2 \\ & = F((1-\theta_k)x^k + \theta_k x) + \frac{L\theta_k^2}{2}\|x - z^k\|^2 - \frac{L\theta_k^2}{2}\|x - z^{k+1}\|^2, \end{aligned}$$

where

$$z^k = -\left(\frac{1}{\theta_k} - 1\right) x^k + \frac{1}{\theta_k} y^k$$

and the last equality follows from Step 3 in this algorithm together with the definitions of y^k and z^k.

By the convexity of F, we further conclude from the above inequality that

$$F(x^{k+1}) \leqslant (1-\theta_k) F(x^k) + \theta_k F(x) + \frac{L\theta_k^2}{2}\|x - z^k\|^2 - \frac{L\theta_k^2}{2}\|x - z^{k+1}\|^2,$$

i.e.,

$$F(x^{k+1}) - F(x) \leqslant (1-\theta_k)\left[F(x^k) - F(x)\right] + \frac{L\theta_k^2}{2}\|x - z^k\|^2 - \frac{L\theta_k^2}{2}\|x - z^{k+1}\|^2.$$

Dividing θ_k^2 on both sides, we have

$$\frac{1}{\theta_k^2}\left[F(x^{k+1}) - F(x)\right] \leqslant \left(\frac{1}{\theta_k^2} - \frac{1}{\theta_k}\right)\left[F(x^k) - F(x)\right] + \frac{L}{2}\|x - z^k\|^2 - \frac{L}{2}\|x - z^{k+1}\|^2$$

Recall that $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$ for any $k \geq 0$. Thus,

$$\frac{1}{\theta_k^2} = \frac{1}{\theta_{k+1}^2} - \frac{1}{\theta_{k+1}}, \tag{3.20}$$

which further yields for all $k \geqslant 0$,

$$\begin{aligned}
&\left(\frac{1}{\theta_{k+1}^2} - \frac{1}{\theta_{k+1}}\right)\left[F(x^{k+1}) - F(x)\right] \\
&\leqslant \left(\frac{1}{\theta_k^2} - \frac{1}{\theta_k}\right)\left[F(x^k) - F(x)\right] + \frac{L}{2}\|x - z^k\|^2 - \frac{L}{2}\|x - z^{k+1}\|^2.
\end{aligned}$$

Summing this from $k = 0$ to $k = N-1$, we obtain that

$$\begin{aligned}
\frac{1}{\theta_{N-1}^2}\left[F(x^N) - F(x)\right] &= \left(\frac{1}{\theta_N^2} - \frac{1}{\theta_N}\right)\left[F(x^N) - F(x)\right] \\
&\leqslant \left(\frac{1}{\theta_0^2} - \frac{1}{\theta_0}\right)\left[F(x^0) - F(x)\right] + \frac{L}{2}\|x - z^0\|^2 - \frac{L}{2}\|x - z^N\|^2 \\
&\leqslant \frac{L}{2}\|x - x^0\|^2,
\end{aligned}$$

where the equality follows from the setting (3.20) and the second inequality follows from $\theta_0 = 1$ and $z^0 = -\left(\frac{1}{\theta_0} - 1\right) x^0 + \frac{1}{\theta_0} y^0 = x^0$. Thus,

$$F(x^N) \leqslant F(x) + \theta_{N-1}^2 \frac{L}{2} \|x - x^0\|^2.$$

By Lemma 2.3, we know that for any $N \geqslant 1$, $\theta_N \leqslant \frac{2}{N+2}$, which establishes the global complexity of FISTA. □

3.3 Iteratively reweighted algorithms

To solve (1.7), the iteratively reweighted algorithms were introduced; see [26, 27, 32, 33, 56, 69, 71, 74, 97]. The iteratively reweighted algorithms can be divided into two kinds: the iteratively reweighted ℓ_2 algorithm and the iteratively reweighted ℓ_1 algorithm. The subproblem of the iteratively reweighted ℓ_2 algorithm takes the following form:

$$x^{k+1} \in \operatorname*{Arg\,min}_{\{x:\ Ax=b\}} \sum_{i=1}^{n} w_i^k x_i^2, \tag{3.21}$$

where $\{w_i^k\} \subseteq \mathbb{R}_{++}$ variate in different literatures, and we discuss some possible choices of $\{w_i^k\}$ below. In [27], for solving

$$\min_{Ax=b} \sum_{i=1}^{n} |x_i|^p, \tag{3.22}$$

Chartrand and Yin proposed an iteratively reweighted ℓ_2 algorithm whose subproblem takes the form (3.21) with $w_i^k = \left((x_i^k)^2 + \epsilon\right)^{p/2-1}$ for some $\epsilon > 0$. Applying these weights, in the numerical tests of [27], their algorithm had a better recovery than the one with the weight $w_i^k = \left((x_i^k)^2\right)^{p-2}$. In addition, it was also shown in [27] that when the sparsity of the signal is under a threshold and A satisfies some assumptions, the solution to the model

$$\min_{\{x:\ Ax=b\}} \sum_{i=1}^{n} w_i x_i^2$$

with $w_i = \left((x_i^k)^2 + \epsilon_j\right)^{p/2-1}$ converges to the solution of (3.22) as $\epsilon_j \downarrow 0$. In [33], the weight w_i^k in (3.21) is chosen as $[(x_i^k)^2 + \epsilon_k^2]^{-1/2}$ with specially chosen $\epsilon_k > 0$. It was proved that using these weights and under some additional assumptions, one of which is that A satisfies the restricted isometry property (RIP) in [33], the sequence generated by their iteratively reweighted ℓ_2 algorithm converges globally. In addition, under further assumptions on ϵ_k, it was proved in [33] that the sequence generated by (3.21) converges to a solution of (1.1). The convergence rate was also given in [33] under suitable assumptions. More applications and variants of the iteratively reweighted ℓ_2 algorithm can be found in [71, 74, 97].

On the other hand, the iteratively reweighted ℓ_1 algorithm can also be applied to solving (3.22) and other variants of (3.22) with other regularizers like the logistic function (see [26, 97]) in place of ℓ_p. The subproblem of the iteratively reweighted ℓ_1 algorithm is given as follows:

$$x^{k+1} \in \operatorname*{Arg\,min}_{\{x:\ Ax=b\}} \sum_{i=1}^{n} w_i^k |x_i|, \tag{3.23}$$

where different works set different $\{w_i^k\} \subseteq \mathbb{R}_{++}$. For instance, to solve (3.22) with the logistic function in place of ℓ_1, i.e., the following model with some $\epsilon > 0$,

$$\min_{\{x:\ Ax=b\}} \sum_{i=1}^{n} \log(|x_i| + \epsilon), \tag{3.24}$$

We set $w_i^k = \frac{1}{|x_i^k|+\epsilon}$ in (3.23) for all $i = 1, 2, \ldots, n$. In this case, [26, Figure. 4(a)] shows that the iteratively reweighted ℓ_1 algorithm outperforms the one with all weights being 1 in sparse signal recovery. One explanation in [26] for this behavior is that $\sum_{i=1}^n w_i|x_i|$ with $w_i = \frac{1}{|x_i^*|+\epsilon}$, where x^* is the local minimizer of problem (3.24), is the first-order approximation of the logistic regularization function in (3.24).

The subproblem of iteratively reweighted ℓ_1 algorithm can also be viewed as a

variant of the convex relaxation (1.1) itself, which is as follows:

$$\min_{\{x:\ Ax=b\}} \sum_{i=1}^{n} w_i|x_i|, \tag{3.25}$$

where we use the "weighted" ℓ_1 norm in place of the ℓ_1 norm with $w_i \geq 0$. When $w_i = 1$ for all $i = 1, 2, \ldots, n$, this model reduces to (1.1). In [26], the picture [26, Figure. 1] intuitively shows that by properly choosing the weights, the weighted l_1 in (3.25) may be more efficient in inducing sparsity than l_1.

The reweighted ℓ_1 scheme in (3.23) has also been incorporated into the proximal gradient methods to solve (1.7). The iteratively reweighted ℓ_1 algorithm in [26, 27, 32, 56] outperforms the proximal gradient method in experimental results and the convergence rate of the sequence generated was studied under mild assumptions. In [69], the convergence of the sequence generated by the iteratively reweighted ℓ_1 algorithm to a stationary point was proved under KL assumptions and other conditions. In [32], Chen and Zhou proved that any accumulation point of $\{x^{\epsilon_k}\}$ is a global minimizer of (1.7) if $\epsilon_k \to 0^+$, where x^{ϵ_k} is the global minimizer of

$$\min \|Ax - b\|_2^2 + \sum_{i=1}^{n} (|x_i| + \epsilon_k)^p. \tag{3.26}$$

They also established the convergence rate of their algorithms under some conditions on p in (1.7) and the assumption that the sequence generated by their iteratively reweighted ℓ_1 algorithm converges to a local minimizer of the problem (3.26) for each ϵ_k; see [32, Theorem 4]. On the other hand, in [56], instead of using $\sum_{i=1}^{n}(|x_i| + \epsilon_k)^p$ with a dynamic ϵ_k to approach the ℓ_p and applying the iteratively reweighted ℓ_1 algorithm to (3.26) for each ϵ_k, Lu proposed several variants of the iteratively reweighted ℓ_1 algorithm to solve (3.26) with ϵ_k being constant. Specifically, in his paper, he started with the following model:

$$\min F(x) = f(x) + \lambda \sum_{i=1}^{n} \|x\|^p, \tag{3.27}$$

f is bounded below and ∇f is Lipschitz continuous; $\lambda > 0$ and $p \in (0,1)$. When $f(x) = \|Ax - b\|^2$, this model reduces to (1.7). He then adopted a Lipschitz continuous approximation to ℓ_p, which results in the following problem:

$$\min F_\epsilon(x) := f(x) + \lambda \sum_{i=1}^{n} h_{u_\epsilon}(x_i), \tag{3.28}$$

where

$$h_{u_\epsilon}(t) = \min_{0 \leqslant s \leqslant u_\epsilon} p\left(|t|s - \frac{s^q}{q}\right), \; u_\epsilon = \left(\frac{\epsilon}{\lambda n}\right)^{\frac{1}{q}},$$

and $\epsilon > 0$. Note that h_{u_ϵ} is a Lipschitzation of ℓ_p and this problem uses h_{u_ϵ} as an approximation of ℓ_p with a fixed ϵ instead of an approximation of ℓ_p with a dynamic ϵ_k as in (3.26). In [56], Lu proved that using a fixed ϵ, the sequence generated by the 7th algorithm of [56] accumulates at a stationary point of (3.27) under suitable assumptions, which is different from the convergence results in [32] that use a dynamic ϵ_k. The 7th algorithm of [56] is precisely as follows:

A variant of new iterative reweighted l_1 ($\mathbf{IRL_1}$) minimization method for (3.28)

Step 0. Let $0 < L_{\min} < L_{\max}$, $\tau > 1$ and $c > 0$ be given. Let q be such that $q^{-1} + p^{-1} = 1$. Input an initial point x^0, $\epsilon > 0$ and set $k = 0$.

Step 1. Choose $L_k^0 \in [L_{\min}, L_{\max}]$ and set $L_k = L_k^0$.

Step 2. Set

$$s_i^{k+1} = \min\left\{\left(\frac{\epsilon}{\lambda n}\right)^{q^{-1}}, \left|x_i^k\right|^{(q-1)^{-1}}\right\} \text{ for } i = 1, \ldots, n;$$

$$x^{k+1} = \arg\min_x \left\{ f(x^k) + \left\langle \nabla f(x^k), y - x^k \right\rangle + \frac{L_k}{2}\|y - x^k\|^2 + \lambda p \sum_{i=1}^{n} s_i^{k+1}|y_i| \right\}.$$

Step 3. If

$$F_\epsilon(x^{k+1}) - F_\epsilon(x^k) \geqslant \frac{c}{2}\|x^{k+1} - x^k\|^2,$$

let $L_k = \tau L_k$, and go to Step 2.

Step 4. If a termination criterion is not met, set $k = k + 1$ and go to Step 1.

Note that the subproblem of this algorithm minimizes the sum of a quadratic and the weighted ℓ_1 norm. Also, line search is incorporated for empirical acceleration in this algorithm.

In the next chapter, we will propose three new iteratively reweighted ℓ_1 algorithms whose subproblems involve the sum of a quadratic and the weighted ℓ_1 norm, for a large class of problems including (3.28). However, instead of using line-search techniques to numerically accelerate our algorithms, we adapt three types of extrapolation schemes in our reweighted proximal gradient algorithms. As we introduced in section 3.2, the proximal gradient method with suitably incorporated extrapolation techniques provably converges at a rate of $O(1/k^2)$. Therefore, it is of interest to investigate how extrapolation techniques behave when adapted to the iteratively reweighted ℓ_1 algorithm.

Chapter 4

Iteratively reweighted ℓ_1 algorithms with extrapolation techniques

4.1 Introduction

In the last chapter, we introduced the iteratively reweighted ℓ_1 algorithm which has been widely studied for minimizing optimization models such as (1.3), (1.4) and (1.5) that attempt to induce sparsity in their solutions. In this chapter we will discuss how extrapolation techniques behave when adapted in the iteratively reweighted ℓ_1 algorithm. Here, we consider this algorithm for solving the following class of optimization problems

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth convex function with a Lipschitz continuous gradient whose Lipschitz modulus is L, C is a nonempty closed convex set, and $x \mapsto \Phi(|x|)$ is a sparsity inducing function: specifically, we assume that $\Phi(y) = \sum_{i=1}^n \phi(y_i)$, where $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a continuous concave function with $\phi(0) = 0$ that is continuously differentiable on $(0, \infty)$, and the limit $\ell := \lim_{t \downarrow 0} \phi'(t)$ exists. We also assume that F is level-bounded and hence $v > -\infty$ as a consequence. Problem (4.1) covers many

applications we mentioned in Section 1.1 such as compressed sensing [25, 40] and statistical variable selections [38, 84, 104, 106], where f is typically a loss function for data fidelity, C is a simple closed convex set (encoding, for example, nonnegativity constraints or box constraints), and ϕ can be, for example, the smoothly clipped absolute deviation (SCAD) function [38], the minimax concave penalty (MCP) function [104] or the log penalty function [67] which we mentioned in Section 1.1.2.[1]

When the iteratively reweighted ℓ_1 algorithm is applied to (4.1), the sparsity inducing regularizer in the objective is approximated by a weighted ℓ_1 norm in each iteration, and the resulting subproblem is then approximately solved to produce the next iterate. These subproblems may not have closed form solutions in general due to the function f, and a variant of the algorithm was proposed in [56] that allows majorization of the smooth function f by a quadratic function with a constant Hessian in each iteration to simplify the subproblem. Moreover, a line-search strategy was incorporated for empirical acceleration; see [56, Algorithm 7].

As we introduced in Section 3.2, first-order methods with extrapolation techniques exhibit a function value convergence rate of $O(1/k^2)$, where k is the number of iterations. In view of the success in accelerating first-order methods such as the proximal gradient algorithm empirically via extrapolations, in this chapter, we investigate how extrapolation techniques can be suitably incorporated into iteratively reweighted ℓ_1 algorithms for solving (4.1). We specifically consider extrapolation techniques motivated from three popular optimal first-order methods: the fast iterative soft-thresholding algorithm (FISTA) [10, 68], the method by Auslender and Teboulle [6] and the method by Lan, Lu and Monteiro [53]. We call the corresponding iteratively reweighted ℓ_1 algorithms with extrapolation IRL$_1e_1$, IRL$_1e_2$ and IRL$_1e_3$,

[1] Note that when f is the least squares loss function and $\Phi(|\cdot|)$ is the MCP or SCAD function, the function $f(\cdot) + \Phi(|\cdot|)$ is not level-bounded (though it necessarily has a minimizer). However, the level-boundedness of F can still be enforced by picking C to be a huge box, i.e., $C = [-M, M]^n$ for a sufficiently large $M > 0$ so that C intersects $\operatorname{Arg\,min}_x\{f(x) + \Phi(|x|)\}$. For this choice of C, the optimal value of F is the same as that of $f(\cdot) + \Phi(|\cdot|)$.

respectively. For each algorithm, we show that the sequence generated clusters at a stationary point of (4.1) under certain condition on the extrapolation parameters. These conditions are satisfied by many choices of extrapolation parameters: for instance, one can pick the parameters as in FISTA with fixed restart [70] in IRL_1e_1. Furthermore, under some additional assumptions such as the Kurdyka-Łojasiewicz property on some suitable potential functions (see for example, [3, 4]), we show that the whole sequence generated by IRL_1e_1 and IRL_1e_3 are indeed convergent. We then perform numerical experiments comparing our algorithms (with our proposed choices of extrapolation parameters) against the general iterative shrinkage and thresholding algorithm (GIST) [46] and an adaptation of the iteratively reweighted ℓ_1 algorithm [56, Algorithm 7] with nonmonotone line-search (IRL_1ls) for solving log penalty regularized least squares problems on random instances. In our experiments, our iteratively reweighted ℓ_1 algorithms with extrapolation usually outperform GIST and IRL_1ls in both CPU time and solution quality. Moreover, IRL_1e_1 and IRL_1e_3 usually perform better than IRL_1e_2.

The rest of this chapter is organized as follows. In Section 4.2, we prove an axillary lemma which will be used in the convergence analysis of IRL_1e_1, IRL_1e_2 and IRL_1e_3. We present the convergence analysis of IRL_1e_1, IRL_1e_2 and IRL_1e_3 in Sections 4.3, 4.4 and 4.5 respectively, and our numerical experiments are presented in Section 4.6. Finally, some concluding remarks are given in Section **??**.

4.2 The sum rule of the subdifferential

In this section, we prove an axillary lemma which will be used in the following sections. This auxiliary lemma has to do with the first-order necessary condition of (4.1). Recall that a point $\bar{x}$ is said to satisfy the first-order necessary condition of

(4.1) if

$$0 \in \partial \left(f(\cdot) + \delta_C(\cdot) + \Phi(|\cdot|) \right)(\bar{x}). \tag{4.2}$$

Such a point is called a stationary point, and it is known from [79, Theorem 10.1] that any local minimizer of (4.1) is a stationary point. Note that computing the subdifferential in (4.2) directly from definition can be complicated. In our next lemma, we show that the subdifferential in (4.2) equals the sum of $\nabla f(\bar{x})$, the normal cone of C at $\bar{x}$ and the set $\Phi'_+(|\bar{x}|) \circ \partial|\bar{x}|$; here and throughout this paper, for any $y \in \mathbb{R}^n$, we write

$$\Phi'_+(|y|) := (\phi'_+(|y_1|), \phi'_+(|y_2|), \ldots, \phi'_+(|y_n|)) \in \mathbb{R}^n_+,$$
$$\partial|y| := \partial|y_1| \times \partial|y_2| \times \cdots \times \partial|y_n| \subset \mathbb{R}^n.$$

Notice that $\Phi'_+(|y|) \in \mathbb{R}^n_+$ for all $y \in \mathbb{R}^n$ is a consequence of Lemma 2.4(i).

Lemma 4.1. *The objective of* (4.1) *satisfies the following equation for any* $x \in C$*:*

$$\partial F(x) = \nabla f(x) + N_C(x) + \Phi'_+(|x|) \circ \partial|x|.$$

Proof. Since $\partial\delta_C(x) \neq \emptyset$ at any $x \in C$, by [79, Corollary 8.11] and [79, Proposition 8.12], we know that f and δ_C are regular at any point in C.

Let $\widetilde{\phi} : \mathbb{R} \to \mathbb{R}$ be defined so that $\widetilde{\phi}(t) = \phi(t)$ when $t > 0$ and $\widetilde{\phi}(t) = \ell t$ otherwise. Then $\widetilde{\phi}$ is continuously differentiable in view of Lemma 2.4(i) and $\phi(|t|) = \max\{\widetilde{\phi}(t), \widetilde{\phi}(-t)\}$ for all $t \in \mathbb{R}$. Using these, we deduce from [79, Example 10.24(e)] that $\phi(|\cdot|)$ is amenable. This together with [79, Exercise 10.26(a)] implies that $\Phi(|\cdot|)$ is amenable. Consequently, $\Phi(|\cdot|)$ is regular, thanks to [79, Exercise 10.25(a)].

Therefore, using [79, Corallary 10.9], we see that

$$\partial F(x) = \nabla f(x) + N_C(x) + \partial\Phi(|\cdot|)(x) = \nabla f(x) + N_C(x) + \Phi'_+(|x|) \circ \partial|x|,$$

where the second equality follows from [79, Proposition 10.5] and Lemma 2.4(ii). □

4.3 Iteratively reweighted ℓ_1 algorithm with type-I extrapolation

In this section, we propose and analyze an iteratively reweighted ℓ_1 algorithm with an extrapolation technique motivated from FISTA [10, 68]: this technique has been widely studied in both convex and nonconvex settings; see, for example, [10, 11, 68, 90, 96]. We call the algorithm based on this extrapolation technique the iteratively reweighted ℓ_1 algorithm with type-I extrapolation (IRL$_1e_1$). This algorithm is presented in Algorithm 4.3 below.

Iteratively reweighted ℓ_1 algorithm with type-I extrapolation (IRL$_1e_1$)

Step 0. Input an initial point $x^0 = x^{-1} \in C$ and $\{\beta_k\} \subset [0, 1)$. Set $k = 0$.

Step 1. Set

$$\begin{aligned} s^{k+1} &= \Phi'_+(|x^k|); \\ y^k &= x^k + \beta_k(x^k - x^{k-1}); \\ x^{k+1} &= \operatorname*{arg\,min}_{y \in C} \left\{ \langle \nabla f(y^k), y - y^k \rangle + \frac{L}{2}\|y - y^k\|^2 + \sum_{i=1}^{n} s_i^{k+1}|y_i| \right\}. \end{aligned} \tag{4.3}$$

Step 2. If a termination criterion is not met, set $k = k + 1$ and go to Step 1.

We next present our global convergence analysis. We will first characterize the cluster points of the sequence generated by the algorithm under suitable assumptions on $\{\beta_k\}$, and then show that the whole sequence is convergent under further assumptions. Our analysis makes extensive use of the following auxiliary function, and is similar to the analysis in [96]:

$$H_1(x, y) := f(x) + \delta_C(x) + \Phi(|x|) + \frac{L}{2}\|x - y\|^2. \tag{4.4}$$

We start by showing that any accumulation point of the sequence $\{x^k\}$ generated by IRL$_1e_1$ is a stationary point of (4.1) under the additional assumption that $\sup_k \beta_k < 1$. This assumption is general enough to accommodate a wide variety of

choices of extrapolation parameters such as those used in FISTA with both fixed and adaptive restart strategies [70]. This latter choice of $\{\beta_k\}$ was shown empirically to be highly effective in accelerating the proximal gradient algorithm for convex composite optimization problems [70] and the proximal difference-of-convex algorithm for a class of difference-of-convex optimization problems [96].

Theorem 4.1. *Suppose that* $\sup_{k\geqslant 0}\beta_k < 1$ *and let* $\{x^k\}$ *be the sequence generated by IRL*$_1$*e*$_1$ *for solving* (4.1)*. Then the following statements hold:*

(i) $\{H_1(x^k, x^{k-1})\}_{k\geqslant 0}$ *is a nonincreasing convergent sequence. Moreover, there exists a positive constant* D_1 *such that*

$$H_1(x^k, x^{k-1}) - H_1(x^{k+1}, x^k) \geqslant D_1\|x^k - x^{k-1}\|^2. \tag{4.5}$$

(ii) *The sequence* $\{x^k\}$ *is bounded and* $\lim_k \|x^{k+1} - x^k\| = 0$.

(iii) *Any accumulation point of* $\{x^k\}$ *is a stationary point of* (4.1).

Proof. First we prove (i). We write $l_f(x; y) := f(y) + \langle \nabla f(y), x - y\rangle$ for notational simplicity. Recall that $\nabla f(x)$ is Lipschitz continuous with modulus L. Then, we have

$$\begin{aligned} F(x^{k+1}) &\leqslant l_f(x^{k+1}; y^k) + \frac{L}{2}\|x^{k+1} - y^k\|^2 + \Phi(|x^{k+1}|) \\ &\leqslant l_f(x^{k+1}; y^k) + \frac{L}{2}\|x^{k+1} - y^k\|^2 + \Phi(|x^k|) + \sum_{i=1}^n s_i^{k+1}(|x_i^{k+1}| - |x_i^k|), \end{aligned}$$

where the second inequality follows from the concavity of ϕ and the definition of s^{k+1}. Notice that x^{k+1} is the minimizer of a strongly convex objective function by its definition in (4.3). Using this together with the above inequality, we see further

that

$$\begin{aligned} F(x^{k+1}) &\leqslant l_f(x^k; y^k) + \frac{L}{2}\|x^k - y^k\|^2 + \Phi(|x^k|) - \frac{L}{2}\|x^{k+1} - x^k\|^2 \\ &\leqslant f(x^k) + \frac{L}{2}\|x^k - y^k\|^2 + \Phi(|x^k|) - \frac{L}{2}\|x^{k+1} - x^k\|^2 \qquad (4.6) \\ &= f(x^k) + \Phi(|x^k|) + \frac{L}{2}\beta_k^2\|x^k - x^{k-1}\|^2 - \frac{L}{2}\|x^{k+1} - x^k\|^2, \end{aligned}$$

where the second inequality follows from the convexity of f, while the equality follows from the definition of y^k in (4.3).

Rearranging (4.6) and invoking the definition of H_1 and the fact that $\sup_k \beta_k < 1$, we have

$$\begin{aligned} 0 &\leqslant \frac{L}{2}(1 - \sup_k \beta_k^2)\|x^k - x^{k-1}\|^2 \leqslant \frac{L}{2}(1 - \beta_k^2)\|x^k - x^{k-1}\|^2 \\ &\leqslant \left[F(x^k) + \frac{L}{2}\|x^k - x^{k-1}\|^2\right] - \left[F(x^{k+1}) + \frac{L}{2}\|x^{k+1} - x^k\|^2\right] \qquad (4.7) \\ &= H_1(x^k, x^{k-1}) - H_1(x^{k+1}, x^k), \end{aligned}$$

which implies that $\{H_1(x^k, x^{k-1})\}$ is nonincreasing and (4.5) holds with $D_1 = \frac{L}{2}(1 - \sup_k \beta_k^2) > 0$. In addition, since $\inf F \geqslant v > -\infty$, we know that $\left\{H_1(x^k, x^{k-1})\right\}$ is bounded from below. Thus, $\lim_k H_1(x^k, x^{k-1})$ exists and (i) holds.

In addition, we have from (4.7) and the definition of H_1 that

$$F(x^k) \leqslant H_1(x^k, x^{k-1}) \leqslant H_1(x^0, x^{-1}) = F(x^0) < \infty,$$

Since F is level-bounded, we conclude from this inequality that $\{x^k\}$ is bounded. Moreover, summing (4.7) from $k = 0$ to ∞, we obtain

$$\frac{L}{2}\sum_{k=0}^{\infty}\left[1 - (\sup_k \beta_k)^2\right]\|x^k - x^{k-1}\|^2 \leqslant H_1(x^0, x^{-1}) - \lim_k H_1(x^{k+1}, x^k) < \infty.$$

Since $\sup_k \beta_k < 1$, we have $\lim_k \|x^{k+1} - x^k\| = 0$. This proves (ii).

Now we prove (iii). Let $\widetilde{x}$ be an accumulation point of $\{x^k\}$ and let $\{x^{k_j}\}$ be a subsequence such that $x^{k_j} \to \widetilde{x}$. Using the first-order optimality condition of the subproblem in (4.3), we have

$$0 \in \nabla f(y^{k_j}) + N_C(x^{k_j+1}) + L(x^{k_j+1} - y^{k_j}) + s^{k_j+1} \circ \partial|x^{k_j+1}|;$$

here we made use of the subdifferential calculus rules in [79, Proposition 10.5] and [79, Proposition 10.9]. Combining this with the definition of y^k in (4.3) and rearranging terms, we deduce that

$$-L[(x^{k_j+1} - x^{k_j}) - \beta_{k_j}(x^{k_j} - x^{k_j-1})] \in \nabla f(y^{k_j}) + N_C(x^{k_j+1}) + s^{k_j+1} \circ \partial|x^{k_j+1}|. \quad (4.8)$$

Next, we claim that

$$\lim_j s^{k_j+1} = \Phi'_+(|\widetilde{x}|). \quad (4.9)$$

To prove this, we first consider those i corresponding to $\widetilde{x}_i \neq 0$. Since ϕ is continuously differentiable, we have from the definition of s^{k_j+1} that $\lim_j s_i^{k_j+1} = \lim_j \phi'_+(|x_i^{k_j}|) = \phi'_+(|\widetilde{x}_i|)$. On the other hand, for those i corresponding to $\widetilde{x}_i = 0$, we have $\lim_j s_i^{k_j+1} = \lim_j \phi'_+(|x_i^{k_j}|) = \ell = \phi'_+(0)$, thanks to Lemma 2.4(i). Thus, (4.9) holds.

Now, notice that $\Phi'_+(x) \circ \partial|x| \subseteq [-\ell, \ell]^n$ for all $x \in \mathbb{R}^n$, meaning that the set-valued mapping $x \rightrightarrows \Phi'_+(x) \circ \partial|x|$ is bounded. Using this, [79, Proposition 5.51], the closedness of convex subdifferentials, (4.9) and the fact that $\lim_k \|x^{k+1} - x^k\| = 0$ from (ii), we see by passing to the limit in (4.8) that

$$0 \in \nabla f(\widetilde{x}) + N_C(\widetilde{x}) + \Phi'_+(|\widetilde{x}|) \circ \partial|\widetilde{x}| = \partial F(\widetilde{x}),$$

where the last equality follows from Lemma 4.1. Thus (iii) holds. □

Corollary 4.1. *Suppose that* $\sup_{k\geqslant 0} \beta_k < 1$ *and let* $\{x^k\}$ *be the sequence generated by IRL*$_1e_1$ *for solving* (4.1)*. Then the set of accumulation points of* $\{(x^k, x^{k-1})\}$*, denoted by* Ω_1*, is a nonempty compact subset of* $\operatorname{dom} \partial H_1$*, and* $H_1 \equiv \lim_k H_1(x^k, x^{k-1})$ *on* Ω_1*.*

Proof. From Theorem 4.1(ii), we know that the set of accumulation points of $\{x^k\}$, denoted by Λ_1, is nonempty and compact. Moreover, since $\lim_k \|x^{k+1} - x^k\| = 0$, we see that $\Omega_1 = \{(x, x) : \ x \in \Lambda_1\}$, which is clearly nonempty and compact. Furthermore, since Λ_1 belongs to $\{x : \ 0 \in \partial F(x)\} \subseteq \operatorname{dom} \partial F$ according to Theorem 4.1(iii), it is routine to check that $\Omega_1 \subset \operatorname{dom} \partial H_1$.

Next, choose any $(\widetilde{x}, \widetilde{x}) \in \Omega_1$ and let $\{x^{k_j}\}$ be a subsequence of $\{x^k\}$ with $x^{k_j} \to \widetilde{x}$. Then

$$H_1(\widetilde{x}, \widetilde{x}) = F(\widetilde{x}) = \lim_j F(x^{k_j}) + \frac{L}{2}\|x^{k_j} - x^{k_j-1}\|^2 = \lim_j H_1(x^{k_j}, x^{k_j-1}),$$

where the second equality follows from the continuity of F on C and Theorem 4.1(ii). Since $\{H_1(x^k, x^{k-1})\}$ is convergent thanks to Theorem 4.1(i) and $(\widetilde{x}, \widetilde{x}) \in \Omega_1$ is chosen arbitrarily, we obtain that $H_1 \equiv \lim_k H_1(x^k, x^{k-1})$ on Ω_1. This completes the proof. □

Next, we prove under additional assumptions on H_1 and ϕ'_+ that the whole sequence $\{x^k\}$ generated by $\mathrm{IRL}_1 e_1$ is convergent to a stationary point of (4.1). We start with an auxiliary lemma.

Lemma 4.2. *Suppose that* $\sup_{k\geqslant 0} \beta_k < 1$ *and that* ϕ'_+ *is Lipschitz continuous on* $[0, \infty)$. *Let* $\{x^k\}$ *be the sequence generated by IRL*$_1 e_1$ *for solving* (4.1). *Then there exists a positive constant* C_1 *such that for all* $k \geqslant 1$,

$$\operatorname{dist}\big((0,0), \partial H_1(x^k, x^{k-1})\big) \leqslant C_1(\|x^{k-1} - x^{k-2}\| + \|x^k - x^{k-1}\|).$$

Proof. First, using the first-order optimality condition of the subproblem in (4.3) and the definition of s_i^k, there exist a $\xi^k \in \partial|x^k|$ and a $\zeta^k \in N_C(x^k)$ such that

$$0 = \nabla f(y^{k-1}) + \zeta^k + \Phi'_+(|x^{k-1}|) \circ \xi^k + L(x^k - y^{k-1}) \tag{4.10}$$

for all $k \geqslant 1$. Define $\eta^k := \nabla f(x^k) + \zeta^k + \Phi'_+(|x^k|) \circ \xi^k + L(x^k - x^{k-1})$. Then we have

$$\begin{aligned}
&(\eta^k, -L(x^k - x^{k-1})) \\
&\in \begin{pmatrix} \nabla f(x^k) + N_C(x^k) + \Phi'_+(|x^k|) \circ \partial|x^k| + L(x^k - x^{k-1}) \\ \{-L(x^k - x^{k-1})\} \end{pmatrix} \\
&= \partial H_1(x^k, x^{k-1}).
\end{aligned}$$

where the equality follows from [79, Exercise 8.8], [79, Proposition 10.5] and Lemma 4.1. Consequently, we have for all $k \geqslant 0$ that

$$\operatorname{dist}\big((0,0), \partial H_1(x^k, x^{k-1})\big) \leqslant \sqrt{\|\eta^k\|^2 + L^2\|x^k - x^{k-1}\|^2}. \tag{4.11}$$

On the other hand, from the definition of η^k and (4.10), we see that

$$\begin{aligned}
\|\eta^k\| &= \big\|\eta^k - \big[\nabla f(y^{k-1}) + \zeta^k + \Phi'_+(|x^{k-1}|) \circ \xi^k + L(x^k - y^{k-1})\big]\big\| \\
&= \big\|\nabla f(x^k) - \nabla f(y^{k-1}) - L(x^{k-1} - y^{k-1}) + \big[\Phi'_+(|x^k|) - \Phi'_+(|x^{k-1}|)\big] \circ \xi^k\big\| \\
&\leqslant \|\nabla f(x^k) - \nabla f(y^{k-1})\| + L\|x^{k-1} - y^{k-1}\| + \|\Phi'_+(|x^k|) - \Phi'_+(|x^{k-1}|)\| \\
&\leqslant \|\nabla f(x^k) - \nabla f(y^{k-1})\| + L\|x^{k-1} - y^{k-1}\| + \sqrt{\sum_{i=1}^n \rho^2(|x_i^k| - |x_i^{k-1}|)^2} \\
&\leqslant L\|x^k - y^{k-1}\| + L\|x^{k-1} - y^{k-1}\| + \rho\|x^k - x^{k-1}\| \\
&\leqslant (L + \rho)\|x^k - x^{k-1}\| + 2L\|x^{k-1} - x^{k-2}\|,
\end{aligned} \tag{4.12}$$

where the first inequality follows from the elementary inequality $\|a \circ b\| \leq \|b\|_\infty \|a\|$ for any a, $b \in \mathbb{R}^n$ and the fact that $\|\xi^k\|_\infty \leq 1$ since $\xi^k \in \partial|x^k|$, the second inequality follows from the Lipschitz continuity of ϕ'_+ (with modulus ρ); the third inequality holds because ∇f is Lipschitz continuous; and we made use of the definition of y^k and the fact that $\{\beta_k\} \subset [0, 1)$ for the last inequality. The desired conclusion now follows immediately from (4.11) and (4.12). □

We are now ready to prove convergence of the whole sequence $\{x^k\}$ generated by $\text{IRL}_1 e_1$ under suitable assumptions. Our proof is similar to standard convergence

arguments making use of KL property; see, for example, [3,4]. We include the proof for completeness.

Theorem 4.2. *Suppose that* $\sup_{k\geqslant 0}\beta_k < 1$ *and that* ϕ'_+ *is Lipschitz continuous on* $[0,\infty)$. *Suppose in addition that* H_1 *is a KL function. Let* $\{x^k\}$ *be the sequence generated by IRL*$_1$*e*$_1$ *for solving* (4.1). *Then* $\sum_{k=1}^{\infty}\|x^k - x^{k-1}\| < \infty$ *and* $\{x^k\}$ *converges to a stationary point of the problem* (4.1).

Proof. In view of Theorem 4.1(iii), it suffices to show that $\sum_{k=1}^{\infty}\|x^k - x^{k-1}\| < \infty$ (which implies convergence of $\{x^k\}$). To this end, note first from Theorem 4.1(i) that $w_1 := \lim_k H_1(x^k, x^{k-1})$ exists. If there exists k' such that $H_1(x^{k'}, x^{k'-1}) = w_1$, then for all $k \geqslant k'$, we must have $H_1(x^k, x^{k-1}) = H_1(x^{k'}, x^{k'-1}) = w_1$, thanks to the fact that $\{H_1(x^k, x^{k-1})\}$ is nonincreasing by Theorem 4.1(i). Combining this with (4.5), we obtain that $x^k = x^{k'}$ when $k \geqslant k'$, i.e. the sequence generated converges finitely and thus the conclusion of this theorem holds in this case. Thus, from now on, we assume that $H_1(x^k, x^{k-1}) > w_1$ for all k.

According to Corollary 4.1, H_1 is constant (which equals w_1) on the nonempty compact set $\Omega_1 \subseteq \operatorname{dom}\partial H$, where Ω_1 is the set of accumulation points of $\{(x^k, x^{k-1})\}$. This together with the assumption that H_1 is a KL function and Lemma 2.2 implies that there exist ϵ_1, $\eta_1 > 0$ and $\varphi_1 \in \Xi_{\eta_1}$ such that

$$\varphi_1'\left(H_1(x,y) - w_1\right)\operatorname{dist}\left(0, \partial H_1(x,y)\right) \geqslant 1$$

for any (x,y) satisfying $\operatorname{dist}((x,y),\Omega_1) < \epsilon_1$ and $w_1 < H_1(x,y) < w_1 + \eta_1$. In addition, since $\{x^k\}$ is bounded according to Theorem 4.1(ii), there exists k_0 such that whenever $k \geqslant k_0$, we have

$$\operatorname{dist}((x^k, x^{k-1}), \Omega_1) < \epsilon_1.$$

Furthermore, from the definition of w_1 and the assumption that $H_1(x^k, x^{k-1}) > w_1$ for all k, we know there exists a k_1 such that such that whenever $k \geqslant k_1$, $w_1 <$

$H_1(x^k, x^{k-1}) < w_1 + \eta_1$. Let $N = \max\{k_0, k_1\}$. Then for $k > N$, we have

$$\varphi_1'\left(H_1(x^k, x^{k-1}) - w_1\right) \operatorname{dist}\left(0, \partial H_1(x^k, x^{k-1})\right) \geqslant 1.$$

Combining this with the concavity of φ_1 we have

$$\begin{aligned}
&\underbrace{[\varphi_1\left(H_1(x^k, x^{k-1}) - w_1\right) - \varphi_1\left(H_1(x^{k+1}, x^k) - w_1\right)]}_{\Delta_k} \cdot \operatorname{dist}\left(0, \partial H_1(x^k, x^{k-1})\right) \\
&\geqslant \varphi_1'\left(H_1(x^k, x^{k-1}) - w_1\right) \cdot \operatorname{dist}\left(0, \partial H_1(x^k, x^{k-1})\right) \cdot \left(H_1(x^k, x^{k-1}) - H_1(x^{k+1}, x^k)\right) \\
&\geqslant H_1(x^k, x^{k-1}) - H_1(x^{k+1}, x^k) \geqslant D_1 \|x^k - x^{k-1}\|^2,
\end{aligned}$$

where the last inequality follows from (4.5). Using Lemma 4.2 to upper bound the term $\operatorname{dist}\left(0, \partial H_1(x^k, x^{k-1})\right)$ in the above relation, we further deduce that

$$\begin{aligned}
\|x^k - x^{k-1}\|^2 &\leqslant \frac{4C_1}{D_1}\Delta_k \cdot \frac{1}{4}\left(\|x^{k-1} - x^{k-2}\| + \|x^k - x^{k-1}\|\right) \\
&\leqslant \left[\frac{C_1}{D_1}\Delta_k + \frac{1}{4}\left(\|x^{k-1} - x^{k-2}\| + \|x^k - x^{k-1}\|\right)\right]^2,
\end{aligned}$$

where the second inequality follows the relation $4ab \leqslant (a+b)^2$ for a, $b \in \mathbb{R}$. Taking square root on both sides of the above inequality and rearranging terms, we have

$$\frac{1}{2}\|x^k - x^{k-1}\| \leqslant \frac{C_1}{D_1}\Delta_k + \frac{1}{4}\left(\|x^{k-1} - x^{k-2}\| - \|x^k - x^{k-1}\|\right).$$

Summing this inequality from $k = N+1$ to ∞, we obtain that

$$\frac{1}{2}\sum_{k=N+1}^{\infty} \|x^k - x^{k-1}\| \leqslant \frac{C_1}{D_1}\left(\varphi_1\left(H_1(x^{N+1}, x^N) - w_1\right)\right) + \frac{1}{4}\left(\|x^N - x^{N-1}\|\right) < \infty,$$

which also implies the convergence of $\{x^k\}$. This completes the proof. □

4.4 Iteratively reweighted ℓ_1 algorithm with type-II extrapolation

In this section, we propose and analyze another version of iteratively reweighted ℓ_1 algorithm with an extrapolation technique motivated from the method by Auslender

and Teboulle [6]: this was described as the second APG method in [90] and was shown empirically to be the most efficient optimal first-order method in the numerical experiments of [11]. We call the algorithm based on this extrapolation technique the iteratively reweighted ℓ_1 algorithm with type-II extrapolation (IRL$_1e_2$). This method is presented as Algorithm 4.4 below.

Iteratively reweighted ℓ_1 algorithm with type-II extrapolation (IRL$_1e_2$)

Step 0. Input initial points $x^0, z^0 \in C$ and a sequence $\{\theta_k\} \subset (0, 1]$. Set $k = 0$.

Step 1. Set

$$\begin{aligned} s^{k+1} &= \Phi'_+(|x^k|); \\ y^k &= (1-\theta_k)x^k + \theta_k z^k; \\ z^{k+1} &= \arg\min_{x\in C} \left\{ \langle \nabla f(y^k), x - y^k\rangle + \frac{L\theta_k}{2}\|x - z^k\|^2 + \sum_{i=1}^n s_i^{k+1}|x_i| \right\}; \\ x^{k+1} &= (1-\theta_k)x^k + \theta_k z^{k+1}. \end{aligned} \tag{4.13}$$

Step 2. If a termination criterion is not met, set $k = k + 1$ and go to Step 1.

We will show that any accumulation point of the sequence $\{z^k\}$ generated by IRL$_1e_2$ is a stationary point of F under suitable assumptions. Our convergence arguments also make use of H_1 defined in (4.4), and are inspired by [90, Appendix A]. In our analysis below, the parameters $\{\theta_k\}$ in IRL$_1e_2$ have to satisfy (4.14). We will demonstrate in Section 4.6 how such $\{\theta_k\}$ can be chosen in our numerical experiments.

Theorem 4.3. *Suppose that the $\{\theta_k\}$ in IRL$_1e_2$ is chosen so that*

$$\sup_{k\geqslant 1}\{\theta_k^2(1-\theta_{k-1})^2 - \theta_{k-1}^2\} < 0, \tag{4.14}$$

and let $\{x^k, y^k, z^k\}$ be the sequences generated by IRL$_1e_2$ for solving (4.1)*. Then the following statements hold.*

(i) $\{H_1(x^k, x^{k-1})\}_{k\geqslant 1}$ *is a nonincreasing convergent sequence.*

(ii) *It holds that*

$$\lim_k \max\{\|z^{k+1} - x^k\|, \|z^{k+1} - y^k\|, \|z^{k+1} - z^k\|\} = 0. \tag{4.15}$$

(iii) *The sequence* $\{z^k\}$ *is bounded.*

(iv) *Any accumulation point of* $\{z^k\}$ *is a stationary point of* (4.1).

Proof. In this proof, we write $l_f(x; y) := f(y) + \langle \nabla f(y), x - y\rangle$ for notational simplicity. Since ∇f is Lipschitz continuous with modulus $L > 0$, we have

$$\begin{aligned}
F(x^{k+1}) &\leqslant l_f(x^{k+1}; y^k) + \frac{L}{2}\|x^{k+1} - y^k\|^2 + \Phi(|x^{k+1}|) \\
&= l_f(x^{k+1}; y^k) + \frac{L\theta_k^2}{2}\|z^{k+1} - z^k\|^2 + \Phi(|x^{k+1}|) \\
&= (1-\theta_k) l_f(x^k; y^k) + \theta_k l_f(z^{k+1}; y^k) + \frac{L\theta_k^2}{2}\|z^{k+1} - z^k\|^2 + \Phi(|x^{k+1}|) \\
&= (1-\theta_k) l_f(x^k; y^k) + \theta_k \left[l_f(z^{k+1}; y^k) + \frac{L\theta_k}{2}\|z^{k+1} - z^k\|^2 + \sum_{i=1}^n s_i^{k+1}|z_i^{k+1}| \right] \\
&\quad + \sum_{i=1}^n \left[\phi(|x_i^{k+1}|) - \theta_k s_i^{k+1}|z_i^{k+1}|\right] \\
&\leqslant l_f(x^k; y^k) + \theta_k \left[\frac{L\theta_k}{2}\|x^k - z^k\|^2 + \sum_{i=1}^n s_i^{k+1}|x_i^k| - \frac{L\theta_k}{2}\|x^k - z^{k+1}\|^2 \right] \\
&\quad + \sum_{i=1}^n \left[\phi(|x_i^{k+1}|) - \theta_k s_i^{k+1}|z_i^{k+1}|\right],
\end{aligned} \tag{4.16}$$

where the first equality follows from the definitions of x^{k+1} and y^k in (4.13) so that

$$x^{k+1} - y^k = [(1-\theta_k)x^k + \theta_k z^{k+1}] - [(1-\theta_k)x^k + \theta_k z^k] = \theta_k(z^{k+1} - z^k),$$

the second equality follows from the definition of x^{k+1}, and the last inequality follows from the definition of z^{k+1} as a minimizer and the strong convexity of the objective of the minimization problem defining z^{k+1}.

From the convexity of f and (4.16), we see further that

$$\begin{aligned} F(x^{k+1}) &\leqslant f(x^k) + \theta_k \left[\frac{L\theta_k}{2}\|x^k - z^k\|^2 + \sum_{i=1}^n s_i^{k+1}|x_i^k| - \frac{L\theta_k}{2}\|x^k - z^{k+1}\|^2 \right] \\ &\quad + \sum_{i=1}^n \left[\phi(|x_i^{k+1}|) - \theta_k s_i^{k+1}|z_i^{k+1}|\right], \\ &\leqslant f(x^k) + \theta_k \left[\frac{L\theta_k}{2}\|x^k - z^k\|^2 - \frac{L\theta_k}{2}\|x^k - z^{k+1}\|^2 \right] \\ &\quad + \sum_{i=1}^n \left[\phi(|x_i^k|) + s_i^{k+1}(|x_i^{k+1}| - |x_i^k|) - \theta_k s_i^{k+1}|z_i^{k+1}| + \theta_k s_i^{k+1}|x_i^k|\right], \end{aligned} \tag{4.17}$$

where the second inequality follows from the fact that $s^{k+1} = \Phi'_+(|x^k|)$ and the concavity of ϕ.

Next, observe from the last relation in (4.13) that for each $i = 1, \ldots, n$,

$$\begin{aligned} & |x_i^{k+1}| = |(1-\theta_k)x_i^k + \theta_k z_i^{k+1}|, \\ \implies \quad & |x_i^{k+1}| \leqslant (1-\theta_k)|x_i^k| + \theta_k|z_i^{k+1}|, \\ \implies \quad & |x_i^{k+1}| - |x_i^k| - \theta_k|z_i^{k+1}| + \theta_k|x_i^k| \leqslant 0. \end{aligned}$$

Combining this with (4.17) and the nonnegativity of s_i^k, we obtain further that for all $k \geqslant 1$ that

$$\begin{aligned} F(x^{k+1}) &\leqslant f(x^k) + \Phi(|x^k|) + \theta_k \left[\frac{L\theta_k}{2}\|x^k - z^k\|^2 - \frac{L\theta_k}{2}\|x^k - z^{k+1}\|^2 \right] \\ &= F(x^k) + \frac{L\theta_k^2(1-\theta_{k-1})^2}{2}\|x^{k-1} - z^k\|^2 - \frac{L\theta_k^2}{2}\|x^k - z^{k+1}\|^2, \end{aligned} \tag{4.18}$$

where the last equality follows from the last relation in (4.13).

Observe that for $k \geqslant 0$, $x^k \in C$ and $\theta_k(z^{k+1} - x^k) = x^{k+1} - x^k$. Using these and the definition of H_1, we obtain from (4.18) that for $k \geqslant 1$,

$$\begin{aligned} H_1(x^{k+1}, x^k) - H_1(x^k, x^{k-1}) &\leqslant \left(\frac{L\theta_k^2(1-\theta_{k-1})^2}{2} - \frac{L\theta_{k-1}^2}{2} \right) \|x^{k-1} - z^k\|^2 \\ &\leqslant -A_1 \|x^{k-1} - z^k\|^2, \end{aligned} \tag{4.19}$$

where $A_1 := \frac{L}{2} \inf_k \{\theta_{k-1}^2 - \theta_k^2(1-\theta_{k-1})^2\}$, which is positive thanks to (4.14). Thus, $\{H_1(x^k, x^{k-1})\}_{k\geqslant 1}$ is nonincreasing.

In addition, since $x^k \in C$, we have

$$v \leq F(x^k) \leqslant F(x^k) + \frac{L}{2}\|x^k - x^{k-1}\|^2 = H_1(x^k, x^{k-1}),$$

showing that $\{H_1(x^k, x^{k-1})\}$ is bounded from below. Thus, $\{H_1(x^k, x^{k-1})\}_{k\geqslant 1}$ is convergent. This proves (i).

Next, summing (4.19) from $k = 1$ to ∞, we have

$$A_1 \sum_{k=1}^{\infty} \|x^{k-1} - z^k\|^2 \leqslant H_1(x^1, x^0) - \lim_k H_1(x^{k+1}, x^k) < \infty.$$

Therefore, $\lim_k \|x^k - z^{k+1}\| = 0$, which further implies that

$$\begin{aligned} \lim_k x^{k+1} - x^k &= \lim_k \theta_k(z^{k+1} - x^k) = 0; \\ \lim_k x^{k+1} - z^{k+1} &= \lim_k (1-\theta_k)(x^k - z^{k+1}) = 0; \\ \lim_k x^{k+1} - y^{k+1} &= \lim_k \theta_{k+1}(x^{k+1} - z^{k+1}) = 0; \end{aligned} \tag{4.20}$$

where the first and second equalities are due to the last relation in (4.13) and the third equality is due to the second relation in (4.13). Then we have

$$\begin{aligned} \lim_k z^{k+1} - z^k &= \lim_k (z^{k+1} - x^k) + (x^k - z^k) = 0, \\ \lim_k z^{k+1} - y^k &= \lim_k (z^{k+1} - x^k) + (x^k - y^k) = 0. \end{aligned}$$

This proves (ii).

We now prove (iii). Notice from (4.19) and the definition of H_1 that

$$F(x^k) \leqslant H_1(x^k, x^{k-1}) \leqslant H_1(x^1, x^0) = F(x^1) + \frac{L}{2}\|x^1 - x^0\|^2 < \infty.$$

Since F is level-bounded, we conclude from this inequality that $\{x^k\}$ is bounded. In view of this and the second equality in (4.20), we conclude that $\{z^k\}$ is also bounded, i.e., (iii) holds.

Now we prove (iv). Let z^* be an accumulation point of $\{z^k\}$ and let $\{z^{k_j}\}$ be a subsequence such that $z^{k_j} \to z^*$. Clearly $z^* \in C$. From (4.15), we know that

$$z^{k_j-1} \to z^*, \quad y^{k_j-1} \to z^*, \quad x^{k_j-1} \to z^*. \tag{4.21}$$

Using the definition of z^k as the minimizer of the optimization problem in (4.13) and the subdifferential calculus rules in [79, Proposition 10.5] and [79, Proposition 10.9], we have

$$0 \in \nabla f(y^{k_j-1}) + N_C(z^{k_j}) + \theta_{k_j-1}L(z^{k_j} - z^{k_j-1}) + s^{k_j} \circ \partial|z^{k_j}|. \tag{4.22}$$

Next we show that

$$\lim_j s^{k_j} = \Phi'_+(|z^*|). \tag{4.23}$$

We consider two cases. First, for those i satisfying $z_i^* \neq 0$, we have from the definition of s^k and (4.21) that $\lim_j s_i^{k_j} = \phi'_+(|z_i^*|)$. On the other hand, for those i corresponding to $z_i^* = 0$, we have by the definition of s^k that $\lim_j s_i^{k_j} = \lim_j \phi'_+(|x_i^{k_j-1}|) = \ell = \phi'_+(0)$, thanks to Lemma 2.4(i). Therefore, $\lim_j s^{k_j} = \Phi'_+(|z^*|)$.

Now, notice that the set-valued mapping $x \rightrightarrows \Phi'_+(x) \circ \partial|x|$ is bounded because $\Phi'_+(x) \circ \partial|x| \subseteq [-\ell, \ell]^n$ for all $x \in \mathbb{R}^n$. Using this, [79, Proposition 5.51], the closedness of convex subdifferentials, (4.23) and (4.21), we see by passing to the limit in (4.22) that

$$0 \in \nabla f(z^*) + N_C(z^*) + \Phi'_+(|z^*|) \circ \partial|z^*| = \partial F(z^*),$$

where the equality follows from Lemma 4.1. This completes the proof. □

4.5 Iteratively reweighted ℓ_1 algorithm with type-III extrapolation

In this section, we propose and analyze yet another version of iteratively reweighted ℓ_1 algorithm with an extrapolation technique motivated from the method by Lan, Lu and Monteiro [53]: this was stated as algorithm LLM in [11] and was the first of its kinds whose complexity has been established in some nonconvex settings [36, 43]. We refer to the algorithm based on this extrapolation technique as the iteratively reweighted ℓ_1 algorithm with type-III extrapolation (IRL$_1e_3$). The method is presented as Algorithm 4.5 below.

Iteratively reweighted ℓ_1 algorithm with type-III extrapolation (IRL$_1e_3$)

Step 0. Input initial points $x^0, z^0 \in C$ and a sequence $\{\theta_k\} \subset (0, 1]$. Set $k = 0$.

Step 1. Set

$$\begin{aligned}
&s^{k+1} = \Phi'_+(|x^k|);\\
&y^k = (1-\theta_k)x^k + \theta_k z^k;\\
&z^{k+1} = \arg\min_{x\in C}\left\{\left\langle \nabla f(y^k), x - y^k\right\rangle + \frac{L\theta_k}{2}\|x - z^k\|^2 + \sum_{i=1}^n s_i^{k+1}|x_i|\right\};\\
&x^{k+1} = \arg\min_{y\in C}\left\{\left\langle \nabla f(y^k), y - y^k\right\rangle + \frac{L}{2}\|y - y^k\|^2 + \sum_{i=1}^n s_i^{k+1}|y_i|\right\}.
\end{aligned} \tag{4.24}$$

Step 2. If a termination criterion is not met, set $k = k + 1$ and go to Step 1.

Now we present convergence analysis for this algorithm. Our analysis is inspired by [36, Section 8] and relies heavily on the following auxiliary function:

$$H_3(x, y, w) = f(x) + \delta_C(x) + \Phi(|x|) + \frac{L}{2}\|w - y\|^2 + \frac{L}{2}\|w - x\|^2.$$

We start by characterizing the accumulation points of the sequence generated by $\mathrm{IRL}_1 e_3$ under suitable assumptions on $\{\theta_k\}$, and establish the convergence of the whole sequence under additional assumptions. In particular, we require $\{\theta_k\}$ to be chosen so that (4.25) holds: we will demonstrate how such $\{\theta_k\}$ can be chosen to satisfy this condition in our numerical experiments in Section 4.6.

Theorem 4.4. *Suppose that $\{\theta_k\}$ in IRL$_1e_3$ is chosen so that for some $\gamma \in (0,1)$,*

$$\sup_{k\geqslant 1} \max\left\{ \frac{\theta_k^2(1-\theta_{k-1})^2}{\gamma} - \theta_{k-1}^2, \frac{\theta_k^2}{1-\gamma} - 1 \right\} < 0. \tag{4.25}$$

Let $\{x^k, y^k, z^k\}$ be the sequences generated by IRL$_1e_3$ for solving (4.1) and define $w^{k+1} := (1-\theta_k)x^k + \theta_k z^{k+1}$ for $k \geqslant 0$. Then the following statements hold:

(i) *$\{H_3(x^k, x^{k-1}, w^k)\}_{k\geqslant 1}$ is a nonincreasing convergent sequence. Moreover, there exists a positive constant D_3 such that*

$$H_3(x^k, x^{k-1}, w^k) - H_3(x^{k+1}, x^k, w^{k+1}) \geqslant D_3(\|x^{k-1} - z^k\|^2 + \|w^k - x^k\|^2). \tag{4.26}$$

(ii) *The sequence $\{x^k\}$ is bounded.*

(iii) *It holds that*

$$\lim_k \max\{\|x^{k-1} - z^k\|, \|w^k - x^k\|, \|x^k - x^{k-1}\|, \|x^k - y^{k-1}\|\} = 0. \tag{4.27}$$

(iv) *Any accumulation point of $\{x^k\}$ is a stationary point of (4.1).*

Proof. In this proof, we write $l_f(x;y) := f(y) + \langle \nabla f(y), x - y\rangle$ for notational sim-

plicity. Since ∇f is Lipschitz, we have

$$
\begin{aligned}
F(x^{k+1}) &\leqslant l_f(x^{k+1}; y^k) + \frac{L}{2}\|x^{k+1} - y^k\|^2 + \Phi(|x^{k+1}|) \\
&= l_f(x^{k+1}; y^k) + \frac{L}{2}\|x^{k+1} - y^k\|^2 + \sum_{i=1}^{n} s_i^{k+1}|x_i^{k+1}| + \Phi(|x^{k+1}|) - \sum_{i=1}^{n} s_i^{k+1}|x_i^{k+1}| \\
&\leqslant l_f(w^{k+1}; y^k) + \frac{L}{2}\|w^{k+1} - y^k\|^2 + \sum_{i=1}^{n} s_i^{k+1}|w_i^{k+1}| + \Phi(|x^{k+1}|) - \sum_{i=1}^{n} s_i^{k+1}|x_i^{k+1}| \\
&\quad - \frac{L}{2}\|w^{k+1} - x^{k+1}\|^2,
\end{aligned}
$$

where $w^{k+1} = (1-\theta_k)x^k + \theta_k z^{k+1}$, and the second inequality follows from the definition of x^{k+1} as the minimizer of the strongly convex subproblem for the x-update. Plugging the definition of w^{k+1} into the first three terms in the last inequality above and invoking the definition of y^k, we see that

$$
\begin{aligned}
F(x^{k+1}) &\leqslant (1-\theta_k) l_f(x^k; y^k) + \theta_k l_f(z^{k+1}; y^k) + \frac{L\theta_k^2}{2}\|z^{k+1} - z^k\|^2 + \Phi(|x^{k+1}|) \\
&\qquad + \sum_{i=1}^{n} s_i^{k+1}|(1-\theta_k)x_i^k + \theta_k z_i^{k+1}| - \sum_{i=1}^{n} s_i^{k+1}|x_i^{k+1}| - \frac{L}{2}\|w^{k+1} - x^{k+1}\|^2.
\end{aligned}
$$

Applying the relation $|(1-\theta_k)x_i^k + \theta_k z_i^{k+1}| \leq (1-\theta_k)|x_i^k| + \theta_k|z_i^{k+1}|$ to the inequality

above and grouping terms, we obtain further that $F(x^{k+1})$ is bounded above by

$$
\begin{aligned}
&(1-\theta_k) l_f(x^k; y^k) + \theta_k \left[l_f(z^{k+1}; y^k) + \frac{L\theta_k}{2}\|z^{k+1} - z^k\|^2 + \sum_{i=1}^n s_i^{k+1}|z_i^{k+1}| \right] \\
&+ \sum_{i=1}^n (1-\theta_k) s_i^{k+1}|x_i^k| + \Phi(|x^{k+1}|) - \sum_{i=1}^n s_i^{k+1}|x_i^{k+1}| - \frac{L}{2}\|w^{k+1} - x^{k+1}\|^2 \\
&\leqslant l_f(x^k; y^k) + \theta_k \left[\frac{L\theta_k}{2}\|x^k - z^k\|^2 + \sum_{i=1}^n s_i^{k+1}|x_i^k| \right] - \frac{L\theta_k^2}{2}\|x^k - z^{k+1}\|^2 \\
&\quad + \sum_{i=1}^n (1-\theta_k) s_i^{k+1}|x_i^k| + \Phi(|x^{k+1}|) - \sum_{i=1}^n s_i^{k+1}|x_i^{k+1}| - \frac{L}{2}\|w^{k+1} - x^{k+1}\|^2 \\
&= l_f(x^k; y^k) + \sum_{i=1}^n \left[s_i^{k+1}|x_i^k| + \phi(|x_i^{k+1}|) - s_i^{k+1}|x_i^{k+1}| \right] + \frac{L\theta_k^2}{2}\|x^k - z^k\|^2 \\
&\quad - \frac{L}{2}\|w^{k+1} - x^{k+1}\|^2 - \frac{L\theta_k^2}{2}\|x^k - z^{k+1}\|^2,
\end{aligned}
$$

where the inequality follows from the definition of z^{k+1} as the minimizer of the strongly convex subproblem for the z-update.

Applying the convexity of f, the concavity of ϕ and the fact that $s^{k+1} = \Phi'_+(|x^k|)$ to the above upper bound, we further have

$$
\begin{aligned}
F(x^{k+1}) &\leqslant f(x^k) + \sum_{i=1}^n \left[\phi(|x_i^k|) + s_i^{k+1}(|x_i^{k+1}| - |x_i^k|) - s_i^{k+1}|x_i^{k+1}| + s_i^{k+1}|x_i^k| \right] \\
&\quad + \frac{L\theta_k^2}{2}\|x^k - z^k\|^2 - \frac{L}{2}\|w^{k+1} - x^{k+1}\|^2 - \frac{L\theta_k^2}{2}\|x^k - z^{k+1}\|^2 \\
&= F(x^k) + \frac{L\theta_k^2}{2}\|x^k - z^k\|^2 - \frac{L}{2}\|w^{k+1} - x^{k+1}\|^2 - \frac{L\theta_k^2}{2}\|x^k - z^{k+1}\|^2.
\end{aligned} \tag{4.28}
$$

Now, observe that $w^k - x^k = (1-\theta_{k-1})x^{k-1} + \theta_{k-1}z^k - x^k = x^{k-1} - x^k + \theta_{k-1}(z^k - x^{k-1})$ for any $k \geqslant 1$, we thus have

$$
x^k - z^k = x^k - x^{k-1} + x^{k-1} - z^k = (1-\theta_{k-1})(x^{k-1} - z^k) - (w^k - x^k).
$$

Using this and the inequality that $(a+b)^2 \leqslant \frac{a^2}{\gamma} + \frac{b^2}{1-\gamma}$, where $\gamma \in (0,1)$ is as in (4.25), we deduce further from (4.28) that $F(x^{k+1})$ is bounded above by

$$\begin{aligned}
&F(x^k) + \frac{L\theta_k^2(1-\theta_{k-1})^2}{2\gamma}\|x^{k-1}-z^k\|^2 + \frac{L\theta_k^2}{2(1-\gamma)}\|w^k-x^k\|^2 \\
&- \frac{L}{2}\|w^{k+1}-x^{k+1}\|^2 - \frac{L\theta_k^2}{2}\|x^k-z^{k+1}\|^2 \\
&= F(x^k) + \frac{L}{2}\big(\theta_{k-1}^2\|x^{k-1}-z^k\|^2 + \|w^k-x^k\|^2 - \|w^{k+1}-x^{k+1}\|^2 - \theta_k^2\|x^k-z^{k+1}\|^2\big) \\
&\quad + \left(\frac{\theta_k^2(1-\theta_{k-1})^2}{\gamma} - \theta_{k-1}^2\right)\frac{L}{2}\|x^{k-1}-z^k\|^2 + \left(\frac{\theta_k^2}{1-\gamma} - 1\right)\frac{L}{2}\|w^k-x^k\|^2.
\end{aligned}$$

Using the assumptions on θ_k, we then obtain the following estimate:

$$\begin{aligned}
F(x^{k+1}) &\leqslant F(x^k) - A_2(\|x^{k-1}-z^k\|^2 + \|w^k-x^k\|^2) \\
&\quad + \frac{L}{2}\big(\theta_{k-1}^2\|x^{k-1}-z^k\|^2 + \|w^k-x^k\|^2 - \|w^{k+1}-x^{k+1}\|^2 - \theta_k^2\|x^k-z^{k+1}\|^2\big) \\
&= F(x^k) - A_2(\|x^{k-1}-z^k\|^2 + \|w^k-x^k\|^2) \\
&\quad + \frac{L}{2}\big(\|w^k-x^{k-1}\|^2 + \|w^k-x^k\|^2 - \|w^{k+1}-x^{k+1}\|^2 - \|w^{k+1}-x^k\|^2\big),
\end{aligned}$$

where $A_2 = \frac{L}{2}\inf_k \min\left\{\theta_{k-1}^2 - \frac{\theta_k^2(1-\theta_{k-1})^2}{\gamma}, 1 - \frac{\theta_k^2}{1-\gamma}\right\}$, which is positive according to (4.25), and the equality follows from the definition of w^k so that $w^k - x^{k-1} = \theta_{k-1}(z^k - x^{k-1})$. Rearranging terms in the above inequality and invoking the definition of H_3, we have for all $k \geqslant 1$ that

$$A_2(\|x^{k-1}-z^k\|^2 + \|w^k-x^k\|^2) \leqslant H_3(x^k, x^{k-1}, w^k) - H_3(x^{k+1}, x^k, w^{k+1}), \quad (4.29)$$

which means that $\{H_3(x^k, x^{k-1}, w^k)\}_{k\geqslant 1}$ is nonincreasing. In addition, it is not hard to see that $\{H_3(x^k, x^{k-1}, w^k)\}$ is bounded from below. Thus, the sequence $\{H_3(x^k, x^{k-1}, w^k)\}$ is convergent. This proves (i).

Next, we have from (4.29) that for any $k \geqslant 1$ that

$$F(x^k) \leqslant H_3(x^k, x^{k-1}, w^k) \leqslant H_3(x^1, x^0, w^1) < \infty,$$

Since F is level-bounded, we conclude from this inequality that $\{x^k\}$ is bounded and therefore (ii) holds.

We now prove (iii). Summing (4.29) from $k = 1$ to ∞, we obtain

$$A_2 \sum_{k=1}^{\infty} (\|x^{k-1} - z^k\|^2 + \|w^k - x^k\|^2) \leqslant H_3(x^1, x^0, w^1) - \lim_k H_3(x^{k+1}, x^k, w^{k+1}) < \infty.$$

Thus, we have

$$\lim_k \|x^{k-1} - z^k\| = \lim_k \|w^k - x^k\| = 0. \tag{4.30}$$

Combining these relations with the definition of w^k, we have

$$w^k - x^{k-1} = \theta_{k-1}(z^k - x^{k-1}) \to 0. \tag{4.31}$$

Combining this with (4.30), we see further that

$$x^k - x^{k-1} = x^k - w^k + w^k - x^{k-1} = (x^k - w^k) + \theta_{k-1}(z^k - x^{k-1}) \to 0. \tag{4.32}$$

Combining this with the definition of y^k and (4.30), we obtain

$$\begin{aligned} y^k - x^k &= \theta_k(z^k - x^k) = \theta_k(z^k - x^{k-1}) + \theta_k(x^{k-1} - x^k) \\ &= \theta_k(z^k - x^{k-1}) + \theta_k \left[(w^k - x^k) + \theta_{k-1}(x^{k-1} - z^k)\right] \\ &= (\theta_k - \theta_k\theta_{k-1})(z^k - x^{k-1}) + \theta_k(w^k - x^k) \to 0. \end{aligned} \tag{4.33}$$

Using this together with (4.32) and (4.30), we deduce that

$$\begin{aligned} &w^k - y^{k-1} = w^k - x^k + (x^k - x^{k-1}) + (x^{k-1} - y^{k-1}) \\ &= \theta_{k-1}(z^k - x^{k-1}) + (\theta_{k-1}\theta_{k-2} - \theta_{k-1})(z^{k-1} - x^{k-2}) + \theta_{k-1}(x^{k-1} - w^{k-1}) \to 0. \end{aligned} \tag{4.34}$$

Finally, using (4.34), we have

$$\begin{aligned} &x^k - y^{k-1} = x^k - w^k + w^k - y^{k-1} \\ &= (x^k - w^k) + \theta_{k-1}(z^k - x^{k-1}) + (\theta_{k-1}\theta_{k-2} - \theta_{k-1})(z^{k-1} - x^{k-2}) + \theta_{k-1}(x^{k-1} - w^{k-1}), \end{aligned}$$

which also goes to zero thanks to (4.30). This proves (iii).

Finally, we prove (iv). Let $\widetilde{x}$ be an accumulation point of $\{x^k\}$ and let $\{x^{k_j}\}$ be a subsequence such that $x^{k_j} \to \widetilde{x}$. From the first-order optimality condition of the second subproblem in (4.24) and subdifferential calculus rules in [79, Proposition 10.5] and [79, Proposition 10.9], we have

$$0 \in \nabla f(y^{k_j}) + N_C(x^{k_j+1}) + L(x^{k_j+1} - y^{k_j}) + s^{k_j+1} \circ \partial|x^{k_j+1}|. \tag{4.35}$$

Using the same arguments as in the proof of Theorem 4.1(iv), we have $\lim_j s^{k_j+1} = \Phi'_+(|\widetilde{x}|)$.

Now, observe that $\Phi'_+(x) \circ \partial|x| \subseteq [-\ell, \ell]^n$ so that the set-valued mapping $x \rightrightarrows \Phi'_+(x) \circ \partial|x|$ is bounded. Using these, (4.27), the closedness of convex subdifferentials and [79, Proposition 5.51], passing to the limit as j goes to ∞ in (4.35), we have

$$0 \in \nabla f(\widetilde{x}) + N_C(\widetilde{x}) + \Phi'_+(|\widetilde{x}|) \circ \partial|\widetilde{x}| = \partial F(\widetilde{x}),$$

where the last equality follows from Lemma 4.1. Thus (iv) holds and this completes the proof. □

Corollary 4.2. *Suppose that the $\{\theta_k\}$ in IRL$_1$e$_3$ is chosen so that (4.25) holds. Let $\{x^k, z^k\}$ be the sequences generated by IRL$_1$e$_3$ for solving (4.1) and define $w^{k+1} := (1-\theta_k)x^k + \theta_k z^{k+1}$ for $k \geqslant 0$. Then the set of accumulation points of $\{(x^k, x^{k-1}, w^k)\}$, denoted by Ω_3, is a nonempty compact subset of $\operatorname{dom} \partial H_3$. Moreover, it holds that $H_3 \equiv \lim_k H_3(x^k, x^{k-1}, w^k)$ on Ω_3.*

Proof. First, we see from Theorem 4.4(ii) that the set of accumulation points of $\{x^k\}$, denoted by Λ_3, is nonempty and compact. Moreover, in view of Theorem 4.4(iii), we deduce that $\Omega_3 = \{(x, x, x) : \ x \in \Lambda_3\}$, which is clearly nonempty and compact. Finally, since $\Lambda_3 \subseteq \{x : \ 0 \in \partial F(x)\} \subseteq \operatorname{dom} \partial F$ according to Theorem 4.4(iv), it is routine to check that $\Omega_3 \subset \operatorname{dom} \partial H_3$ as required.

Now, fix any $(\widetilde{x}, \widetilde{x}, \widetilde{x}) \in \Omega_3$ and let $\{x^{k_j}\}$ be a subsequence of $\{x^k\}$ with $x^{k_j} \to \widetilde{x}$.

Then

$$H_3(\widetilde{x},\widetilde{x},\widetilde{x}) = F(\widetilde{x}) = \lim_j F(x^{k_j}) + \frac{L}{2}\|w^{k_j} - x^{k_j}\|^2 + \frac{L}{2}\|w^{k_j} - x^{k_j-1}\|^2$$

$$= \lim_j H_3(x^{k_j}, x^{k_j-1}, w^{k_j}),$$

where the second equality follows from the continuity of F on C and Theorem 4.4(iii). Since $\{H_3(x^k, x^{k-1}, w^k)\}$ is convergent according to Theorem 4.4(i) and $(\widetilde{x},\widetilde{x},\widetilde{x}) \in \Omega_3$ is chosen arbitrarily, we conclude that $H_3 \equiv \lim_k H_3(x^k, x^{k-1}, w^k)$ on Ω_3. This completes the proof. □

Next, we show under some assumptions on H_3 and ϕ'_+ that the sequence $\{x^k\}$ generated by $\mathrm{IRL}_1 e_3$ converges to a stationary point of (4.1). We first prove the following auxiliary lemma.

Lemma 4.3. *Suppose that the $\{\theta_k\}$ in IRL_1e_3 is chosen so that (4.25) holds and that ϕ'_+ is Lipschitz continuous. Let $\{x^k, z^k\}$ be the sequences generated by IRL_1e_3 for solving (4.1) and define $w^{k+1} := (1-\theta_k)x^k + \theta_k z^{k+1}$ for $k \geqslant 0$. Then there exists a positive constant C_3 such that for all $k \geqslant 2$,*

$$\operatorname{dist}\big((0,0,0), \partial H_3(x^k, x^{k-1}, w^k)\big)$$

$$\leqslant C_3(\|x^k - w^k\| + \|z^k - x^{k-1}\| + \|x^{k-1} - w^{k-1}\| + \|z^{k-1} - x^{k-2}\|).$$

Proof. First, using the optimality condition of the x-update in (4.24) and the definition of s^k, there exist a $\xi^k \in \partial|x^k|$ and a $\zeta^k \in N_C(x^k)$ such that for all $k \geqslant 2$,

$$0 = \nabla f(y^{k-1}) + \zeta^k + \Phi'_+(|x^{k-1}|) \circ \xi^k + L(x^k - y^{k-1}). \tag{4.36}$$

Define $\eta^k := \nabla f(x^k) + \zeta^k + \Phi'_+(|x^k|) \circ \xi^k + L(x^k - w^k)$. Then we have

$$(\eta^k, -L(w^k - x^{k-1}), L(w^k - x^{k-1}) + L(w^k - x^k))$$

$$\in \begin{pmatrix} \nabla f(x^k) + N_C(x^k) + \Phi'_+(|x^k|)\partial|x^k| + L(x^k - w^k) \\ \{-L(w^k - x^{k-1})\} \\ \{L(w^k - x^{k-1}) + L(w^k - x^k)\} \end{pmatrix}$$

$$= \partial H_3(x^k, x^{k-1}, w^k);$$

here the equality follows from [79, Exercise 8.8], [79, Proposition 10.5] and Lemma 4.1. Hence, there exists a $C_0 > 0$ so that for all $k \geqslant 1$,

$$\operatorname{dist}\big((0,0,0), \partial H_3(x^k, x^{k-1}, w^k)\big) \leqslant C_0(\|\eta^k\| + \|w^k - x^{k-1}\| + \|w^k - x^k\|). \quad (4.37)$$

Next, from the definition of η^k and (4.36), we see further that

$$\begin{aligned}
\|\eta^k\| &= \|\eta^k - \big[\nabla f(y^{k-1}) + \zeta^k + \Phi'_+(|x^{k-1}|) \circ \xi^k + L(x^k - y^{k-1})\big]\| \\
&= \|\nabla f(x^k) - \nabla f(y^{k-1}) + \big[\Phi'_+(|x^k|) - \Phi'_+(|x^{k-1}|)\big] \circ \xi^k - L(w^k - y^{k-1})\| \\
&\leqslant \|\nabla f(x^k) - \nabla f(y^{k-1})\| + \|\Phi'_+(|x^k|) - \Phi'_+(|x^{k-1}|)\| + L\|w^k - y^{k-1}\| \\
&\leqslant \|\nabla f(x^k) - \nabla f(y^{k-1})\| + \sqrt{\sum_{i=1}^n \rho^2(|x_i^k| - |x_i^{k-1}|)^2} + L\|w^k - y^{k-1}\| \\
&\leqslant L\|x^k - y^{k-1}\| + \rho\|x^k - x^{k-1}\| + L\|w^k - y^{k-1}\| \\
&\leqslant [L + \rho]\|x^k - x^{k-1}\| + L\|x^{k-1} - y^{k-1}\| + L\|w^k - y^{k-1}\|,
\end{aligned} \quad (4.38)$$

where the first inequality follows from the elementary inequality $\|a \circ b\| \leqslant \|b\|_\infty \|a\|$ for any $a, b \in \mathbb{R}^n$ and the fact that $\|\xi^k\|_\infty \leqslant 1$ since $\xi^k \in \partial|x^k|$; the second inequality follows from the Lipschitz continuity of ϕ'_+ (with modulus ρ); the third inequality holds because ∇f is Lipschitz continuous. The desired conclusion now follows from (4.37), (4.38) and the relations (4.31), (4.32), (4.33), (4.34), which state that $x^k - x^{k-1}$, $w^k - x^{k-1}$, $x^{k-1} - y^{k-1}$ and $w^k - y^{k-1}$ can be written as linear combinations of $z^k - x^{k-1}$, $x^k - w^k$, $z^{k-1} - x^{k-2}$, $x^{k-1} - w^{k-1}$ with coefficients at most 2. □

We will now establish the convergence of the whole sequence $\{x^k\}$ generated by IRL$_1e_3$ under some assumptions. Our analysis is similar to standard convergence analysis based on KL property; see, for example, [3, 4]. We include the proof for the convenience of the readers.

Theorem 4.5. *Suppose that the $\{\theta_k\}$ in IRL$_1e_3$ is chosen so that (4.25) holds, that H_3 is a KL function and that ϕ'_+ is Lipschitz continuous. Let $\{x^k\}$ be the*

sequence generated by IRL$_1$e$_3$ for solving (4.1). *Then* $\sum_{k=1}^{\infty} \|x^k - x^{k-1}\| < \infty$ *and* $\{x^k\}$ *converges to a stationary point of* (4.1).

Proof. In view of Theorem 4.4(iv), it suffices to prove $\sum_{k=1}^{\infty} \|x^k - x^{k-1}\| < \infty$, which implies convergence of $\{x^k\}$. Now, recall from Theorem 4.4(i) that $w_3 := \lim_k H_3(x^k, x^{k-1}, w^k)$ exists. If there exists $k' \geq 1$ such that $H_3(x^{k'}, x^{k'-1}, w^{k'}) = w_3$, then (4.26) implies that for any $k \geqslant k'$, $H_3(x^k, x^{k-1}, w^k) = H_3(x^{k'}, x^{k'-1}, w^{k'}) = w_3$. Invoking (4.26) again together with (4.32), we obtain that $x^k = x^{k'}$ when $k \geqslant k'$, i.e. the sequence generated converges finitely and hence the conclusion of this theorem holds trivially. In what follows, we consider the case where $H_3(x^k, x^{k-1}, w^k) > w_3$ for all k.

Recall from Corollary 4.2 that Ω_3 is a nonempty compact subset of $\operatorname{dom} \partial H_3$ and $H_3 \equiv w_3$. Since H_3 is a KL function, using Lemma 2.2, there exist $\epsilon_3, \eta_3 > 0$ and $\varphi_3 \in \Xi_{\eta_3}$ such that

$$\varphi_3'\left(H_3(x, y, w) - w_3\right) \operatorname{dist}\left(0, \partial H_3(x, y, w)\right) \geqslant 1$$

for any (x, y, w) satisfying $\operatorname{dist}((x, y, w), \Omega_3) < \epsilon_3$ and $w_3 < H_3(x, y, w) < w_3 + \eta_3$. In addition, recall from Corollary 4.2 that Ω_3 is the set of accumulation points of $\{(x^k, x^{k-1}, w^k)\}$. Since $\{(x^k, x^{k-1}, w^k)\}$ is bounded in view of Theorem 4.4(ii) and (iii), there exists k_0 such that whenever $k \geqslant k_0$,

$$\operatorname{dist}((x^k, x^{k-1}, w^k), \Omega_3) < \epsilon_3.$$

Furthermore, it follows from the definition of w_3 that there exists k_1 such that whenever $k \geqslant k_1$, $w_3 < H_3(x^k, x^{k-1}, w^k) < w_3 + \eta_3$. Define $N' = \max\{k_0, k_1\}$. Then for all $k > N'$, we have

$$\varphi_3'(\underbrace{H_3(x^k, x^{k-1}, w^k) - w_3}_{\chi_k}) \cdot \operatorname{dist}\left(0, \partial H_3(x^k, x^{k-1}, w^k)\right) \geqslant 1.$$

Using this and the concavity of φ_3, we have further that

$$\begin{aligned}
&\underbrace{(\varphi_3(\chi_k) - \varphi_3(\chi_{k+1}))}_{\Delta_k} \cdot \operatorname{dist}\left(0, \partial H_3(x^k, x^{k-1}, w^k)\right) \\
&\geqslant \varphi_3'(\chi_k) \cdot \operatorname{dist}\left(0, \partial H_3(x^k, x^{k-1}, w^k)\right) \cdot (\chi_k - \chi_{k+1}) \\
&\geqslant H_3(x^k, x^{k-1}, w^k) - H_3(x^{k+1}, x^k, w^{k+1}) \geqslant D_3\left(\|x^{k-1} - z^k\|^2 + \|w^k - x^k\|^2\right),
\end{aligned}$$

where the last inequality is from (4.26). Now, using the relations that $(a+b)^2 \leqslant 2(a^2+b^2)$ and $4ab \leqslant (a+b)^2$ for $a,\, b \in \mathbb{R}$, we further deduce from this inequality that

$$\begin{aligned}
&\left(\|x^{k-1} - z^k\| + \|w^k - x^k\|\right)^2 \leqslant 2\left(\|x^{k-1} - z^k\|^2 + \|w^k - x^k\|^2\right) \\
&\leqslant \frac{8C_3}{D_3}\Delta_k \cdot \frac{1}{4C_3}\operatorname{dist}\left(0, \partial H_3(x^k, x^{k-1}, w^k)\right) \\
&\leqslant \frac{8C_3}{D_3}\Delta_k \cdot \frac{1}{4}\left(\|x^k - w^k\| + \|z^k - x^{k-1}\| + \|x^{k-1} - w^{k-1}\| + \|z^{k-1} - x^{k-2}\|\right) \\
&\leqslant \left[\frac{2C_3}{D_3}\Delta_k + \frac{1}{4}\left(\|x^k - w^k\| + \|z^k - x^{k-1}\| + \|x^{k-1} - w^{k-1}\| + \|z^{k-1} - x^{k-2}\|\right)\right]^2,
\end{aligned}$$

where the third inequality follows from Lemma 4.3. Taking square root on both sides of the above inequality and rearrange terms, we obtain

$$\begin{aligned}
&\frac{1}{2}\left(\|x^{k-1} - z^k\| + \|w^k - x^k\|\right) \\
&\leqslant \frac{2C_3}{D_3}\Delta_k + \frac{1}{4}\left[\|x^{k-1} - w^{k-1}\| + \|z^{k-1} - x^{k-2}\| - \|x^k - w^k\| - \|z^k - x^{k-1}\|\right].
\end{aligned}$$

Summing this inequality from $k = N'+1$ to ∞, using (4.32) and the fact that $H_3(x^k, x^{k-1}) > w_3$ for all k, we obtain that

$$\begin{aligned}
\frac{1}{2}\sum_{k=N'+1}^{\infty}\|x^k - x^{k-1}\| &= \frac{1}{2}\sum_{k=N'+1}^{\infty}\|x^k - w^k + \theta_{k-1}(z^k - x^{k-1})\| \\
&\leqslant \frac{1}{2}\sum_{k=N'+1}^{\infty}\left(\|x^{k-1} - z^k\| + \|w^k - x^k\|\right) \\
&\leqslant \frac{2C_3}{D_3}\varphi_3\left(\chi_{N'+1}\right) + \frac{1}{4}\left(\|x^{N'} - w^{N'}\| + \|z^{N'} - x^{N'-1}\|\right) < \infty,
\end{aligned}$$

which implies the convergence of $\{x^k\}$ and $\sum_{k=1}^{\infty} \|x^k - x^{k-1}\| < \infty$. $\square$

4.6 Numerical test

In this section, we perform numerical experiments to study the behaviors of IRL_1e_1, IRL_1e_2 and IRL_1e_3. All codes are written in Matlab, and the experiments are performed in Matlab 2015b on a 64-bit PC with an Intel(R) Core(TM) i7-4790 CPU (3.60GHz) and 32GB of RAM.

We consider the following log penalty regularized least squares problem [46]:

$$\min \ F_{\log}(x) := \frac{1}{2}\|Ax - b\|^2 + \sum_{i=1}^{n} \left(\lambda \log(|x_i| + \epsilon) - \lambda \log \epsilon\right), \tag{4.39}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $\lambda > 0$, $\epsilon > 0$. This is a special case of (4.1) with f being the least squares loss function, $C = \mathbb{R}^n$ and $\phi(t) = \lambda \log(t + \epsilon) - \lambda \log \epsilon$. Thus, we deduce from Theorem 4.3 that the sequence generated by IRL_1e_2 for a choice of $\{\theta_k\}$ satisfying (4.14) clusters at a stationary point of (4.39). In addition, one can check that the corresponding H_1 and H_3 are continuous subanalytic functions [37, Section 6.6], and hence they are KL functions [13, Theorem 3.1]. Consequently, we know from Theorem 4.2 (resp., Theorem 4.5) that if $\sup_k \beta_k < 1$ (resp., $\{\theta_k\}$ satisfies (4.25)), then the whole sequence generated by IRL_1e_1 (resp., IRL_1e_3) converges to a stationary point of (4.39).

In our experiments below, we compare IRL_1e_1, IRL_1e_2 and IRL_1e_3 with two other state-of-the-art algorithms for solving (4.39): the general iterative shrinkage and thresholding algorithm (GIST) [46] and an adaptation of the iteratively reweighted ℓ_1 algorithm [56, Algorithm 7] with nonmonotone line-search (IRL_1ls). We discuss the implementation details of these algorithms below.

IRL$_1e_1$. For this algorithm, we set $L = \lambda_{\max}(AA^T)$ and choose $\{\beta_k\}$ as in FISTA

[10, 68] with both the adaptive and fixed restart schemes [70]:[2]

$$\beta_k = \theta_k(\theta_{k-1}^{-1} - 1) \text{ with } \theta_{k+1} = \frac{2}{1+\sqrt{1+4/\theta_k^2}} \text{ and } \theta_0 = \theta_{-1} = 1$$

and we reset $\theta_{k-1} = \theta_k = 1$ every 200 iterations, or when $\left\langle y^{k-1} - x^k, x^k - x^{k-1} \right\rangle > 0$. It is clear that $\sup_k \beta_k < 1$. We initialize this algorithm at the origin and terminate it when

$$\frac{2L\|x^{k+1} - y^k\| + \ell\|x^{k+1} - x^k\|}{\max\{1, \|x^{k+1}\|\}} < 10^{-4}.$$

Notice from (4.8) that this termination criterion implies that $\operatorname{dist}(0, \partial F(x^{k+1})) \leqslant 10^{-4} \max\left\{1, \|x^{k+1}\|\right\}$.

IRL$_1$e$_2$. For this algorithm, we set $L = \lambda_{\max}(AA^T)$, and let θ_k be as in FISTA [10, 68] for the first 50 iterations, i.e., $\theta_0 = 1$ and $\theta_{k+1} = \frac{2}{1+\sqrt{1+4/\theta_k^2}}$ for $0 \leqslant k \leqslant 48$, $\theta_{50} = \theta_{49}$, and we update $\theta_k = \theta_{99-k}$ for $51 \leqslant k \leqslant 99$ and set $\theta_k = \theta_{\mathsf{mod}(k,100)}$ for $k \geqslant 100$. It can be verified with simple computation that this choice of $\{\theta_k\}$ satisfies (4.14). We initialize the algorithm at the origin and terminate it when

$$\frac{L\|z^{k+1} - y^k\| + \ell\|x^k - z^{k+1}\| + L\|x^{k+1} - y^k\|}{\max\{1, \|z^{k+1}\|\}} < 10^{-4}.$$

Observe from (4.22) that this termination criterion implies that $\operatorname{dist}(0, \partial F(z^{k+1})) \leqslant 10^{-4} \max\left\{1, \|z^{k+1}\|\right\}$.

IRL$_1$e$_3$. For this algorithm, we set $L = \lambda_{\max}(AA^T)$, and we generate a sequence $\{\rho_k\}$ as in FISTA in the first 57 iterations and fix it from then on, i.e., $\rho_0 = 1$ and

$$\rho_{k+1} = \begin{cases} \frac{2}{1+\sqrt{1+4/\rho_k^2}} & 0 \leqslant k \leqslant 55, \\ \rho_{56} & k \geqslant 56. \end{cases}$$

[2] In our experiments, this quantity is computed in matlab with code lambda=norm(A*A'), when $m < 2000$ and by opts.issym = 1; lambda = eigs(A*A',1,'LM',opts); otherwise.

We then set $\theta_k = \rho_{k+6}$ for all $k \geq 0$. It can be verified that the above $\{\theta_k\}$ satisfies (4.25) with $\gamma = 0.95$. We initialize the algorithm at the origin and terminate it when

$$\frac{2L\|x^{k+1} - y^k\| + \ell\|x^{k+1} - x^k\|}{\max\{1, \|x^{k+1}\|\}} < 10^{-4}.$$

Note from (4.35) that this termination criterion implies that $\operatorname{dist}(0, \partial F(x^{k+1})) \leqslant 10^{-4} \max\{1, \|x^{k+1}\|\}$.

GIST. This algorithm was proposed in [46]; see also [98]. Following the notation in [29, Appendix A, Algorithm 1], here we set $c = 10^{-4}, \tau = 2, M = 4$ and set $L_0^0 = 1$ and

$$L_k^0 = \min\left\{10^8, \max\left\{\frac{\|A(x^k - x^{k-1})\|^2}{\|x^k - x^{k-1}\|^2}, 10^{-8}\right\}\right\} \tag{4.40}$$

for $k \geqslant 1$. Note that the subproblem in [29, Appendix A, (A.4)] now becomes

$$\min_{x \in \mathbb{R}^n}\left\{\left\langle A^T(Ax^k - b), x - x^k\right\rangle + \frac{L_k}{2}\|x - x^k\|^2 + \sum_{i=1}^n (\lambda \log(|x_i| + \epsilon) - \lambda \log \epsilon)\right\}$$

whose closed form solution can be found in [46]. We initialize the algorithm at the origin and terminate it when

$$\frac{\|\nabla f(x^k) - \nabla f(x^{k+1})\| + L_k\|x^k - x^{k+1}\|}{\max\{1, \|x^{k+1}\|\}} < 10^{-4};$$

this condition implies that $\operatorname{dist}(0, \partial F(x^{k+1})) \leqslant 10^{-4} \max\{1, \|x^{k+1}\|\}$.

IRL$_1$*ls*. This algorithm is an adaptation of [56, Algorithm 7], which was originally designed for solving (4.1) with $\phi(t) = \lambda \min\{p(ts - \frac{s^q}{q}) : \ 0 \leqslant s \leqslant (\frac{\epsilon}{\lambda n})^{\frac{1}{q}}\}$ for some $p \in (0, 1)$, $q = \frac{p}{p-1}$, $\lambda > 0$ and $\epsilon > 0$. For ease of presentation, we present our adaptation below in Algorithm 4.6. Its convergence can be proved by adapting the convergence analysis of [56, Algorithm 7] and that of [98].

Iteratively reweighted ℓ_1 algorithm with nonmonotone line-search for (4.39) ($\mathbf{IRL_1}ls$)

Step 0. Let $0 < L_{\min} < L_{\max}$, $\tau > 1$ and $c > 0$ be given. Input an initial point x^0 and set $k = 0$.

Step 1. Choose $L_k^0 \in [L_{\min}, L_{\max}]$ and set $L_k = L_k^0$.

Step 2. Set

$$s_i^{k+1} = \frac{\lambda}{|x_i^k| + \epsilon} \text{ for } i = 1, \ldots, n;$$

$$x^{k+1} = \arg\min_{y \in C} \left\{ \langle \nabla f(x^k), y - x^k \rangle + \frac{L_k}{2} \|y - x^k\|^2 + \sum_{i=1}^{n} s_i^{k+1} |y_i| \right\}.$$

Step 3. If

$$F_{\log}(x^{k+1}) > \max_{[k-M]_+ \leqslant s \leqslant k} F_{\log}(x^s) - \frac{c}{2} \|x^{k+1} - x^k\|^2,$$

let $L_k = \tau L_k$, and go to Step 2.

Step 4. If a termination criterion is not met, set $k = k + 1$ and go to Step 1.

For this algorithm, we let $L_{\min} = 10^{-8}, L_{\max} = 10^8, c = 10^{-4}, \tau = 2, M = 4$ and set $L_0^0 = 1$ and for $k \geqslant 1$, we set L_k^0 as in (4.40). We initialize the algorithm at the origin and terminate it when

$$\frac{\|\nabla f(x^k) - \nabla f(x^{k+1})\| + (L_k + \ell)\|x^k - x^{k+1}\|}{\max\{1, \|x^{k+1}\|\}} < 10^{-4}.$$

From Step 2 in Algorithm 4.6, one can observe that this termination criterion implies that $\operatorname{dist}(0, \partial F(x^{k+1})) \leqslant 10^{-4} \max\left\{1, \|x^{k+1}\|\right\}$ at termination.

We compare the above algorithms on random instances. We first generate an $m \times n$ matrix A with i.i.d. standard Gaussian entries and then normalize this matrix to have unit column norms. A subset T of size $r = [\frac{m}{9}]$ is then chosen uniformly at random from $\{1, 2, 3, ..., n\}$ and an r-sparse vector $y \in \mathbb{R}^m$ supported on T with i.i.d. standard Gaussian entries is generated. We then set $b = Ay + 0.01 \cdot \omega$, where $\omega \in \mathbb{R}^m$ has i.i.d. standard Gaussian entries.

In our experiments, we set $(m, n) = (720i, 2560i)$, with $i = 1, \ldots, 10$. We pick $\lambda = 5 \times 10^{-4}$ in (4.39) and experiment with $\epsilon = 0.1$ and 0.5. We present the corresponding results in Tables 4.1 and 4.2, respectively, where we report the time for computing $\lambda_{\max}(A^T A)$ (t_0), the CPU times in seconds (time) and the function values at termination (fval), averaged over 20 random instances. One can see that our algorithms are usually faster than GIST and IRL$_1ls$ and return slightly better function values at termination. Moreover, IRL$_1e_1$ and IRL$_1e_3$ are usually faster than IRL$_1e_2$.

Table 4.1: $\lambda = 5e-4, \epsilon = 0.5$

Problem Size			**time**					**fval**				
m	n	t_0	GIST	IRL$_1ls$	IRL$_1e_1$	IRL$_1e_2$	IRL$_1e_3$	GIST	IRL$_1ls$	IRL$_1e_1$	IRL$_1e_2$	IRL$_1e_3$
720	2560	0.1	1.7	1.5	0.7	0.9	0.6	3.7918e-02	3.7918e-02	3.7897e-02	3.7900e-02	3.7896e-02
1440	5120	0.7	7.0	6.6	3.3	4.3	2.6	7.5904e-02	7.5900e-02	7.5859e-02	7.5867e-02	7.5858e-02
2160	7680	0.6	15.0	14.3	7.2	9.4	5.7	1.1443e-01	1.1443e-01	1.1436e-01	1.1437e-01	1.1436e-01
2880	10240	1.3	25.9	25.0	12.6	16.6	9.9	1.5224e-01	1.5224e-01	1.5215e-01	1.5217e-01	1.5215e-01
3600	12800	2.4	39.4	38.7	19.9	25.6	15.5	1.8805e-01	1.8805e-01	1.8794e-01	1.8796e-01	1.8794e-01
4320	15360	3.8	56.7	55.8	28.1	36.4	21.9	2.2774e-01	2.2773e-01	2.2761e-01	2.2763e-01	2.2761e-01
5040	17920	6.2	75.9	75.0	38.4	48.6	29.7	2.6491e-01	2.6491e-01	2.6474e-01	2.6478e-01	2.6475e-01
5760	20480	8.0	99.8	99.6	50.4	63.6	39.1	3.0627e-01	3.0627e-01	3.0609e-01	3.0614e-01	3.0609e-01
6480	23040	11.1	124.7	123.7	62.8	80.4	48.8	3.4231e-01	3.4233e-01	3.4212e-01	3.4216e-01	3.4212e-01
7200	25600	14.7	157.4	154.9	79.2	101.0	61.4	3.8133e-01	3.8133e-01	3.8111e-01	3.8116e-01	3.8111e-01

Table 4.2: $\lambda = 5e-4, \epsilon = 0.1$

Problem Size			**time**					**fval**				
m	n	t_0	GIST	IRL$_1ls$	IRL$_1e_1$	IRL$_1e_2$	IRL$_1e_3$	GIST	IRL$_1ls$	IRL$_1e_1$	IRL$_1e_2$	IRL$_1e_3$
720	2560	0.1	0.6	0.5	0.3	0.4	0.3	9.3307e-02	9.3307e-02	9.3305e-02	9.3302e-02	9.3303e-02
1440	5120	0.7	2.1	2.0	1.4	1.8	1.5	1.8663e-01	1.8663e-01	1.8663e-01	1.8662e-01	1.8662e-01
2160	7680	0.7	4.5	4.2	3.0	4.0	3.2	2.7940e-01	2.7940e-01	2.7939e-01	2.7938e-01	2.7938e-01
2880	10240	1.3	8.0	7.5	5.2	6.9	5.7	3.7401e-01	3.7401e-01	3.7401e-01	3.7400e-01	3.7400e-01
3600	12800	2.4	12.3	11.6	8.2	10.9	8.9	4.6537e-01	4.6537e-01	4.6536e-01	4.6535e-01	4.6535e-01
4320	15360	3.9	17.4	16.7	11.6	15.3	12.6	5.6347e-01	5.6347e-01	5.6346e-01	5.6344e-01	5.6344e-01
5040	17920	6.2	23.2	21.8	15.6	20.9	17.0	6.5319e-01	6.5320e-01	6.5318e-01	6.5316e-01	6.5317e-01
5760	20480	8.0	30.1	28.4	20.1	26.8	22.0	7.4930e-01	7.4931e-01	7.4930e-01	7.4927e-01	7.4927e-01
6480	23040	10.8	37.9	36.3	25.5	33.9	27.8	8.3961e-01	8.3961e-01	8.3960e-01	8.3957e-01	8.3957e-01
7200	25600	14.5	47.7	45.1	32.0	42.6	34.9	9.3472e-01	9.3472e-01	9.3471e-01	9.3467e-01	9.3468e-01

Chapter 5

Conclusion

In this thesis, we introduced sparsity inducing models that are adopted in many real-world problems such as compressed sensing and statistical problems. Then, existing first-order methods including the proximal gradient method and the iteratively reweighted methods were introduced to solve those sparsity inducing models. Since the proximal gradient method can be slow in many cases as suggested in [5, 10, 64–66, 68], extrapolation techniques were adapted for empirical and possible theoretical acceleration. However, there is not much existing work adapting extrapolation techniques in the iteratively reweighted algorithms.

In this thesis, we incorporated three classical extrapolation techniques presented in [6,10,53,68] into the iteratively reweighted ℓ_1 algorithm. The resulting algorithms are named IRL$_1e_1$, IRL$_1e_2$ and IRL$_1e_3$ respectively and their convergence properties under suitable assumptions on the extrapolation parameters were analyzed. When analyzing IRL$_1e_2$, the global convergence of the sequence generated by IRL$_1e_2$ was not established because of the failure of proving the distance from zero to the subdifferential of the potential function we used there is bounded by the successive changes of the generated sequence. On the other hand, global sequential convergence for IRL$_1e_1$ and IRL$_1e_3$ was established under additional smoothness assumption on the objective function and KL assumption on some suitable potential functions.

In our experiments, we empirically compared IRL_1e_1, IRL_1e_2 and IRL_1e_3 with GIST proposed in [46] and an adaptation of [56, Algorithm 7]. With properly chosen parameters, we found that our algorithms with extrapolation techniques are slightly faster than the other two without these techniques. We also applied our algorithms, GIST proposed in [46] and an adaptation of [56, Algorithm 7] to other models with *log* penalty function in (4.39) being replaced by the MCP function and SCAD function. In those experiments, the iteratively algorithms, IRL_1e_1, IRL_1e_2 and IRL_1e_3 and an adaptation of [56, Algorithm 7], did not perform as GIST in time or the ability of recovery.

Bibliography

[1] F. Alvarez and H. Attouch. An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. *Set-valued Anal.* 9:3–11, 2011.

[2] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions invoving analytic features. *Math. Program.* 116:5–16, 2009.

[3] 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. *Math. Oper. Res.* 35:438–457, 2010.

[4] 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. *Math. Program.* 137:91–129, 2013.

[5] J.-F. Aujol and Ch. Dossal. Stability of over-relaxations for the forward-backward algorithm, application to FISTA. *SIAM J. Optim.* 25:2408–2433, 2015.

[6] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim.* 16:697–725, 2006.

[7] W. Bajwa, J. Haupt, A. Sayeed and R. Nowak. Compressive wireless sensing. Technical report, the 5th International Conference on Information Processing in Sensor Networks, 2006.

[8] D. Baron, M. F. Duarte, M. B. Wakin, S. Sarvotham and R. G. Baraniuk. Distributed compressed sensing. Technical report, Proc. 39th Asilomar Conf. Signals, Systems, and Computers, Pacific Grove, CA, 2005.

[9] H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke and H. Wolkowicz. *Fixed-point algorithms for inverse problems in science and engineering.* chapter 10, 2011.

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

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

[12] E. G. Birgin, J. M. Martinez and M. Raydan. Spectral projected gradient methods: review and perspectives. *J. Stat. Softw.* 60:1–21, 2014.

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

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

[15] J. F. Bonnans and C. Pola. A trust region interior point algorithm for linear constrained optimization. *SIAM J. Optim.* 7:717–731, 1997.

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

[17] J. Borwein and Q. Zhu. *Techniques in Variational Analysis.* Springer, 2005.

[18] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[19] A. B. Brown. A Proof of the Lebesgue Condition for Riemann Integrability Am. Math. Mon. 43:396-398, 1936.

[20] J-F Cai, E. J. Candès and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* 20:1956–1982, 2008.

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

[22] E. J. Candès and P. A. Randall. Highly robust error correction by convex programming. *IEEE T. Inform. Theory.* 54:2829–2840, 2008.

[23] E. J. Candès, J. Romberg and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE T. Inform. Theory.* 52:489–509, 2006.

[24] E. J. Candès and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies. *IEEE T. Inform. Theory.*, 52:5406–5425, 2006.

[25] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE T. Inform. Theory.* 51:4203–4215, 2005.

[26] E. J. Candès, M. B. Wakin and S. P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *J. Fourier Anal. Appl.* 14:877–905, 2008.

[27] R. Chartrand and W. Yin. Iteratively reweighted algorithms for compressive sensing. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2008.

[28] S. Chen, D. L. Donoho and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* 20:33–61, 1999.

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

[30] X. Chen, Q. Lin, S. Kim, J. G. Carbonell and E. P. Xing. Smoothing proximal gradient method for general structure sparse regression. *Ann. Appl. Stat.* 6:719–752, 2012.

[31] X. Chen, Y. Ye, Z. Wang and D. Ge. Complexity of unconstrained $L_2 - L_p$ minimization. *Math. Program.* 143:371–383, 2014.

[32] X. Chen and W. Zhou. Convergence of the reweighted $\ell_1$minimization algorithm for $\ell_2 - \ell_p$ minimization. *Comput. Optim. and Appl.* 59:47–61, 2013.

[33] I. Daubechies, R. DeVore, M. Fornasier and C. S. G*ü*nt*ü*rk. Iteratively reweighted least squares minimization for sparse recovery. *Commun. Pur. Appl. Math.* 63:1-38, 2010.

[34] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage,biometrika. *SIAM J. Sci. Comput.* 81:425–455, 1994.

[35] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* 81:1200–1224, 1995.

[36] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. Preprint. Available at https://arxiv.org/abs/1605.00125, 2016.

[37] F. Facchinei and J-S Pang. *Finite-dimensional variational inequalities and complementarity problems Vol I.* Springer, New York, 2013.

[38] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* 96:1348–1360, 2001.

[39] A. Forsgren, P. E. Gill and M. H. Wright. Interior methods for nonlinear optimizatin. *SIAM Rev.* 44:525–597, 2002.

[40] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing.* Springer, New York, 2013.

[41] J. Friedman, T. Hastie and R. Tibshirani. A note on the group lasso and a sparse group lasso. *Stat. Theory (math.ST).* Submitted on 5 Jan 2010.

[42] D. Ge, X. Jiang and Y. Ye. A note on the complexity of L_p minimization. *Math. Program.* 129:285–299, 2011.

[43] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.* 156:59–99, 2016.

[44] G. Ghen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman funcitons. *SIAM J. Optimiz.* 3:538–543, 1993.

[45] P. Gong and C. Zhang. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. *Proc. Int. Conf. Mach. Learn.* 28:37–45, 2013.

[46] P. Gong, C. Zhang, Z. Lu, J. Z. Huang and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. *Proc. Int. Conf. Mach. Learn.* 28:37–45, 2013.

[47] J. Huang, S. Ma and C. Zhang. The iterated lasso for high-dimensional logistic regression. The University of Iowa, Department of Statistics and Actuarial Science, Technical Report No. 392, November 5, 2008.

[48] L. Jacob, G. Obozinski and J-P Vert. Group lasso with overlap and graph LASSO. Appearing in Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, 2009.

[49] P. R. Johnstone and P. Moulin. Local and global convergence of an inertial version of forward-backward splitting. *SIAM J. Optimiz.* See https://arxiv.org/abs/1502.02281, 2017.

[50] Y. Kim, H. Choi and H-S Oh. Smoothly clipped absolute deviation on high dimensions. *J. Amer. Statist. Assoc. (Theory and Methods).* 103:1665–1673, 2008.

[51] Y. Kim, J. Kim and Y. Kim. Blockwise sparse regression. *Stat. Sinica.* 16:375–390, 2006.

[52] K. Koh, S-J Kim and S. P. Boyd. An interior-point method for large-scale l_1-regularized logistic regression. *J. Mach. Learn. Res.* 8:1519–1555, 2007.

[53] G. Lan, Z. Lu and R. D. C. Monteiro. Primal-dual first-order methods with $O(1/\epsilon)$ iteration-complexity for cone programming. *Math. Program.* 126:1–29, 2011.

[54] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.* 16:964–979, 1979.

[55] Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM. J. Matrix Anal. and Appl.* 31:1235–1256, 2010.

[56] Z. Lu. Iterative reweighted minimization methods for l_p regularized unconstrained nonlinear programming. *Math. Program.* 147:277–307, 2014.

[57] Z. Lu, R. D. C. Monteiro and M. Yuan. Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Math. Program.* 131:163–194, 2012.

[58] Z. Luo and P. Tseng. On the convergence rate of dual ascent methods for linearly constrained convex minimization. *Comput. Optim. and Appl.* 18:846–867, 1993.

[59] M. Lustig, D. L. Donoho and J. M. Pauly. Sparse MRI: the application of compressed sensing for rapid mrimaging. *Magn. Reson. Med.* 58:1182–95, 2007.

[60] S. Ma, D. Goldfarb and L. Chen. Fixed point and bregman iterative methods for matrix rank minimization. *Math. Program.* 128:321–353, 2011.

[61] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, San Diego, CA, 1999.

[62] L. Meier, S. Geer and P. Bühlmann. The group lasso for logistic regression. *J. R. Statist. Soc. B.* 70:53–71, 2008.

[63] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.* 24:227–234, 1995.

[64] Y. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk. SSSR.* 269:543–547, 1983.

[65] Y. Nesterov. *Introductory Lectures on Convex Programming.* Kluwer Academic Publisher, 2004.

[66] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Math. Program.* 120:221–259, 2009.

[67] M. Nikolova. Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares. *Multiscale Model. Sim.* 4:960–991, 2005.

[68] Y. Nesterov. Gradient methods for minimizing composite objective function. *Math. Program.* 140:125–161, 2013.

[69] P. Ochs, A. Dosovitskiy, T. Brox and T. Pock. On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision. *SIAM J. Imaging Sci.* 8:331–372, 2015.

[70] B. O'Donoghue and E. J. Candès. Adaptive restart for accelerated gradient schemes. *Found. Comput. Math.* 15:715–732, 2015.

[71] J. Palmer, D. Wipf and K. Kreutz-Delgado and B. Rao. Variational EM algorithms for non-gaussian latent variable models. *Adv. Neur. In.* 18:1059–1066, 2006.

[72] N. Parikh and S. P. Boyd. Proximal algorithms. *Found. Trends Optimiz.* 1:123–231, 2013.

[73] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Comp. Math. Math.* 4:1–17, 1964.

[74] B. Rao and K. Kreutz-Delgado. An affine scaling methodology for best basis selection. *IEEE T. Signal. Proces.* 47:187–200, 1999.

[75] B. Recht and E. J. Candès. Exact matrix completion via convex optimization. *Found. Comput. Math.* 9:717–772, 1995.

[76] B. Recht, M. Fazel and P. A. Parrilo. Guaranteed minimum rank solutions of matrix equation via nuclear norm minimization. *SIAM Rev.* 52:471–501, 2010.

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

[78] R. T. Rockafellar. Monotone operator and the proximal point algorithm. *SIAM J. Control Optim.* 14:877–898, 1976.

[79] R. T. Rockafellar and R. J-B. Wets. *Variational Analysis*. Springer, 3rd printing 2009.

[80] S. Sardy, A. G. Bruce and P. Tseng. Block coordinate relaxation methods for nonparametric wavelet denoising. *J. Comput. Graph. Stat.* 9:361–379, 1995.

[81] N. Simon, J. Friedman and T.Hastieand R. Tibshirani. A sparse-group lasso. *J. Comput. Graph. Stat.* 22:231–245, 2013.

[82] D. Takhar, V. Bansal, M. Wakin, M. Duarte, D. Baron, J. Laska, K. F. Kelly and R. G. Baraniuk. A compressed sensing camera: new theory and an implementation using digital micromirrors. *Proc. Computational Imaging IV*. 6065:1–10, 2006.

[83] S. Tao, D. Boley and S. Zhang. Local linear convergence of ISTA and FISTA on the lasso problem. *SIAM J. Optimiz.* 26:313–336, 2016.

[84] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.* B. 58:267–288, 1996.

[85] R. Tibshirani. The lasso method for variable selection in the Cox model. *Stat. Med.* 16:385–395, 1997.

[86] K. C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pac. J. Optim.* 6:615-640, 2010.

[87] Y. Tsaig and D. L. Donoho. Extensions of compressed sensing. *Signal Process.* 86:533–548, 2006.

[88] P. Tseng. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM J. Control Optim.* 29:119–138, 1991.

[89] P. Tseng. On accelerated proximal gradient methods for convexconcave optimization. Technical report, 2008.

[90] P. Tseng. Approximation accuracy, gradient methods and error bound for structured convex optimization. *Math. Program.* 125:263–295, 2010.

[91] P. Tseng, I. M. Bomze and W. Schachinger. A first-order interior-point method for linearly constrained smooth optimization. *Math. Program.* 127:399–424, 2011.

[92] P. Tseng and S. Yun. A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. *Comput. Optim. and Appl.* 47:179–206, 2010.

[93] C. Vonesch and M. Unser. A fast thresholded landweber algorithm for wavelet-regularized multidimensional deconvolution. *IEEE T. Image Process.* 17:539–549, 2008.

[94] H. Wang, R. Li and C-L Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika.* 94:553–568, 2007.

[95] B. Wen, X. Chen and T.K. Pong. Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. *SIAM J. Optim.* 17:124–145, 2017.

[96] B. Wen, X. Chen and T. K. Pong. A proximal difference-of-convex algorithm with extrapolation. *Comput. Optim. Appl.* 69:297–324, 2018.

[97] D. Wipf and S. Nagarajan. Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. *IEEE J. Sel. Top. Signa.* 4:317–329, 2010.

[98] S. J. Wright, R. Nowak and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE T. Signal Process.* 57:2479-2493, 2009.

[99] Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM J. Imaging Sci.* 6:1758–1789, 2013.

[100] Y. Ye. *Interior-Point Algorithm: Theory and Analysis*. First edition, 1997.

[101] P. Yu and T. K. Pong. Iteratively reweighted ℓ_1 algorithms with extrapolation. Available at https://arxiv.org/abs/1710.07886.

[102] M. Yuan and L. Yi. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc.* B 68:49–67, 2006.

[103] C. Zalinescu. *Convex Analysis in General Vector Spaces.* World Scientific Publishing Co.Pte.Ltd, 2002.

[104] C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* 38:894–942, 2010.

[105] P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.* 7:2541–2563, 2006.

[106] H. Zou and H. Trevor. Regularization and variable selection via the elastic net. *J. R. Stat. Soc.* B 67:301–320, 2005.