4-4-2023

First-Order Algorithms for Nonlinear Structured Optimization

Miao Zhang
Louisiana State University and Agricultural and Mechanical College

FIRST-ORDER ALGORITHMS FOR NONLINEAR STRUCTURED OPTIMIZATION

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Mathematics

by
Miao Zhang
B.S., Huazhong Agricultural University, 2016
M.S., Huazhong Agricultural University, 2018
M.S., Louisiana State University, 2020
May 2023

This thesis is dedicated to my family.

Acknowledgments

First of all, I could not have undertaken this journey without my advisor Professor Hongchao Zhang. I am extremely grateful for his invaluable patience, guidance and support during my graduate study. His immense knowledge and plentiful experience have encouraged me in all the time of my research and life.

I would like to thank Professor Li-yeng Sung, Professor Xiaoliang Wan and Professor Jerry Trahan for their willingness to serve on my general and dissertation committee and suggestions.

I would like to thank Department of Mathematics for providing a great environment and numerous resources for graduates to enhance skills in teaching and knowledge in mathematics.

I would also like to thank my grandmother, my parents and my aunts for their endless love, and my friends in China and Baton Rouge for their friendship and support.

Table of Contents

Abstract

Nonlinear optimization is a critical branch in applied mathematics and has attracted wide attention due to its popularity in practical applications. In this work, we present two methods which use first-order information to solve two typical classes of nonlinear structured optimization problems.

For a class of unconstrained nonconvex composite optimization problems where the objective is the sum of a smooth but possibly nonconvex function and a convex but possibly nonsmooth function, we propose a unified proximal gradient method with extrapolation, which provides unified treatment to convex and nonconvex problems. The method achieves the best-known convergence rate for first-order methods when solving convex optimization problems. In the case that the problem is nonconvex, the method performs as a proximal gradient method with extrapolation, and a linear convergence rate of the objective values and the generated iterates is obtained under additional proper assumptions. The efficiency of the algorithm is shown by numerical experiments.

For a family of nonconvex separable optimization problems with linear constraints where the objective function is the sum of a smooth but possibly nonconvex function and a possibly nonsmooth nonconvex function, an inexact alternating direction method of multipliers is designed. The method solves subproblems to adaptive error criteria. An expansion step and more flexible dual stepsize are exploited to accelerate the convergence of the algorithm. A linear convergence rate of the generated iterates is guaranteed under proper conditions. Numerical examples illustrate the better performance of the method compared with state-of-the-art ADMM algorithms.

Chapter 1. Introduction

1.1. Nonlinear Structured Optimization Problems

Nonlinear optimization aims to solve the optimization problems where objective function is nonlinear or some of the constraints are not linear. It is a critical branch of applied mathematics and science as plenty of practical applications, especially the ones in the field of machine learning, can be formulated into nonlinear optimization problems. For example, the objective function can be the logistic loss function for logistic regression, the hinge loss for support vector machine and the squared loss for least squares regression [69]. Due to the practical needs of dealing with huge number of samples or data of high dimensions, it is quite common to add regularization terms to the existing loss functions. For example, l_1 penalty is added to the squared loss to obtain sparse solution. The objective function therefore becomes the sum of the squared loss and the l_1 regularization. This problem is also known as l_1 regularized least squared problem (LASSO). Additionally, such problems, namely problems which have special structure, are called structured optimization problems. For instance, LASSO is a typical example of nonlinear structured optimization problems as it has a composite nonlinear objective function. Therefore, many methods have been proposed to accelerate computation by exploiting the structure of these optimization problems.

In this thesis, we focus on two classes of nonlinear structured optimization problems. One is unconstrained composite optimization problems with objective function being the sum of a Lipschitz continuously differentiable function but possibly nonconvex and a proper closed convex but possibly nonsmooth function. The other is a class of linearly constrained separable optimization problems where the objective function is given by the summation of a smooth but possibly nonconvex function and a possibly nonconvex nonsmooth function.

1.2. Literature Review

These two types of problems attracted wide attention in recent years because of the rapid development of data science and machine learning. Tremendous work has been made to solve these nonconvex and nonsmooth optimization problems in the last few decades.

Gradient-based methods have been greatly used to solve general unconstrained optimization problems. Gradient descent method was first suggested by Cauchy in 1847 [19]. It is extremely simple and achieves $\mathcal{O}(1/k)$ complexity bound in terms of functional optimality gap when solving convex smooth optimization problems, where k is the number of total iterations [49, 102]. Nesterov in 1983 proposed an accelerated gradient method to further improve the complexity bound to $\mathcal{O}(1/k^2)$, which exhibits the best-known complexity bound for solving convex smooth problems by only using first-order information [100, 49]. Ghadimi and Lan in [49] generalized Nesterov's accelerated gradient method to nonconvex and possibly stochastic problems by properly specifying a stepsize policy. The modified accelerated gradient method achieves the optimal convergence rate for convex problems while possessing the rate of convergence obtained in [101, 18] for solving general smooth NLP problems. For solving unconstrained composite problems, this modified method can employ the stepsize in the accelerated gradient method even when a component of the objective function is possibly nonconvex.

Similar to the method in [49], which takes advantage of the composite structure of the objective function and is extended to nonconvex problems and further accelerates convergence, many algorithms were proposed to deal with unconstrained composite problems of minimizing the sum of possibly nonconvex or nonsmooth components.

In the case that the objective function is the sum of a smooth but possibly nonconvex function and a nonsmooth function which is a composition of a proper closed function and a linear operator, a proximal augmented Lagrangian method was proposed in [36] which adds an additional optimization variable to replace the linear operator in the composition and then minimizes a differentiable function (i.e., the introduced proximal augmented Lagrangian) over one decision variable rather than a joint nonconvex and nondifferentiable subproblem when applying the method of multipliers. An adapted alternating direction method of multipliers in [86] solves this family of composite problems by adding a proximal term to the second subproblem in the usual ADMM. This modification allows the cluster point of the generated sequence gives a stationary point of the nonconvex problem when choosing a sufficiently large parameter of the augmented Lagrangian. In a special case where the linear operator is the identity, the proximal gradient method can be applied and it is shown any cluster point is a stationary point even with a slightly more flexible constant stepsize [86]. Proximal gradient algorithm with extrapolation in [123] also solves this special class of problems and the paper shows, by choosing the extrapolation parameter under some given threshold, the sequence generated converges R-linearly to a stationary point of the problem and the objective values converge R-linearly as well. There are other variants of proximal gradient method, we refer to [87, 9, 11, 10, 115] and the references therein for details of those algorithms, for example, the fast iterative shrinkage-thresholding algorithm (FISTA) [9]. Ghadimi presented a conditional gradient type method in [48] to solve composite problems when a component is a (weakly) smooth term and the other is a (strongly) convex regularization term, which achieves the best-known complexity results for the first time when the weakly smooth term is nonconvex and nearly optimal complexity when it is convex. The work in [50] provides a generic frame to solve more general nonlinear, possibly nonconvex, optimization problems. A unified accelerated gradient method was proposed to solve composite optimization problem where one component is Lipschitz continuous and the other is possibly nonsmooth but convex. It achieves best-known rate of $\mathcal{O}(1/k)$ in terms of the projected gradient and the optimal rate when the problem is convex. Moreover, a unified prox-level method was also presented in the work to tackle problems where one component has Hölder continuous gradient, which obtains the optimal complexity bound and the best-known iteration complexity for both convex and nonconvex optimization. Additionally, problems, where one of the components in the objective function of composite problems is the finite average of n functions where n can be extremely large, arise widely in machine learning, statistics, and operations research, e.g., [141, 15, 142, 113, 28]. This special class of composite optimization problems are investigated by variants of proximal-based methods (e.g., primal-dual proximal algorithm) and stochastic methods (e.g., mini-batch stochastic approximation method), which can be found in [130, 51, 82, 47].

For the other type of problems we consider in this thesis, i.e., the linearly constrained separable optimization problems, the study of algorithms to solve it can date back to the middle of last century [42, 16, 32]. When the objective function and all constraints are linear in a constrained problem, the optimization problem is called a *linear programming* problem (also referred as *linear optimization*), which is the simplest type of constrained optimization [42, 97]. Since linear optimization is not the purpose of this thesis, we refer to [29, 31, 99, 111, 34, 103, 110, 61, 13, 60] and the references therein for methods of solving linear programming problems, such as the simplex method and the interior point method.

In the case that the objective is separable and nonlinear, and the problem has linear equality constraints, there have been many Lagrangian-based algorithms to solve this special class of constrained problems. Augmented Lagrangian method (ALM) (also known as the method of multipliers) was introduced to obtain convergence under more mild conditions than dual ascent [16]. It first minimizes augmented Lagrangian with respect to primal variables, then updates the dual variable, see, e.g., [73, 43, 22, 96]. The alternating direction method of multipliers (ADMM) was proposed to utilize the decomposition structure of primal variables in the objective function, which results in solving relatively easier subproblems with respect to each primal variable, see [45, 55, 67, 54, 40] and the references therein. However, ALM and ADMM can not solve subproblems in parallel due to the quadratic term in the augmented Lagrangian function. The predictor corrector proximal multiplier method (PCPM) was proposed to address the shortcoming. PCPM first introduces a predictor variable, then minimizes the sum of the Lagrangian, evaluated at the updated predictor rather than the Lagrange multiplier in the original ADMM, and two proximal terms with respect to the two primal variables separately. The coupling term in augmented Lagrangian was ignored in such a way, thus parallel computing is enabled. Finally it updates the Lagrange multiplier as that in ADMM [25, 27]. Another issue with ADMM is that the convergence of the generated primal iterates is not guaranteed in general (although the convergence can be shown under additional assumptions [16]).

The proximal alternating direction method of multipliers (PADMM) introduces proximal terms into the subproblems in ADMM to overcome it [112, 26]. We also refer to [26, 23] for other proximal ADMMs.

Many methods on solving separable linearly constrained optimization problems are directly or closely related to the idea of ADMM as we can observe above. ADMM is indeed a benchmark method for solving linearly constrained separable optimization problems. ADMM was introduced by Glowinski and Marroco in [56] and Gabay and Mercier in [46] in 1970s. The convergence results of the residual, objective values and dual variable can be obtained under modest assumptions for 2-block convex optimization. The generated primal iterates do not necessarily converge to the optimal solution, but the convergence can be shown under additional assumptions (e.g., the strongly convexity or some error bound conditions) [16, 23, 63]. Comparing to the extensive research made for ADMM when solving convex problems, there is limited work for nonconvex cases. ADMM was shown in [75] to converge to the set of stationary points for certain types of nonconvex problems (the consensus and sharing problems) with a sufficiently large penalty parameter. Li and Pong proved the iterates generated by ADMM converge to a stationary point with additional assumptions that both components in the objective function are semi-algebraic, a matrix in the equality constraints is the identity and strongly convexity of one component or full-rank of the other matrix in the constraints hold [85]. The Bregman modification of ADMM (BADMM) presented in [120] converges to a stationary point of the associated augmented Lagrangian with additional assumptions that one matrix in the constraint is injective or a suitable Bregman distance is chosen.

Note that the dominant computation happens to find solutions of the subproblems when applying ADMM, so variants of ADMM were proposed to solve the subproblems inexactly to reduce computation cost while still obtaining convergence results. An inexact ADMM with relative error criteria was proposed which exploits subgradient information at the candidate solutions and two constant parameters to control the accuracy of inexact solutions and error tolerance for convex separable problems [128]. Hager and Zhang introduced an inexact ADMM that uses more adaptive stopping criteria based on both current and accumulated iteration change in the subproblems [65]. Another inexact proximal ADMM in [70] allows the proximal and penalty parameters to change at each iteration. Another drawback of ADMM is that ADMM cannot be directly applied to multiblock problems [24]. Thus much research has also been made to add additional assumptions to the objective function and the constraints or propose variants of ADMM in order to show the convergence of ADMM-type methods for solving multiblock problems, which can be found in [75, 63, 131, 67, 67, 71, 62]. For the separable problems with constraints involving coupling terms, we refer to [74, 30, 75] and the references therein.

In this thesis, motivated by the wide appearances of structured problems in practical applications, we propose two algorithms for nonconvex composite optimization problems and nonconvex nonsmooth separable problems, respectively. We present a unified proximal gradient method with extrapolation for nonconvex composite optimization, which is on the basis of proximal gradient method and extrapolation technique. It is a unified method to handle both convex and nonconvex composite problems. When the problem is convex, it is automatically reduced to an optimal gradient method. If the problem is possibly nonconvex, the algorithm proceeds as a proximal gradient method with extrapolation technique. We also introduce an inexact ADMM for separable nonconvex and nonsmooth linearly constrained optimization problems. The inexact ADMM exploits an expansion linesearch step and adaptive accuracy while global convergence and a linear convergence rate is guaranteed under proper conditions.

1.3. Outline of the Thesis

This thesis is organized as follows. In Chapter 2, we review some fundamental definitions, properties and significant methods closely related to our algorithms or the optimization problems our proposed methods solve. In Chapter 3, we introduce the unified gradient method with extrapolation for nonconvex composite optimization and show the convergence results. Numerical experiments are included as well. In Chapter 4, we present the inexact ADMM for separable nonconvex and nonsmooth linearly constrained optimization problems, the convergence analysis and the numerical examples. In Chapter 5, we conclude the major work in this thesis and discuss some future work that may be of interests.

Chapter 2. Preliminaries

2.1. Some Definitions and Properties

In this section, we review some fundamental definitions and properties of functions. We refer to [84, 101, 42, 12, 6] for more detail.

Throughout the thesis, $\mathbb{R}$, $\mathbb{R}^n$, and $\mathbb{R}^{n\times m}$ denote the sets of real numbers, n dimensional real column vectors, and $n \times m$ real matrices, respectively. Let $\mathbf{I}$ be the identity matrix and $\mathbf{0}$ denote zero matrix or vector. The range of a matrix $\mathcal{Q}$ is denoted as *Range*$(\mathcal{Q})$. We use the notations $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ for the standard Euclidean norm in $\mathbb{R}^n$ and the associated inner product respectively, which are defined as

$$\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} \mathbf{x}_i^2} \text{ and } \langle\mathbf{x},\mathbf{y}\rangle = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{y}_i \tag{2.1}$$

for $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$. In addition, the $\mathcal{Q}$-norm is defined as

$$\|\mathbf{x}\|_{\mathcal{Q}} = \sqrt{\mathbf{x}^{\mathsf{T}}\mathcal{Q}\mathbf{x}} \tag{2.2}$$

for $\mathbf{x} \in \mathbb{R}^n$ and $\mathcal{Q} \in \mathbb{R}^{n\times n}$ is a positive semidefinite matrix. The definition of positive (semi)definite matrix is given in Definition 2.1.14 later.

First, we need following basic definitions to construct our optimization problems.

Definition 2.1.1. *A set $\mathcal{X} \subset \mathbb{R}^n$ is closed if for every convergent sequence $\{\mathbf{x}_t\}$ taken from $\mathcal{X}$, we have $\lim_{t\to\infty}\mathbf{x}_t \in \mathcal{X}$ as well.*

Definition 2.1.2. *For a nonempty subset $\mathcal{X}$ of $\mathbb{R}^n$, $\mathcal{X}$ is convex if and only if $(1-\lambda)\mathbf{x}_1 + \lambda\mathbf{x}_2 \in \mathcal{X}$ for any $\mathbf{x}_1,\mathbf{x}_2 \in \mathcal{X}$ and arbitrary $\lambda \in [0,1]$.*

Definition 2.1.3. *Let $\mathcal{X}$ be s subset of $\mathbb{R}^n$. A real-valued function f is Lipschitz continuous on $\mathcal{X}$ with the constant L if*

$$|f(\mathbf{x}) - f(\mathbf{y})| \le L\|\mathbf{x}-\mathbf{y}\| \tag{2.3}$$

for any $\mathbf{x}$, $\mathbf{y} \in \mathcal{X}$.

The next lemma is an important statement for the geometric interpretation of Lipschitz continuously differentiable functions.

Lemma 2.1.1. *Suppose a real-valued function* f *is Lipschitz continuously differentiable on* $\mathcal{X} \subset \mathbb{R}^n$. *Then for any* $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, *we have*

$$|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{2.4}$$

Remark 2.1.1. *Geometrically we can have following illustration on Lipschitz continuously differentiable functions based on Lemma 2.1.1. Let us define two quadratic functions and fix some* $\mathbf{x}_0 \in \mathcal{X}$*:*

$$f_1(\mathbf{x}) = f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_0\|^2,$$
$$f_2(\mathbf{x}) = f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle - \frac{L}{2}\|\mathbf{x} - \mathbf{x}_0\|^2.$$

Then the graph of f *lies between the graphs of* f_1 *and* f_2, *i.e.,*

$$f_2(\mathbf{x}) \leq f(\mathbf{x}) \leq f_1(\mathbf{x}), \ \forall \mathbf{x} \in \mathcal{X}.$$

Before giving the definition of proper closed functions, we first introduce the effective domain and the epigraph of functions. From here, we consider functions f from $\mathcal{X} \subset \mathbb{R}^n$ to the extended-real-line $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ unless specified.

Definition 2.1.4. *Given a function* f, *the effective domain (or domain) and epigraph of* f *are defined as*

$$dom(f) = \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) < +\infty\},$$

$$epi(f) = \{(\mathbf{x}, r) \in \mathcal{X} \times \mathbb{R} : f(\mathbf{x}) \leq r\},$$

respectively.

Definition 2.1.5. *For any convex set $\mathcal{X}$ in $\mathbb{R}^n$, the interior of $\mathcal{X}$ relative to its affine hull is called the relative interior of $\mathcal{X}$, denoted by int($\mathcal{X}$), where the affine hull of $\mathcal{X}$ is the smallest affine set that includes $\mathcal{X}$, which is the intersection of all the affine sets that include $\mathcal{X}$.*

Remark 2.1.2. *The relative interior coincides with the true interior when the affine hull is all of $\mathbb{R}^n$, but is able to serve as a robust substitute for the true interior when the true interior is empty.*

Next, we have the definition of proper functions and a lemma on the closedness of functions as follows.

Definition 2.1.6. *A function $f : \mathcal{X} \to \overline{\mathbb{R}}$ is called proper if it never takes the value $-\infty$ and there exists at least one $\mathbf{x} \in \mathcal{X}$ such that $f(\mathbf{x}) < +\infty$, i.e., dom(f) $\neq \emptyset$.*

Lemma 2.1.2. *A function f is closed if and only if epi(f) is a closed set.*

Here are a few definitions to construct some useful properties of functions.

Definition 2.1.7. *(Lower semicontinuity.) A function f is called lower semicontinuous at a vector $\mathbf{x} \in \mathcal{X}$ if*

$$f(\mathbf{x}) \leq \lim\inf_{k\to\infty} f(\mathbf{x}_k)$$

for every sequence $\{\mathbf{x}_k\} \subset \mathcal{X}$ with $\{\mathbf{x}_k\} \to \mathbf{x}$ as $k \to \infty$.

Definition 2.1.8. *A function f is lower semicontinuous if it is lower semicontinuous at every point $\mathbf{x}$ in its domain.*

Definition 2.1.9. *For any subset $\mathcal{C} \subseteq \mathcal{X}$, the indicator function of $\mathcal{C}$ is defined as*

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

For an indicator function, its domain is where the function value is 0, i.e., $\text{dom}(\delta_{\mathcal{C}}) = \mathcal{C}$.

Definition 2.1.10. *For a nonempty closed set $\mathcal{X} \subseteq \mathbb{R}^n$, the Euclidean distance from $\mathbf{y}$ to $\mathcal{X}$, denoted as dist$(\mathbf{y}, \mathcal{X})$, is defined by*

$$\textit{dist}(\mathbf{y}, \mathcal{X}) = \inf_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|.$$

Now let us define convex, strongly and weakly convex functions.

Definition 2.1.11. *(Convex functions.) A function $f : \mathcal{X} \to \overline{\mathbb{R}}$ is called convex if epi(f) is a convex set.*

Remark 2.1.3. *Here is another statement for describing convex functions. A proper function f is convex if and only if*

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \tag{2.5}$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ and $\lambda \in [0,1]$. This inequality is a special case of Jensen's inequality which states that for any $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ and $\sum_{i=1}^{n} \lambda_i = 1$ where $\lambda_i \geq 0$ for $i = 1, \ldots, n$, we have

$$f\left(\sum_{i=1}^{n} \lambda_i \mathbf{x}_i\right) \leq \sum_{i=1}^{n} \lambda_i f\left(\mathbf{x}_i\right).$$

A geometric interpretation of convex functions in terms of the inequality (2.5) is that the graph of function f between any two points lies below the line segment between the two points. Below is another equivalent interpretation of convex functions.

Remark 2.1.4. *A differentiable function $f : \mathcal{X} \to \overline{\mathbb{R}}$ is convex if and only if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^{\mathsf{T}}(\mathbf{y} - \mathbf{x}) \tag{2.6}$$

for any $\mathbf{x}$, $\mathbf{y} \in \mathcal{X}$, where $\mathcal{X}$ is an open and convex set.

Definition 2.1.12. *(Strongly convexity.) A function $f : \mathcal{X} \to (-\infty, +\infty]$ is called ν-strongly convex for a given $\nu > 0$ if dom(f) is convex and the following inequality holds for any $\mathbf{x}$, $\mathbf{y} \in$ dom(f) and $\lambda \in [0,1]$:*

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) - \frac{\nu}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2. \tag{2.7}$$

Note that the strong convexity parameter ν depends on the underlying norm. The norm we use here is the Euclidean norm in Definition 2.1. Strongly convexity obviously indicates convexity.

Definition 2.1.13. *[118, weak convexity.] If (2.7) holds for $\nu < 0$, then the function f is called weakly convex.*

Then we define positive semidefinite (and definite) matrix.

Definition 2.1.14. *For a symmetric matrix A, A is positive semidefinite if*

$$\langle A\mathbf{x}, \mathbf{x}\rangle \geq 0, \ \forall \mathbf{x} \in \mathbb{R}^n.$$

We use notation $A \succeq \mathbf{0}$ to indicate a matrix A is positive semidefinite and $A \succ \mathbf{0}$ for A is positive definite (meaning the above inequality must be strict for $\mathbf{x} \neq \mathbf{0}$).

Remark 2.1.5. *By Definition 2.1.14, for matrices A and B, we say $A \succeq B$ if $A - B \succeq \mathbf{0}$ and $A \succ B$ if $A - B \succ \mathbf{0}$.*

Definition 2.1.15. *(Subgradient.) Let $f : \mathcal{X} \to (-\infty, \infty]$ be a proper function and let $\mathbf{x} \in$ dom(f). Any vector $\mathbf{g}$ satisfying*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^T(\mathbf{y} - \mathbf{x}) \ \textit{for all} \ \mathbf{y} \in \mathcal{X}$$

is called a subgradient of the function f at $\mathbf{x}$.

Definition 2.1.16. *(Subdifferential.) The set of all subgradients of f at $\mathbf{x}$ is called the subdifferential of f at $\mathbf{x}$ and is denoted by $\widehat{\partial} f(\mathbf{x})$:*

$$\widehat{\partial} f(\mathbf{x}) = \{\mathbf{g} : f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^T(\mathbf{y} - \mathbf{x}) \ for \ all \ \mathbf{y} \in \mathcal{X}\}. \tag{2.8}$$

Definition 2.1.17. *[109, Definition 8.3 (b)] For a proper lower semicontinuous function f, its limiting subdifferential at $\mathbf{x} \in dom(f)$, denoted as $\partial f(\mathbf{x})$, is defined as*

$$\partial f(\mathbf{x}) = \left\{\boldsymbol{\nu} \in \mathbb{R}^n : \exists \mathbf{x}^k \to \mathbf{x}, f(\mathbf{x}^k) \to f(\mathbf{x}), \boldsymbol{\nu}^k \to \boldsymbol{\nu} \ with \ \boldsymbol{\nu}^k \in \widehat{\partial} f(\mathbf{x}^k)\right\}, \tag{2.9}$$

where $\widehat{\partial} f(\mathbf{x})$ denotes the regular subdifferential of f.

Definition 2.1.18. *(Gradient.) The gradient of a function $f(\mathbf{x})$ on $\mathbb{R}^n$ is defined as*

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \dots, \frac{\partial f(\mathbf{x})}{\partial x_n}\right)^T.$$

Remark 2.1.6. *Subgradient gives affine global underestimator of f. If the function $f(\mathbf{x})$ is differentiable (not necessarily convex) at $\mathbf{x}$, $\partial f(\mathbf{x})$ reduces to a unique vector $\nabla f(\mathbf{x})$. However, if the function is nondifferentiable at the point $\mathbf{x}$, then its subdifferential at $\mathbf{x}$ may contain multiple vectors.*

Remark 2.1.7. *The limiting subdifferential plays a much wider role in nonsmooth and nonconvex analysis and optimization. For example, the Fermat's optimality condition in Theorem 2.1.5 for nonconvex function, that is, if $\mathbf{x}$ is a local minimizer of f, then $\mathbf{0} \in \partial f(\mathbf{x})$ [109].*

Definition 2.1.19. *A point $\mathbf{x}$ is a cluster point of sequence $\{\mathbf{x}_t\}$ if there is a subsequence $\{\mathbf{x}_{t_k}\}$ that converges to $\mathbf{x}$.*

In the following, we state a few important and fundamental facts in optimization.

Theorem 2.1.1. *(First-order optimality condition.) Let $\mathbf{x}^*$ be a local minimizer of a differentiable function $f(\mathbf{x})$. Then*

$$\nabla f(\mathbf{x}^*) = \mathbf{0}.$$

The above theorem is only a necessary condition of a local minimizer for general unconstrained optimization. The points satisfying such condition are called stationary points of function f. Such points are not always the local minimizers. For example, if we look at function $f(x) = x^3$ where $x \in \mathbb{R}$, we have $x = 0$ is a stationary point but not a local solution.

For general constrained optimization, we also have similar first-order necessary conditions, which are known as the Karush-Kuhn-Tucker conditions or KKT conditions for short.

Let us first give a general form of constrained optimization

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & c_i(\mathbf{x}) = 0, i \in \mathcal{E}, \\ & c_i(\mathbf{x}) \geq 0, i \in \mathcal{I}, \end{aligned} \tag{2.10}$$

where f and functions c_i are differentiable on a subset of $\mathbb{R}^n$, and $\mathcal{E}$, $\mathcal{I}$ are sets of finite indices. The Lagrangian function associated with the problem (2.10) is defined as

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \sum_{i\in\mathcal{E}\cup\mathcal{I}} \lambda_i c_i(\mathbf{x}).$$

Definition 2.1.20. *(Active set.) The active set $\mathcal{A}(\mathbf{x})$ at any feasible $\mathbf{x}$ is the union of the set $\mathcal{E}$ with the indices of the active inequality constraints, that is,*

$$\mathcal{A}(\mathbf{x}) = \mathcal{E} \cup \{i \in \mathcal{I} | c_i(\mathbf{x}) = 0\}. \tag{2.11}$$

Definition 2.1.21. *(LICQ.) Given the point $\mathbf{x}^*$ and the active set $\mathcal{A}(\mathbf{x}^*)$ defined by (2.11), we say that the linear independence constraint qualification (LICQ) holds if the set of active constraint gradients $\{\nabla c_i(\mathbf{x}^*), i \in \mathcal{A}(\mathbf{x}^*)\}$ is linearly independent.*

Definition 2.1.22. *A point $\mathbf{x}^*$ is a local solution of the problem (2.10) if $\mathbf{x}^* \in \Omega$ and there is a neighborhood $\mathcal{N}$ of $\mathbf{x}^*$ such that $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for $\mathbf{x} \in \mathcal{N} \cap \Omega$, where Ω is the feasible set of (2.10).*

Now we are ready to state the KKT conditions.

Theorem 2.1.2. *Suppose that $\mathbf{x}^*$ is a local solution of (2.10) and that the LICQ holds at $\mathbf{x}^*$. Then there is a Lagrange multiplier vector $\boldsymbol{\lambda}^*$, with components λ_i^*, $i \in \mathcal{E} \cup \mathcal{I}$, such that the following conditions are satisfied at $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$*

$$\nabla_x \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0}, \tag{2.12a}$$

$$c_i(\mathbf{x}^*) = 0, \quad \forall i \in \mathcal{E}, \tag{2.12b}$$

$$c_i(\mathbf{x}^*) \geq 0, \quad \forall i \in \mathcal{I}, \tag{2.12c}$$

$$\lambda_i^* \geq 0, \quad \forall i \in \mathcal{I}, \tag{2.12d}$$

$$\lambda_i^* c_i(\mathbf{x}^*) = 0, \quad \forall i \in \mathcal{E} \cup \mathcal{I}. \tag{2.12e}$$

In the KKT conditions, (2.12b) and (2.12c) state that $\mathbf{x}^*$ is primal feasible, (2.12d) indicates $\boldsymbol{\lambda}^*$ is dual feasible, and (2.12e) is called complementarity condition, which can be rewritten as

$$\lambda_i^* c_i(\mathbf{x}^*) = 0, \quad \forall i \in \mathcal{I}.$$

Remark 2.1.8. *There might be many vectors $\boldsymbol{\lambda}^*$ satisfying the conditions (2.12) for a given problem of the form (2.10) and solution point $\mathbf{x}^*$. However, when LICQ holds, the optimal $\boldsymbol{\lambda}^*$ is unique.*

Corollary 2.1.1. *Let* $\mathbf{x}^*$ *be a local minimizer of differentiable function* $f(\mathbf{x})$ *subject to linear equality constraints*

$$\{\mathbf{x} \in \mathbb{R}^n | A\mathbf{x} = \mathbf{b}\} \neq \emptyset,$$

where $A \in \mathbb{R}^{m \times n}$ *and* $\mathbf{b} \in \mathbb{R}^m$*,* $m < n$*. Then there exists a vector of multipliers* λ^* *such that*

$$\nabla f(\mathbf{x}^*) = A^T \lambda^*.$$

Theorem 2.1.3. *(Second-order optimality condition.) Let* $\mathbf{x}^*$ *be a local minimizer of twice differentiable function* $f(\mathbf{x})$*. Then*

$$\nabla f(\mathbf{x}^*) = \mathbf{0}, \ H_f(\mathbf{x}^*) \succeq \mathbf{0},$$

where $H_f(\mathbf{x})$ *is the Hessian of function* f *at* $\mathbf{x}$*.*

The above theorem is again only a necessary condition of a local minimizer. Now let us give a sufficient condition.

Theorem 2.1.4. *Let function* f *be twice differentiable function on* $\mathbb{R}^n$ *and* $\mathbf{x}^*$ *satisfy the following conditions:*

$$\nabla f(\mathbf{x}^*) = \mathbf{0}, \ H_f(\mathbf{x}^*) \succ \mathbf{0}.$$

Then $\mathbf{x}^*$ *is a strict local minimizer of* $f(\mathbf{x})$*.*

The above theorems state the optimality conditions for differentiable functions. More generally, we have the following theorem which presents an optimality condition for possibly nondifferentiable functions.

Theorem 2.1.5. *Let* $f : \mathcal{X} \to (-\infty, +\infty]$ *be a proper convex function. Then*

$$\mathbf{x}^* \in \arg\min\{f(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$$

if and only if $\mathbf{0} \in \widehat{\partial} f(\mathbf{x}^*)$.

Next, let us introduce an important definition, proximal mapping (also called proximal operator).

Definition 2.1.23. *Given a function* $g : \mathcal{X} \to (-\infty, +\infty]$*, the proximal mapping of* g *is the operator given by*

$$\mathit{prox}_{t,g}(\mathbf{x}) = \arg\min_{\mathbf{u}} \left\{ \frac{1}{2t} \|\mathbf{u} - \mathbf{x}\|^2 + g(\mathbf{u}) \right\}.$$

Proximal mapping gives a closed-form solution to problems with a wide choices of function g. We list a few examples here. For the sake of simplicity, the value of t is set to 1 and is omitted in the subscript of proximal operator.

- When g is affine, i.e., $g(\mathbf{u}) = \mathbf{a}^T\mathbf{u} + c$, where $\mathbf{a} \in \mathbb{R}^n$ and $c \in \mathbb{R}$, then

$$\text{prox}_g(\mathbf{x}) = \arg\min_{\mathbf{u}} \left\{ \frac{1}{2} \|\mathbf{u} - \mathbf{x}\|^2 + \mathbf{a}^T\mathbf{u} + c \right\} = \mathbf{x} - \mathbf{a}.$$

- If g is convex quadratic given by $g(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T A\mathbf{u} + \mathbf{b}^T\mathbf{u} + c$, where A is positive definite, $\mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$, then

$$\text{prox}_g(\mathbf{x}) = \arg\min_{\mathbf{u}} \left\{ \frac{1}{2} \|\mathbf{u} - \mathbf{x}\|^2 + \frac{1}{2}\mathbf{u}^T A\mathbf{u} + \mathbf{b}^T\mathbf{u} + c \right\} = (A + \mathbf{I})^{-1}(\mathbf{x} - \mathbf{b}).$$

- Let g be the l_1 norm with parameter $\lambda > 0$, we have

$$\text{prox}_g(\mathbf{x}) = \arg\min_{\mathbf{u}} \left\{ \frac{1}{2} \|\mathbf{u} - \mathbf{x}\|^2 + \lambda\|\mathbf{u}\|_1 \right\} = \mathcal{S}_\lambda(\mathbf{x}),$$

  where $\mathcal{S}_\lambda(y)$ is called soft thresholding function defined as

$$\mathcal{S}_\lambda(y) = \begin{cases} y - \lambda, & y \geq \lambda, \\ 0, & |y| < \lambda, \\ y + \lambda, & y \leq -\lambda. \end{cases}$$

The following theorem states the nonexpansivity of proximal operator.

Theorem 2.1.6. *Let f be a proper closed and convex function. Then for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we have*

$$\|prox_f(\mathbf{x}) - prox_f(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|.$$

In the following, we present some frequently used properties and definitions when proving convergence of algorithms or convergence rate.

Definition 2.1.24. *A sequence $\{\mathbf{x}_n\}$ is called Cauchy sequence if for any $\epsilon > 0$, there exists an N such that*

$$\|\mathbf{x}_n - \mathbf{x}_m\| < \epsilon$$

for all m, $n \geq N$.

For Cauchy sequence , we have following lemmas.

Lemma 2.1.3. *Any Cauchy sequence is bounded.*

Lemma 2.1.4. *Every convergent sequence is a Cauchy sequence.*

Lemma 2.1.5. *Every real Cauchy sequence possesses a limit.*

We also have an important inequality stated in the theorem below.

Theorem 2.1.7. *Suppose $\mathbf{x}$ and $\mathbf{y}$ are vectors in $\mathbb{R}^n$. The Cauchy-Schwarz inequality says*

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|.$$

Equality occurs if and only if $\|\mathbf{y}\|$ is a multiple of $\|\mathbf{x}\|$ or vice versa.

Convergence rate is one of the key features to measure the performance of an algorithm. We list a few here and refer to [105] for more details. Suppose we have a sequence $\{\mathbf{x}_t\}$ in $\mathbb{R}^n$ that converges to $\mathbf{x}^*$, then we have the following definitions on describing the convergence rate.

Definition 2.1.25. *The convergence is Q-linear if there is a constant $r \in (0,1)$ such that*

$$\frac{\|\mathbf{x}_{t+1} - \mathbf{x}^*\|}{\|\mathbf{x}_t - \mathbf{x}^*\|} \leq r,$$

for all t sufficiently large.

The above definition states that the distance from current step to the solution $\mathbf{x}^*$ decreases at each iteration by at least a constant factor. The prefix Q stands for "quotient".

Definition 2.1.26. *The convergence is Q-superlinear if*

$$\lim_{t\to\infty} \frac{\|\mathbf{x}_{t+1} - \mathbf{x}^*\|}{\|\mathbf{x}_t - \mathbf{x}^*\|} = 0.$$

for all t sufficiently large.

Definition 2.1.27. *The convergence is Q-quadratic if*

$$\frac{\|\mathbf{x}_{t+1} - \mathbf{x}^*\|}{\|\mathbf{x}_t - \mathbf{x}^*\|^2} \leq M,$$

for all t sufficiently large, where M is a positive constant and not necessarily less than 1.

Here are a few examples to illustrate what kind of sequence has the above convergence rate:

- The sequence $\{1 + 0.5^t\}$ converges Q-linearly to 1.
- The sequence $\{1 + t^{-t}\}$ converges Q-superlinearly to 1.
- The sequence $\{1 + 0.5^{2^t}\}$ converges Q-quadratically to 1.

The convergence rate of some well-known methods has been established. For example, quasi-Newton methods converge Q-superlinearly, Newton's method has Q-quadratically rate, and steepest descent algorithms converge at a Q-linear rate. There is also a weaker form of convergence rate, which we state in the next few definitions.

Definition 2.1.28. *The convergence is R-linear if there is a sequence of nonnegative scalars* $\{\nu_t\}$ *such that*

$$\|\mathbf{x}_t - \mathbf{x}^*\| \leq \nu_t$$

for all t*, and* $\{\nu_t\}$ *converges Q-linearly to zero. The sequence* $\{\|\mathbf{x}_t - \mathbf{x}^*\|\}$ *is said to be dominated by* $\{\nu_t\}$*.*

Definition 2.1.29. *We say the sequence* $\{\mathbf{x}_t\}$ *converges R-superlinearly to* $\mathbf{x}^*$ *if* $\{\|\mathbf{x}_t - \mathbf{x}^*\|\}$ *is dominated by a Q-superlinear sequence.*

Definition 2.1.30. *We say the sequence* $\{\mathbf{x}_t\}$ *converges R-quadratically to* $\mathbf{x}^*$ *if* $\{\|\mathbf{x}_t - \mathbf{x}^*\|\}$ *is dominated by a Q-quadratic sequence.*

Note that the prefix R here stands for "root". R-rates convergence concerns about the overall rate of decrease in the error, instead of the decrease at each iteration. For instance, the sequence

$$\mathbf{x}_t = \begin{cases} 1 + 0.5^t, & t \text{ even}, \\ 1, & t \text{ odd}, \end{cases} \tag{2.13}$$

converges R-linearly to 1. However, the error does not decrease at every step. It actually increases at every second iteration.

2.2. Proximal Gradient Methods and Acceleration

In this section, we introduce some methods closely related to our unified proximal gradient method (UPG) which will be discussed in detail in Chapter 3. UPG is proposed based on the scheme of proximal gradient method with proper modification. UPG reduces to an optimal gradient method if the problem is convex and performs as a proximal gradient method with extrapolation for the nonconvex case. Therefore, we give brief introductions to the proximal gradient method and some variants, and an optimal gradient method in the next few subsections. More details can be found in [101, 49, 9, 6].

2.2.1. Proximal Gradient Method

Considering a general unconstrained optimization problem, gradient descent method solves it by taking steps towards the negative gradient direction by small stepsizes. If the gradient does not exist for the objective function, then the subgradient method can be applied, which has the same idea as the gradient descent method with simple modification in the update steps, i.e., the subgradient method uses subgradient (see Definition 2.1.15) instead of gradient to be the descent direction.

Simply taking gradient or subgradient of the objective function as a search direction does not exploit the structure of the objective function that an optimization problem may carry. For example, the objective function can be a sum of two components. Thus proximal gradient method is proposed to leverage such composite structure of the problem.

Suppose we have following unconstrained composite optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ F(\mathbf{x}) := f(\mathbf{x}) + p(\mathbf{x}), \tag{2.14}$$

where $f(\mathbf{x})$ is convex and differentiable with $\text{dom}(f) = \mathbb{R}^n$ and $p(\mathbf{x})$ is convex but possibly nondifferentiable. Instead of making quadratic approximation of the function $F(\mathbf{x})$, like that in gradient descent method, the proximal gradient method only approximates the differentiable component $f(\mathbf{x})$ of the objective function and keeps the nondifferentiable part $p(\mathbf{x})$.

By doing so, the proximal gradient method actually tries to find the minimizer of the following problem at k-th iteration

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{2t_k}\|\mathbf{x} - \mathbf{x}_k\|^2 + p(\mathbf{x}) \right\}, \tag{2.15}$$

where $t_k \in \mathbb{R}$ is a positive parameter. We can see from (2.15) the quadratic approximation of $f(\mathbf{x})$ is in fact the sum of the linearization of f around the previous output $\mathbf{x}_k$ and a proximal term.

After some algebraic manipulation and cancellation of some constant terms, by the definition of proximal mapping in Definition 2.1.23, the minimization (2.15) can be reformulated into

$$\begin{aligned} \mathbf{x}_{k+1} &= \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2t_k}\|\mathbf{x} - (\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k))\|^2 + p(\mathbf{x}) \right\} \\ &= \operatorname{prox}_{t_k,p}(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)). \end{aligned} \tag{2.16}$$

And the above method is called proximal gradient method.

The proximal gradient method can be viewed as two steps. First, it applies gradient descent method with stepsize of t_k to the differentiable part of the objective function, which is the function $f(\mathbf{x})$ in (2.14). This step corresponds to the input of the proximal mapping in (2.16). The second step is to evaluate the proximal mapping at the output from the first step. This stage only needs the information of the nondifferentiable component, namely $p(\mathbf{x})$, and does not depend on $f(\mathbf{x})$.

The update step (2.16) in the proximal gradient method can also be compactly written as

$$\mathbf{x}_{k+1} = \mathcal{T}_{t_k}(\mathbf{x}_k)$$

where $\mathcal{T}_t(\mathbf{x})$ is called prox-grad operator defined by

$$\mathcal{T}_t(\mathbf{x}) = \operatorname{prox}_{t,p}\left(\mathbf{x} - t\nabla f(\mathbf{x})\right).$$

This compact writing gives an important notion that is frequently used in the convergence analysis of the proximal gradient method and its variants, that is the so-called gradient mapping. The gradient mapping is defined as

$$\mathcal{G}(\mathbf{x}) = \frac{1}{t}\left(\mathbf{x} - \mathcal{T}_t(\mathbf{x})\right). \tag{2.17}$$

The gradient mapping can be demonstrated as a generalization of the idea of gradient. When functions f and p are differentiable, the gradient mapping is just the ordinary gradient. For more general cases, it is shown that the magnitude of the gradient mapping vanishes as the solution to the proximal subproblem approaches to a stationary point of the primal problem (2.14).

Remark 2.2.1. *The proximal gradient method has close connection to gradient descent method and projected gradient method. In the case that p is 0, the proximal gradient method reduces to the gradient descent method. If p is some indicator function, then the proximal gradient method proceeds as projected gradient method.*

The proximal gradient method has an $\mathcal{O}(1/k)$ rate of convergence of the generated function values to the optimal value for the convex case and a linear convergence rate for strongly convex case.

One of the most popular applications of the proximal gradient method is to solve LASSO, which is given by

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|^2 + \lambda\|\mathbf{x}\|_1, \tag{2.18}$$

where $\mathbf{y} \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$ and $\lambda > 0$. It is easy to see the first term and the l_1 regularization in (2.18) correspond to the differentiable component f and nonsmooth part p in (2.14), respectively.

If we directly apply the proximal gradient method, namely making quadratic approximation of the l_2 term around a point $\mathbf{z}$ and doing some algebraic manipulation, we can rewrite (2.18) as

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2}\|\mathbf{y} - A\mathbf{z}\|^2 + \langle -A^\mathsf{T}(\mathbf{y} - A\mathbf{z}), \mathbf{x} - \mathbf{z}\rangle + \frac{1}{2t}\|\mathbf{x} - \mathbf{z}\|^2 + \lambda\|\mathbf{x}\|_1 \right\},$$

where $t > 0$. Then it is equivalent to

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2t\lambda}\|\mathbf{x} - \mathbf{z} - tA^\mathsf{T}(\mathbf{y} - A\mathbf{z})\|^2 + \|\mathbf{x}\|_1 \right\}, \tag{2.19}$$

where $\frac{1}{2}\|\mathbf{y} - A\mathbf{z}\|^2$ is dropped since it does not depend on $\mathbf{x}$. We need to point out that moving λ to the first term does not affect the minimizer, i.e., (2.19) has the same solution as that of the reformulation of (2.18) without dividing λ in the objective function. An alternative way to obtain (2.19) is to replace ∇f by $\nabla f(\mathbf{z}) = -A^\mathsf{T}(\mathbf{y} - A\mathbf{z})$ in (2.16) and divide λ in both terms. Note that here we consider an arbitrary point $\mathbf{z} \in \mathbb{R}^n$ instead of a specific point, e.g., $\mathbf{x}_k$, and ignore the difference of t_k and t.

By Definition 2.1.23 of the proximal mapping and letting $\mathbf{z}$ be the previous iterate $\mathbf{x}_k$, then (2.19) is equivalent to

$$\mathbf{x}_{k+1} = \text{prox}_{t\lambda,\|\cdot\|_1}(\mathbf{x}_k + tA^T(\mathbf{y} - A\mathbf{x}_k)). \tag{2.20}$$

The above method is called the iterative shrinkage-thresholding algorithm (ISTA). Then we can directly see the solution to (2.20) is

$$\mathbf{x}_{k+1} = \mathcal{S}_{t\lambda}(\mathbf{x}_k + tA^T(\mathbf{y} - A\mathbf{x}_k)),$$

where $\mathcal{S}_t(\mathbf{x})$ is the soft thresholding function.

2.2.2. Accelerated Proximal Gradient Method

The proximal gradient method obtains a convergence rate of $\mathcal{O}(1/k)$ in terms of function value gap for general convex nonsmooth optimization. Since the best-known convergence rate obtained by first-order methods is $\mathcal{O}(1/k^2)$ for convex smooth problems, a lot of work has been made to further improve the convergence rate of the proximal gradient method to reach the optimal rate, among which the fast iterative shrinkage-thresholding algorithm (FISTA) by Beck and Teboulle in [9] is a typical accelerated variant of the proximal gradient method.

FISTA can be seen as a generalization of ISTA. Both FISTA and ISTA involve finding solutions of a proximal subproblem. However, ISTA solves the proximal subproblem which is evaluated at the previous point $\mathbf{x}_k$, see the update in (2.20), while the subproblem in FISTA is formulated as

$$\mathbf{x}_k = \mathrm{prox}_{t_k,p}(G(t_k, \mathbf{y}_k)), \tag{2.21}$$

where $G(t_k, \mathbf{y}_k)$ is some gradient step of the smooth part f in (2.14) at a new point $\mathbf{y}_k$ with stepsize t_k, and $\mathbf{y}_k$ is given by

$$\begin{aligned} \theta_{k+1} &= \frac{1+\sqrt{1+4\theta_k^2}}{2}, && (2.22) \\ \mathbf{y}_{k+1} &= \mathbf{x}_k + \left(\frac{\theta_k - 1}{\theta_{k+1}}\right)(\mathbf{x}_k - \mathbf{x}_{k-1}). && (2.23) \end{aligned}$$

We can see from (2.23) that $\mathbf{y}_k$ is generated by a linear combination of the two most recent points and is easy to compute. So the introduction of $\mathbf{y}_k$ does not add expensive computational cost compared to ISTA. In addition, two choices of t_k in (2.21) are presented by the authors. One way sets $t_k = L_f$ throughout the iterations where L_f is the Lipschitz constant of f. The other is to choose it by backtracking. Furthermore, θ_k does not have to be that in (2.22). It can be any sequence satisfying (i) $\theta_k \geq \frac{k+2}{2}$; (ii) $\theta_{k+1}^2 - \theta_{k+1} - \theta_k^2 \leq 0$ for any $k \geq 0$. For example, another possible selection of θ_k is $\frac{k+2}{2}$.

With the above structure (2.21) - (2.23) and that t_k is chosen to be constant or by backtracking and θ_k meets the above criteria, it is proved that FISTA achieves an $\mathcal{O}(1/k^2)$ rate of convergence with respect to function value gap, which improves the convergence rate $\mathcal{O}(1/k)$ of ISTA despite ISTA and FISTA have essentially same steps.

2.2.3. Accelerated Gradient Method

In this subsection, we present an accelerated gradient method (AG) proposed by Ghadimi and Lan in [49], which generalized Nesterov's accelerated gradient method originally solving convex smooth optimization to further address both convex and nonconvex optimization problems by applying some stepsize strategy. Since our unified proximal gradient method in Chapter 3 will reduce to this accelerated gradient method if the optimization problem is convex, we only discuss the AG method for convex composite optimization and skip the details for the nonconvex case.

The AG method considers a class of composite problems given by

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) + p(\mathbf{x}) + \mathcal{X}(\mathbf{x}), \tag{2.24}$$

where f and p are both Lipschitz continuously differentiable and convex, and $\mathcal{X}$ is convex but possibly nonsmooth with bounded domain.

The AG method has three important steps. It first updates a point $\widehat{\mathbf{x}}_{k+1}$ at the points from previous iteration by

$$\widehat{\mathbf{x}}_{k+1} = (1 - \alpha_{k+1})\mathbf{x}_k + \alpha_{k+1}\breve{\mathbf{x}}_k,$$

where α is preset by some stepsize policy. Then the negative gradient evaluated at the most recent point $\widehat{\mathbf{x}}_{k+1}$ is used to be the search direction of the gradient steps in two proximal subproblems, which are as follows:

$$\begin{aligned} \breve{\mathbf{x}}_{k+1} &= \operatorname{prox}_{\lambda_{k+1},\mathcal{X}}(\breve{\mathbf{x}}_k - \lambda_{k+1}\nabla\Phi(\widehat{\mathbf{x}}_{k+1})), \\ \mathbf{x}_{k+1} &= \operatorname{prox}_{\beta_{k+1},\mathcal{X}}(\widehat{\mathbf{x}}_{k+1} - \beta_{k+1}\nabla\Phi(\widehat{\mathbf{x}}_{k+1})), \end{aligned}$$

where $\Phi(\mathbf{x}) := f(\mathbf{x}) + p(\mathbf{x})$ is the smooth part of the objective function in (2.24), λ and β are chosen by certain stepsize strategy.

Stepsize policy is significant in gradient descent method and its variants. The authors established criteria that α, λ and β should satisfy for the algorithm to converge and additionally achieve the optimal convergence rate. An example of selecting α, λ and β is given as

$$\alpha_k = \frac{2}{k+1}, \ \beta_k = \frac{1}{2L_\Phi} \text{ and } \lambda_k = \frac{k\beta_k}{2}, \ \forall k \geq 1, \tag{2.25}$$

where L_Φ is the Lipschitz constant of Φ. For convex cases, the choice of λ in (2.25) is more aggressive than that of nonconvex situations since λ in (2.25) is in the order of $\mathcal{O}(k/L_\Phi)$ while it is in the order of $\mathcal{O}(1/L_\Phi)$ for general nonconvex problems. For more selections of stepsize for both cases, we refer to [49].

The AG method with stepsize selection in (2.25) is proved to converge at the rate of $\mathcal{O}(1/k^2)$ in terms of function value gap, which is the optimal rate for first-order methods. Recall the definition of the gradient mapping in (2.17), then the gradient mapping for this particular method (i.e., the AG method) is given by

$$\mathcal{G}(\widehat{\mathbf{x}}_k) = \frac{1}{\beta_k}\left(\widehat{\mathbf{x}}_k - \mathcal{T}_{\beta_k,\mathcal{X}}(\widehat{\mathbf{x}}_k)\right) = \frac{1}{\beta_k}(\widehat{\mathbf{x}}_k - \mathbf{x}_k).$$

It is shown that the norm square of the gradient mapping converges at the rate of $\mathcal{O}(L_\Phi^2/k^3 + L_\Phi L_f/k)$ for this AG method.

2.3. Alternating Direction Method of Multipliers and Related Methods

In this section, we review two well-known algorithms, the augmented Lagrangian method (ALM, also called the method of multipliers) and alternating direction proximal method of multipliers (AD-PMM), that solve optimization with a separable objective function and linear equality constraints. Alternating direction method of multipliers (ADMM) is a special case of AD-PMM. Here we only give brief introductions of ALM and AD-PMM here. We refer to [16, 6] for more details.

2.3.1. Augmented Lagrangian Method

Consider the following optimization problem

$$\begin{aligned} \min_{(\mathbf{x},\mathbf{y})\in\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}} &\ f(\mathbf{x}) + g(\mathbf{y}) \\ \text{s.t.} &\quad A\mathbf{x} + B\mathbf{y} = \mathbf{b}, \end{aligned} \tag{2.26}$$

where f and g are proper closed convex functions, $A \in \mathbb{R}^{m\times n_1}$, $B \in \mathbb{R}^{m\times n_2}$ and $\mathbf{b} \in \mathbb{R}^m$.

Then the Lagrangian function associated with problem (2.26), denoted as $\mathcal{L}(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda})$, is defined by

$$\mathcal{L}(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) = f(\mathbf{x}) + g(\mathbf{y}) - \boldsymbol{\lambda}^{\mathsf{T}}(A\mathbf{x} + B\mathbf{y} - \mathbf{b}), \tag{2.27}$$

where $\boldsymbol{\lambda}$ is called Lagrange multiplier. In addition, the augmented Lagrangian with penalty parameter $\beta > 0$, denoted as $\mathcal{L}_\beta(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda})$, is given by

$$\mathcal{L}_\beta(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) = \mathcal{L}(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) + \frac{\beta}{2}\|A\mathbf{x} + B\mathbf{y} - \mathbf{b}\|^2. \tag{2.28}$$

The augmented Lagrangian can be viewed as the Lagrangian of the following problem

$$\begin{aligned} \min_{(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}} \; & f(\mathbf{x}) + g(\mathbf{y}) + \frac{\beta}{2}\|A\mathbf{x} + B\mathbf{y} - \mathbf{b}\|^2 \\ \text{s.t.} \quad & A\mathbf{x} + B\mathbf{y} = \mathbf{b}. \end{aligned} \tag{2.29}$$

Then we have the associated dual function of the problem (2.29)

$$h_\beta(\boldsymbol{\lambda}) = \inf_{\mathbf{x},\mathbf{y}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \tag{2.30}$$

and the dual problem at k-th iteration

$$\boldsymbol{\lambda}_{k+1} = \arg\max_{\lambda} h_\beta(\boldsymbol{\lambda}). \tag{2.31}$$

By Theorem 2.1.5, we have $\boldsymbol{\lambda}_{k+1}$ satisfies

$$\mathbf{0} \in \partial h_\beta(\boldsymbol{\lambda}_{k+1})$$

if and only if $\boldsymbol{\lambda}$ is updated as

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \beta(A\mathbf{x}_{k+1} + B\mathbf{y}_{k+1} - \mathbf{b}), \tag{2.32}$$

where

$$\mathbf{x}_{k+1} \in \arg\min \{f(\mathbf{x}) - \langle A^T\boldsymbol{\lambda}_{k+1}, \mathbf{x}\rangle\}, \tag{2.33}$$

$$\mathbf{y}_{k+1} \in \arg\min \{g(\mathbf{y}) - \langle B^T\boldsymbol{\lambda}_{k+1}, \mathbf{y}\rangle\}. \tag{2.34}$$

By applying (2.32) and Theorem 2.1.5, (2.33) and (2.34) are equivalent to

$$(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) \in \arg\min_{\mathbf{x},\mathbf{y}} \left\{ f(\mathbf{x}) + g(\mathbf{y}) + \frac{\beta}{2}\left\| A\mathbf{x} + B\mathbf{y} - \mathbf{b} - \frac{1}{\beta}\boldsymbol{\lambda}_k \right\|^2 \right\}. \tag{2.35}$$

Note that if we expand the quadratic term in (2.35) and drop one of the resulting terms $\frac{1}{2\beta}\|\boldsymbol{\lambda}_k\|^2$ as it does not depend on $\mathbf{x}$ or $\mathbf{y}$, then the objective function will be exactly $\mathcal{L}_\beta(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}_k)$. Now we are ready to give the augmented Lagrangian method, which contains two stages:

$$(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) \in \arg\min_{\mathbf{x},\mathbf{y}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}_k), \tag{2.36}$$

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \beta(A\mathbf{x}_{k+1} + B\mathbf{y}_{k+1} - \mathbf{b}). \tag{2.37}$$

The name of ALM comes from the minimization of the augmented Lagrangian function in the first step.

2.3.2. Alternating Direction Proximal Method of Multipliers

If we look closely at the subproblem (2.36), it is easy to observe that finding the solution of the joint subproblem is difficult due to the coupling terms of $\mathbf{x}$ and $\mathbf{y}$ in the quadratic term of the augmented Lagrangian. A practical way to handle the challenge is to decompose the subproblem into two minimization problems, i.e., we first minimize the objective function with respect to $\mathbf{x}$, then minimize it with respect to $\mathbf{y}$. This modification allows the minimization to proceed over only one variable each time instead of a pair in the original subproblem. Besides, a proximity term is added to each resulting subproblem to further generalize this method. Thus, we have the following alternating direction proximal method of multipliers (AD-PMM):

$$\mathbf{x}_{k+1} \in \arg\min_{\mathbf{x}} \left\{ \mathcal{L}_\beta(\mathbf{x}, \mathbf{y}_k, \boldsymbol{\lambda}_k) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_k\|^2_{\mathcal{D}_x} \right\}, \tag{2.38}$$

$$\mathbf{y}_{k+1} \in \arg\min_{\mathbf{y}} \left\{ \mathcal{L}_\beta(\mathbf{x}_{k+1}, \mathbf{y}, \boldsymbol{\lambda}_k) + \frac{1}{2}\|\mathbf{y} - \mathbf{y}_k\|^2_{\mathcal{D}_y} \right\}, \tag{2.39}$$

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \beta(A\mathbf{x}_{k+1} + B\mathbf{y}_{k+1} - \mathbf{b}), \tag{2.40}$$

where $\mathcal{D}_x \in \mathbb{R}^{n_1 \times n_1}$ and $\mathcal{D}_y \in \mathbb{R}^{n_2 \times n_2}$ are positive semidefinite (see Definition 2.1.14).

Adding the proximity term can lead to various algorithms. For example, in the case that $\mathcal{D}_x$ and $\mathcal{D}_y$ are both zero matrices, the proximity terms in (2.38) and (2.39) are actually dropped. AD-PMM reduces to a method named alternating direction method of multipliers (ADMM). ADMM can also be viewed as a direct generalization of ALM with solving the subproblem in ALM coordinate-wise.

When $\mathcal{D}_x = \alpha_1 \mathbf{I} - \beta A^\mathsf{T} A$ and $\mathcal{D}_y = \alpha_2 \mathbf{I} - \beta B^\mathsf{T} B$ with $\alpha_1 \geq \beta \sigma_{\max}(A^\mathsf{T} A)$ and $\alpha_2 \geq \beta \sigma_{\max}(B^\mathsf{T} B)$ where $\sigma_{\max}(W)$ denotes the largest eigenvalue of a matrix W, then it still holds $\mathcal{D}_x$ and $\mathcal{D}_y$ are positive semidefinite. Recall that (2.35) is equivalent to (2.36). The objective function in (2.38) can be reformulated as

$$\begin{aligned}
& f(\mathbf{x}) - \boldsymbol{\lambda}_k^\mathsf{T}(A\mathbf{x} + B\mathbf{y}_k - \mathbf{b}) + \frac{\beta}{2}\left\|A\mathbf{x} + B\mathbf{y}_k - \mathbf{b}\right\|^2 + \frac{1}{2}\|\mathbf{x} - \mathbf{x_k}\|^2_{\mathcal{D}_x} \\
= \; & f(\mathbf{x}) + \frac{\beta}{2}\left\|A\mathbf{x} + B\mathbf{y}_k - \mathbf{b} - \frac{1}{\beta}\boldsymbol{\lambda}_k\right\|^2 + \frac{1}{2}\|\mathbf{x} - \mathbf{x_k}\|^2_{\mathcal{D}_x} \\
= \; & f(\mathbf{x}) + \frac{\beta}{2}\left\|A(\mathbf{x} - \mathbf{x}_k) + A\mathbf{x}_k + B\mathbf{y}_k - \mathbf{b} - \frac{1}{\beta}\boldsymbol{\lambda}_k\right\|^2 + \frac{1}{2}\|\mathbf{x} - \mathbf{x_k}\|^2_{\mathcal{D}_x} \\
= \; & f(\mathbf{x}) + \frac{\beta}{2}\left\|A(\mathbf{x} - \mathbf{x}_k)\right\|^2 + \beta\langle A(\mathbf{x} - \mathbf{x}_k), A\mathbf{x}_k + B\mathbf{y}_k - \mathbf{b} - \frac{1}{\beta}\boldsymbol{\lambda}_k\rangle + \frac{1}{2}\|\mathbf{x} - \mathbf{x_k}\|^2_{\mathcal{D}_x} \\
= \; & f(\mathbf{x}) + \beta\langle A\mathbf{x}, A\mathbf{x}_k + B\mathbf{y}_k - \mathbf{b} - \frac{1}{\beta}\boldsymbol{\lambda}_k\rangle + \frac{\alpha_1}{2}\|\mathbf{x} - \mathbf{x}_k\|^2, \qquad (2.41)
\end{aligned}$$

where the equalities hold up to a constant which does not depend on $\mathbf{x}$. Therefore (2.38) can be written as

$$\mathbf{x}_{k+1} \in \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}) + \beta\langle A\mathbf{x}, A\mathbf{x}_k + B\mathbf{y}_k - \mathbf{b} - \frac{1}{\beta}\boldsymbol{\lambda}_k\rangle + \frac{\alpha_1}{2}\|\mathbf{x} - \mathbf{x}_k\|^2 \right\}. \qquad (2.42)$$

Similarly, we have the $\mathbf{y}$-subproblem (2.39) as

$$\mathbf{y}_{k+1} \in \arg\min_{\mathbf{y}} \left\{ g(\mathbf{y}) + \beta\langle B\mathbf{y}, A\mathbf{x}_{k+1} + B\mathbf{y}_k - \mathbf{b} - \frac{1}{\beta}\boldsymbol{\lambda}_k\rangle + \frac{\alpha_2}{2}\|\mathbf{y} - \mathbf{y}_k\|^2 \right\}. \qquad (2.43)$$

Note that the second term in (2.42), namely the inner product, in fact comes from the linearization of $\frac{\beta}{2}\left\|A\mathbf{x}+B\mathbf{y}_k-\mathbf{b}-\frac{1}{\beta}\boldsymbol{\lambda}_k\right\|^2$ around the point $\mathbf{x}_k$. Same observation holds for that in (2.43). Thus, the steps (2.42) and (2.43), together with the dual variable update (2.40) are called alternating direction linearized proximal method of multipliers (AD-LPMM).

Since both ADMM and AD-LPMM are variants of AD-PMM by choosing different matrices in the proximal terms, we only state the convergence results of AD-PMM. Under mild assumptions, AD-PMM is shown to converge at a rate of $\mathcal{O}(1/k)$ in terms of function value and residual of the equality constraints.

Chapter 3. Unified Proximal Gradient Method for Composite Problems

In this chapter, we propose a unified algorithm that is based on the proximal gradient method and is equipped with extrapolation and linesearch techniques. The algorithm deals with composite optimization problems that are possibly nonsmooth and nonconvex. It gives a unified way to solve convex and nonconvex problems. If the composite problems are convex, the algorithm reduces to an accelerated gradient method similar to Algorithm 2 in [49] for convex cases and obtains the best-known convergence rate of first-order methods. When the problems are nonconvex, a linesearch technique is activated which improves the efficiency of the method, and the algorithm performs as a proximal gradient method with extrapolation.

3.1. Composite Problems

Let us consider the composite optimization problem

$$\min_{\mathbf{x}\in\mathcal{X}} \; F(\mathbf{x}) := f(\mathbf{x}) + p(\mathbf{x}), \tag{3.1}$$

where $\mathcal{X} \subset \mathbb{R}^n$ is a closed convex set, f is Lipschitz continuously differentiable on an open set containing $\mathcal{X}$, but possibly nonconvex and $p : \mathcal{X} \to \mathbb{R}$ is a proper closed convex, but possibly nonsmooth, function. The constraint $\mathbf{x} \in \mathcal{X}$ can also be formulated as an indicator function of $\mathcal{X}$ into the function p. Applications in the form of (3.1) appear frequently in machine learning, statistical inference, and image processing (e.g., [83, 78, 69, 58]).

Motivated by the wide practical applications of problems in the form of (3.1), the advantage of extrapolation and the need of more efficient methods to handle both convex and nonconvex problems, a unified proximal gradient method with extrapolation is proposed in Algorithm 3.1. The contribution of the work mainly lies in the following aspects.

First, the proposed method gives unified treatment to the problem (3.1) for which the problem may be convex or not. Much work has been done to solve the problem (3.1). When both f and p are convex, the accelerated gradient methods are proved to achieve an optimal convergence rate of $\mathcal{O}(1/k^2)$. However, the convergence for the nonconvex case is not fully clear [8, 9]. In the case that f is nonconvex but g is convex, the convergence results are shown for methods like general iterative shrinkage and thresholding in [59], gradient descent with proximal average in [139]. However, the convergence rate of these methods for the convex situations are not analyzed. The analysis of a generalized accelerated gradient method in [49] provides convergence in terms of gradient mapping for the case only when f is possibly nonconvex. If both f and p are nonconvex, APG-like algorithms are proposed in [87] to solve the problems. The algorithms have accelerated convergence rate for convex problems and a linear rate of function values is achieved when the problems are nonconvex. Therefore, the study of unified methods to tackle both convex and nonconvex problems is still limited. The proposed method in Algorithm 3.1 admits unified analysis for both situations. The algorithm obtains the optimal convergence rate for convex optimization problems. When the problems are nonconvex, a linear convergence rate of the generated iterates and function values is presented under additional proper assumptions.

Second, the proposed algorithm is equipped with an extrapolation step where the extrapolation parameter is determined by a linesearch technique. Extrapolation can date back to the extragradient method of Korpelevich in [81] and the heavy ball method by Polyak in [106]. It is now greatly used in optimization methods to accelerate convergence, e.g., [20, 126, 88, 124], while not significantly increasing the computational cost. The extrapolation involves a linear combination of points from previous iteration. Then the gradients in the subproblems in Algorithm 3.1 are evaluated at the extrapolated point. The extrapolation parameter is adaptively chosen by a linesearch technique, so that the extrapolation parameter is not of a fixed form, for example, $\frac{1}{t+1}$.

Third, a stochastic accelerated gradient method with variance reduction is presented in Algorithm 3.6, which is a generalization of the unified proximal gradient method in Algorithm 3.1 to solve stochastic optimization problems that can be possibly nonconvex. Stochastic methods are exceptionally useful when dealing with datasets which have huge number of samples, in other words, n in problem (3.70) is extremely large. A variance reduction step is implemented such that the expectation of the stochastic gradients used in the subproblems is within some range of the gradient in the deterministic case. Numerical experiments illustrate the efficiency of the proposed stochastic method, though the theoretical results are still a work in progress.

3.2. Algorithm Description

We would like to apply Algorithm 3.1 to solve problem (3.1). Algorithm 3.1 is an extension of the accelerated gradient method for solving convex composite optimization to the case that f is not necessarily convex.

Note that, if the problem is convex, the inequality in Step 2 of Algorithm 3.1 is automatically satisfied with $\mu_t = 0$ for any $t \geq 0$ (see Remark 2.1.4). In this way, the linesearch technique is inactivated. Then $\overline{\tau}_t$, $\underline{\tau}_t$, τ_t are all reduced to 0. Therefore, we have $\beta_t = \overline{\beta}_t$ and Step 2 is just an extrapolation step with parameter $\beta_t = \frac{2}{t+1}$. In addition, the parameters in Step 3 are also set to be $\gamma_t = \frac{2L}{t}$ and $\eta_t = L + \frac{1}{t}$. It is easy to verify that the

```
Initialization:  Given $\mathbf{x}_1 \in \mathcal{X}$, $\rho > 1$, $\lambda \in [0,1]$ and $L > \mathcal{L}$;
                 Set $\breve{\mathbf{x}}_1 = \mathbf{x}_1$ and $\mu_0 = 0$.
For $t = 1, 2, 3, \ldots$
1. Set $\overline{\beta}_t = 2/(t+1)$.
2. Choose the smallest integer $j \geq 0$ such that $\mu_t = \min\{\mu_{t-1} + \rho^j - 1, L\}$
   and $\widehat{\mathbf{x}}_t = \beta_t \breve{\mathbf{x}}_t + (1-\beta_t)\mathbf{x}_t$ satisfy
      $f(\mathbf{x}_t) - f(\widehat{\mathbf{x}}_t) - \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_t - \widehat{\mathbf{x}}_t \rangle \geq -\frac{\mu_t}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2$, where $\beta_t = \max\{\overline{\beta}_t, \tau_t\}$
   with $\tau_t = \lambda \underline{\tau}_t + (1-\lambda)\overline{\tau}_t$, $\underline{\tau}_t = \frac{1}{2}\left(1 - \sqrt{\frac{L-\mu_t}{L+\mu_t}}\right)$ and $\overline{\tau}_t = \frac{\mu_t}{L+\mu_t}$.
3. Set $\gamma_t = \beta_t \eta_t$, where $\eta_t = 2L/(2-\beta_t)$.
4. $\breve{\mathbf{x}}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x} \rangle + \frac{\gamma_t}{2}\left\|\mathbf{x} - \breve{\mathbf{x}}_t\right\|^2 + p(\mathbf{x}) \right\}$.
5. If $\beta_t = 1$, let $\mathbf{x}_{t+1} = \breve{\mathbf{x}}_{t+1}$;
6. Else      $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x} \rangle + \frac{\eta_t}{2}\left\|\mathbf{x} - \widehat{\mathbf{x}}_t\right\|^2 + p(\mathbf{x}) \right\}$.
end
```

Algorithm 3.1. A unified algorithm for nonconvex composite optimization with extrapolation

subproblem in Step 4 is equivalent to

$$\breve{\mathbf{x}}_{t+1} = \operatorname{prox}_{\frac{1}{\gamma_t}, p+\delta_{\mathcal{X}}}\left(\breve{\mathbf{x}}_t - \frac{1}{\gamma_t}\nabla f(\widehat{\mathbf{x}}_t)\right), \tag{3.2}$$

where $\delta_{\mathcal{X}}$ is an indicator function of $\mathcal{X}$. Observe that the stepsize sequence $\left\{\frac{1}{\gamma_t}\right\}$ in the subproblem (3.2) is in the order of $\mathcal{O}(\frac{t}{L})$, which is an aggressive stepsize policy since the stepsize gets larger as iteration goes on. Furthermore, Lipschitz constants of nonlinear functions are usually difficult to get or it takes much computational effort to obtain. So, no knowledge of Lipschitz constant is required throughout the proposed algorithm for both convex and nonconvex cases. Instead, we only need a constant L such that $L > \mathcal{L}$ where $\mathcal{L}$ is the Lipschitz constant of the function f.

3.3. Global Convergence Analysis

In this section, we discuss the global convergence of Algorithm 3.1. We have the following assumptions throughout the chapter.

Assumption 3.3.1. *The gradient of f is Lipschitz continuous, i.e., there exists a constant*

$\mathcal{L} > 0$ *such that*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \mathcal{L}\|\mathbf{x} - \mathbf{y}\| \tag{3.3}$$

for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$*.*

Assumption 3.3.2. *Assume* p *has a strongly convex modulus* $\nu \geq 0$*, i.e., for any* $\mathbf{x} \in \mathcal{X}$*,* $\mathbf{y} \in \mathcal{X}$ *and* $\mathbf{p} \in \widehat{\partial} p(\mathbf{x})$*, it has*

$$p(\mathbf{y}) - p(\mathbf{x}) - \langle \mathbf{p}, \mathbf{y} - \mathbf{x} \rangle \geq \frac{\nu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \tag{3.4}$$

where $\widehat{\partial} p(\mathbf{x})$ *is the subdifferential of the proper closed convex function* p *at* $\mathbf{x}$*.*

Assumption 3.3.3. *Assume the function value of* F *on* $\mathcal{X}$ *is bounded below, i.e., we have* $\overline{F} > -\infty$*, where* $\overline{F} := \inf_{x \in \mathcal{X}} F(\mathbf{x})$*.*

Note that based on Assumption 3.3.1, there exists a constant $\mu \in [0, \mathcal{L}]$ such that

$$-\frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{\mathcal{L}}{2}\|\mathbf{x} - \mathbf{y}\|^2, \tag{3.5}$$

for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

When f is a convex function, Algorithm 3.1 will be just reduced to an accelerated gradient method for solving convex composite optimization. In this case the convergence properties of Algorithm 3.1 is standard and had been established in the literature [49]. Hence, we just state the following convergence theorem and only provide a sketch of its proof.

Theorem 3.3.1. *Suppose the Assumptions 3.3.1 and 3.3.2 hold, and* f *is a convex function. If the solution set of problem (3.1) is not empty, for the iterates generated by Algorithm 3.1, we have*

$$F(\mathbf{x}_{k+1}) - F(\mathbf{x}^*) \leq \frac{2L}{k(k+1)}\|\mathbf{x}^* - \mathbf{x}_1\|^2 \tag{3.6}$$

and

$$\min_{t\in\{1,\ldots,k\}} \|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 \leq \frac{24L^3}{(L-\mathcal{L})k^2(k+1)}\|\mathbf{x}^* - \mathbf{x}_1\|^2, \tag{3.7}$$

where $\mathbf{g}(\widehat{\mathbf{x}}_t) = \eta_t(\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1})$ *and* $\mathbf{x}^*$ *is any optimal solution of (3.1).*

Proof. Since $\mu_0 = 0$ and f is a convex function, we can see from Algorithm 3.1 that $\mu_t = 0$ for all $t \geq 0$, which implies $\underline{\tau}_t = \frac{1}{2}\left(1 - \sqrt{(L-\mu_t)/(L+\mu_t)}\right) = 0$ and $\overline{\tau}_t = \mu_t/(L+\mu_t) = 0$ for all $t \geq 1$. Hence, we have $\tau_t = 0$ and $\beta_t = \overline{\beta}_t$ for all $t \geq 1$. So, in this case, Algorithm 3.1 is just reduced to an accelerated gradient method for solving convex composite optimization. Then, following the convergence results in the literature, the iterates generated by Algorithm 3.1 have the following property: for any $\mathbf{x} \in \mathcal{X}$, we have

$$\begin{aligned} F(\mathbf{x}_{t+1}) - F(\mathbf{x}) \;\; &\leq \;\; (1-\beta_t)(F(\mathbf{x}_t) - F(\mathbf{x})) + \frac{\beta_t\gamma_t}{2}\left[\|\mathbf{x} - \breve{\mathbf{x}}_t\|^2 - \|\mathbf{x} - \breve{\mathbf{x}}_{t+1}\|^2\right] \\ &\quad - \frac{\eta_t - \mathcal{L}}{2\eta_t^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\eta_t}{2}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2, \end{aligned} \tag{3.8}$$

where $\widetilde{\mathbf{x}}_{t+1} = \beta_t\breve{\mathbf{x}}_{t+1} + (1-\beta_t)\mathbf{x}_t$, $\eta_t - \mathcal{L} = 2L/(2-\overline{\beta}_t) - \mathcal{L} = L(t+1)/t - \mathcal{L} > L - \mathcal{L} > 0$ and $\mathbf{g}(\widehat{\mathbf{x}}_t) = \eta_t(\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1})$.

Dividing $\Gamma_t = 2/(t(t+1))$, $t \geq 1$, on both sides of (3.8), for $t \geq 2$, we obtain

$$\begin{aligned} &\frac{1}{\Gamma_t}\left(F(\mathbf{x}_{t+1}) - F(\mathbf{x})\right) + \frac{\eta_t - \mathcal{L}}{2\eta_t^2\Gamma_t}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 + \frac{\eta_t}{2\Gamma_t}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 \\ \leq \;\; &\frac{1}{\Gamma_{t-1}}(F(\mathbf{x}_t) - F(\mathbf{x})) + \frac{\beta_t\gamma_t}{2\Gamma_t}\left[\|\mathbf{x} - \breve{\mathbf{x}}_t\|^2 - \|\mathbf{x} - \breve{\mathbf{x}}_{t+1}\|^2\right], \end{aligned}$$

which by $\eta_t = L(t+1)/t$, $\beta_t = 2/(t+1)$ and $\gamma_t = 2L/t$ can be simplified to

$$\begin{aligned} &\frac{1}{\Gamma_t}\left(F(\mathbf{x}_{t+1}) - F(\mathbf{x})\right) + \frac{L - \mathcal{L}}{4L^2(t+1)/t^3}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 + \frac{L(t+1)^2}{4}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 \\ \leq \;\; &\frac{1}{\Gamma_{t-1}}(F(\mathbf{x}_t) - F(\mathbf{x})) + \|\mathbf{x} - \breve{\mathbf{x}}_t\|^2 - \|\mathbf{x} - \breve{\mathbf{x}}_{t+1}\|^2. \end{aligned} \tag{3.9}$$

When $t = 1$, by (3.8) and $\beta_1 = 1$, we have

$$\begin{aligned}
& \frac{1}{\Gamma_1}\left(F(\mathbf{x}_2) - F(\mathbf{x})\right) + \frac{L-\mathcal{L}}{8L^2}\|\mathbf{g}(\widehat{\mathbf{x}}_1)\|^2 + L\|\mathbf{x}_2 - \widetilde{\mathbf{x}}_{t+1}\|^2 \\
\leq \;& \|\mathbf{x} - \breve{\mathbf{x}}_1\|^2 - \|\mathbf{x} - \breve{\mathbf{x}}_2\|^2. \qquad (3.10)
\end{aligned}$$

Adding (3.9) and (3.10) for $t = 1, \ldots, k$, we have

$$\begin{aligned}
& \sum_{t=1}^{k}\left(\frac{L-\mathcal{L}}{4L^2(t+1)/t^3}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 + \frac{L(t+1)^2}{4}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2\right) \\
& + \frac{1}{\Gamma_k}\left(F(\mathbf{x}_{k+1}) - F(\mathbf{x})\right) \\
\leq \;& \|\mathbf{x} - \breve{\mathbf{x}}_1\|^2 = \|\mathbf{x} - \mathbf{x}_1\|^2, \qquad (3.11)
\end{aligned}$$

for any $\mathbf{x} \in \mathcal{X}$. Then, taking $\mathbf{x} = \mathbf{x}^*$ in (3.11), we can derive (3.6) and (3.7) by direct calculations. □

We add a few observations about the results in Theorem 3.3.1. For the convex case, UPG achieves the convergence rate $\mathcal{O}(1/k^2)$ in (3.6) in terms of function values, which is the best-known convergence rate for methods only using first-order information. Additionally, the gradient mapping converges at the rate of $\mathcal{O}(1/k^3)$ by (3.7). In other words, UPG can find a solution $\overline{\mathbf{x}}$ such that $\|\mathbf{g}(\overline{\mathbf{x}})\|^2 \leq \epsilon$ in $\mathcal{O}(1/\epsilon^{\frac{1}{3}})$ iterations.

In the following we focus on studying the convergence of Algorithm 3.1 when f is not necessarily a convex function. We first have the following lemma.

Lemma 3.3.1. *Suppose the Assumptions 3.3.1 and 3.3.2 hold. Then, for the iterates generated by Algorithm 3.1, we have*

$$\begin{aligned}
F(\mathbf{x}_{t+1}) \;\leq\;& F(\mathbf{x}_t) + \frac{\mu_t + \gamma_t/\beta_t}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 - \frac{\gamma_t/\beta_t}{2}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 \\
& - \frac{\beta_t\nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 - \frac{\eta_t - \mathcal{L}}{2\eta_t^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\eta_t + \nu}{2}\left\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\right\|^2, \qquad (3.12)
\end{aligned}$$

where

$$\widetilde{\mathbf{x}}_{t+1} = \beta_t \breve{\mathbf{x}}_{t+1} + (1 - \beta_t)\mathbf{x}_t. \tag{3.13}$$

Proof. We first observe that all the iterates $\mathbf{x}_t$, $\breve{\mathbf{x}}_t$ and $\widehat{\mathbf{x}}_t$ are contained in $\mathcal{X}$ and $\beta_t \in (0, 1]$ for all $t \geq 1$. Then, by the definition of $\widetilde{\mathbf{x}}_{t+1}$ in (3.13), we also have $\widetilde{\mathbf{x}}_{t+1} \in \mathcal{X}$, since $\mathcal{X}$ is a convex set. By (3.5), the following relations hold

$$\begin{aligned}
f(\mathbf{x}_{t+1}) &\leq f(\widehat{\mathbf{x}}_t) + \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t \rangle + \frac{\mathcal{L}}{2} \left\| \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t \right\|^2 \\
&= f(\widehat{\mathbf{x}}_t) + \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_t - \widehat{\mathbf{x}}_t \rangle + \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{\mathcal{L}}{2} \left\| \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t \right\|^2 \\
&\leq f(\mathbf{x}_t) + \frac{\mu_t}{2} \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \langle \nabla f(\widehat{\mathbf{x}}_t), \widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t \rangle + \frac{\mathcal{L}}{2} \left\| \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t \right\|^2 \\
&\quad + \left\langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1} \right\rangle.
\end{aligned} \tag{3.14}$$

Note that $\eta_t = 2L/(2 - \beta_t) > L > \mathcal{L}$. Since

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x} \rangle + \frac{\eta_t}{2} \left\| \mathbf{x} - \widehat{\mathbf{x}}_t \right\|^2 + p(\mathbf{x}) \right\} \tag{3.15}$$

and $\widetilde{\mathbf{x}}_{t+1} \in \mathcal{X}$, we obtain

$$\begin{aligned}
&\langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1} \rangle + p(\mathbf{x}_{t+1}) \\
\leq\;& \frac{\eta_t}{2} \left(\|\widetilde{\mathbf{x}}_{t+1} - \widehat{\mathbf{x}}_t\|^2 - \|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|^2 \right) + p(\widetilde{\mathbf{x}}_{t+1}) - \frac{\eta_t + \nu}{2} \left\| \mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1} \right\|^2.
\end{aligned} \tag{3.16}$$

By the definition (3.13) of $\widetilde{\mathbf{x}}_{t+1}$ and $\widehat{\mathbf{x}}_t = \beta_t \breve{\mathbf{x}}_t + (1 - \beta_t)\mathbf{x}_t$, we have

$$\beta_t(\breve{\mathbf{x}}_{t+1} - \widehat{\mathbf{x}}_t) + (1 - \beta_t)(\mathbf{x}_t - \widehat{\mathbf{x}}_t) = \widetilde{\mathbf{x}}_{t+1} - \widehat{\mathbf{x}}_t = \beta_t \mathbf{s}_t, \tag{3.17}$$

where $\mathbf{s}_t = \breve{\mathbf{x}}_{t+1} - \breve{\mathbf{x}}_t$.

Let us define

$$\mathbf{g}(\widehat{\mathbf{x}}_t) = \eta_t(\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1}). \tag{3.18}$$

Then, it follows from (3.16), (3.17) and (3.18) that

$$
\begin{aligned}
& \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1} \rangle \\
\leq \;& \frac{\eta_t \beta_t^2}{2} \|\mathbf{s}_t\|^2 - \frac{1}{2\eta_t} \|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 + p(\widetilde{\mathbf{x}}_{t+1}) - p(\mathbf{x}_{t+1}) - \frac{\eta_t + \nu}{2} \|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 .
\end{aligned}
$$

So, by (3.14) and (3.17), $\widetilde{\mathbf{x}}_{t+1} = \beta_t \breve{\mathbf{x}}_{t+1} + (1-\beta_t)\mathbf{x}_t$, and the convexity of the function p, we have

$$
\begin{aligned}
& F(\mathbf{x}_{t+1}) = f(\mathbf{x}_{t+1}) + p(\mathbf{x}_{t+1}) \\
\leq \;& \beta_t \left[f(\mathbf{x}_t) + \langle \nabla f(\widehat{\mathbf{x}}_t), \breve{\mathbf{x}}_{t+1} - \mathbf{x}_t \rangle + p(\breve{\mathbf{x}}_{t+1}) \right] + (1-\beta_t) \left[f(\mathbf{x}_t) + p(\mathbf{x}_t) \right] \\
& + \frac{\mu_t}{2} \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \frac{\eta_t \beta_t^2}{2} \|\mathbf{s}_t\|^2 - \frac{\eta_t - \mathcal{L}}{2\eta_t^2} \|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\eta_t + \nu}{2} \|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 \\
= \;& \beta_t \Big[f(\mathbf{x}_t) + \langle \nabla f(\widehat{\mathbf{x}}_t), \breve{\mathbf{x}}_{t+1} - \mathbf{x}_t \rangle + \frac{\gamma_t}{2} \|\mathbf{s}_t\|^2 + p(\breve{\mathbf{x}}_{t+1}) \Big] + (1-\beta_t) F(\mathbf{x}_t) \\
& + \frac{\mu_t}{2} \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \frac{\eta_t \beta_t^2 - \gamma_t \beta_t}{2} \|\mathbf{s}_t\|^2 - \frac{\eta_t - \mathcal{L}}{2\eta_t^2} \|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 \\
& - \frac{\eta_t + \nu}{2} \|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 \\
= \;& \beta_t \Big[f(\mathbf{x}_t) + \langle \nabla f(\widehat{\mathbf{x}}_t), \breve{\mathbf{x}}_{t+1} - \mathbf{x}_t \rangle + \frac{\gamma_t}{2} \|\mathbf{s}_t\|^2 + p(\breve{\mathbf{x}}_{t+1}) \Big] + (1-\beta_t) F(\mathbf{x}_t) \\
& + \frac{\mu_t}{2} \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 - \frac{\eta_t - \mathcal{L}}{2\eta_t^2} \|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\eta_t + \nu}{2} \|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 , \qquad (3.19)
\end{aligned}
$$

where the last equality follows from $\gamma_t \beta_t - \eta_t \beta_t^2 = 0$. Now, it follows from

$$
\breve{\mathbf{x}}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \Big\{ \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x} \rangle + \frac{\gamma_t}{2} \|\mathbf{x} - \breve{\mathbf{x}}_t\|^2 + p(\mathbf{x}) \Big\} ,
$$

$\mathbf{s}_t = \breve{\mathbf{x}}_{t+1} - \breve{\mathbf{x}}_t$, $\mathbf{x}_t \in \mathcal{X}$ and (3.4) that

$$
\begin{aligned}
& \langle \nabla f(\widehat{\mathbf{x}}_t), \breve{\mathbf{x}}_{t+1} - \mathbf{x}_t \rangle + \frac{\gamma_t}{2} \|\mathbf{s}_t\|^2 + p(\breve{\mathbf{x}}_{t+1}) \\
\leq \;& \frac{\gamma_t}{2} \left(\|\mathbf{x}_t - \breve{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|^2 \right) + p(\mathbf{x}_t) - \frac{\nu}{2} \|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|^2 .
\end{aligned}
$$

Hence, by (3.19), we have

$$\begin{aligned}
F(\mathbf{x}_{t+1}) \;\; &\leq \;\; \beta_t \left[f(\mathbf{x}_t) + \frac{\gamma_t}{2} \left(\|\mathbf{x}_t - \breve{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|^2 \right) + p(\mathbf{x}_t) - \frac{\nu}{2} \|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|^2 \right] \\
&\quad + (1-\beta_t) F(\mathbf{x}_t) + \frac{\mu_t}{2} \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 - \frac{\eta_t - \mathcal{L}}{2\eta_t^2} \|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\eta_t + \nu}{2} \|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 \\
&\leq \;\; F(\mathbf{x}_t) + \frac{\mu_t}{2} \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \frac{\beta_t \gamma_t}{2} \left(\|\mathbf{x}_t - \breve{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|^2 \right) \\
&\quad - \frac{\beta_t \nu}{2} \|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|^2 - \frac{\eta_t - \mathcal{L}}{2\eta_t^2} \|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\eta_t + \nu}{2} \|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 . \qquad (3.20)
\end{aligned}$$

Note that

$$\breve{\mathbf{x}}_t - \mathbf{x}_t = \frac{1}{\beta_t}(\widehat{\mathbf{x}}_t - \mathbf{x}_t) \quad \text{and} \quad \breve{\mathbf{x}}_{t+1} - \mathbf{x}_t = \frac{1}{\beta_t}(\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t). \qquad (3.21)$$

Then, we have from (3.20) that (3.12) holds. □

Lemma 3.3.1 gives an important relationship on the function value reduction of F. It provides the results of how the objective function value reduces with respect to iterates. Based on Lemma 3.3.1, we have the following global convergence results of Algorithm 3.1 for the nonconvex case of problem (3.1).

Theorem 3.3.2. *Suppose the Assumptions 3.3.1 and 3.3.2 hold. Then, for the iterates generated by Algorithm 3.1, there exists an index $t_0 \geq 1$ such that*

$$E_{t+1} \leq E_t - \frac{L - \mathcal{L}}{8L^2} \|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - c\eta_t \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 - \frac{\beta_t \nu}{2} \|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 \qquad (3.22)$$

for all $t \geq t_0$, where $\widetilde{\mathbf{x}}_t$ is defined in (3.13) and $c > 0$ is a constant, and

$$E_t = F(\mathbf{x}_t) + \frac{\eta_{t-1}}{2} \|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{\eta_{t-1} + \nu}{2} \|\widetilde{\mathbf{x}}_t - \mathbf{x}_t\|^2 . \qquad (3.23)$$

Furthermore, if Assumption 3.3.3 holds, we have

$$\min_{t \in \{t_0, t_0+1, \ldots, T\}} \|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 \leq \frac{8L^2(E_{t_0} - \overline{F})}{L - \mathcal{L}} \frac{1}{T - t_0} = \mathcal{O}(1/T), \qquad (3.24)$$

where $\mathbf{g}(\widehat{\mathbf{x}}_t) = \eta_t(\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1})$ defined in (3.18).

Proof. For $t \geq 2$, by (3.21) we obtain

$$\begin{aligned}
\widehat{\mathbf{x}}_t - \mathbf{x}_t &= \beta_t(\breve{\mathbf{x}}_t - \mathbf{x}_t) = \beta_t\left((\breve{\mathbf{x}}_t - \mathbf{x}_{t-1}) + (\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t)\right) + \beta_t\left(\widetilde{\mathbf{x}}_t - \mathbf{x}_t\right) \\
&= \beta_t\left(\frac{1}{\beta_{t-1}}(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}) + (\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t)\right) + \beta_t\left(\widetilde{\mathbf{x}}_t - \mathbf{x}_t\right) \\
&= \theta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}) + \beta_t\left(\widetilde{\mathbf{x}}_t - \mathbf{x}_t\right), \qquad (3.25)
\end{aligned}$$

where $\theta_t = \frac{\beta_t}{\beta_{t-1}}(1 - \beta_{t-1})$. By defining $\beta_0 = 1$, $\mathbf{x}_0 = \mathbf{x}_1$ and $\widetilde{\mathbf{x}}_1 = \mathbf{x}_1$, we can see (3.25) also holds for $t = 1$. Hence, for $t \geq 1$, it follows from (3.12) and $L \leq \eta_t < 2L$ that

$$\begin{aligned}
F(\mathbf{x}_{t+1}) &\leq F(\mathbf{x}_t) + \frac{\gamma_t/\beta_t + \mu_t}{2}\|\theta_t\left(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\right) + \beta_t\left(\widetilde{\mathbf{x}}_t - \mathbf{x}_t\right)\|^2 - \frac{\gamma_t/\beta_t}{2}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 \\
&\quad - \frac{\beta_t\nu}{2\beta_t}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 - \frac{\eta_t - \mathcal{L}}{2\eta_t^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\eta_t + \nu}{2}\left\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\right\|^2 \\
&\leq F(\mathbf{x}_t) + \frac{\gamma_t/\beta_t + \mu_t}{2}\|\theta_t\left(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\right) + \beta_t\left(\widetilde{\mathbf{x}}_t - \mathbf{x}_t\right)\|^2 - \frac{\gamma_t/\beta_t}{2}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 \\
&\quad - \frac{\beta_t\nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 - \frac{L - \mathcal{L}}{8L^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\eta_t + \nu}{2}\left\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\right\|^2. \qquad (3.26)
\end{aligned}$$

Now, since $\mu_t = \min\{\mu_{t-1} + \rho^j - 1, L\}$ for some $\rho > 1$ and $j \geq 0$, it follows from $L > \mathcal{L} \geq \mu$ and (3.5) that the sequence $\{\mu_t\}$ is monotonically nondecreasing with upper bound $\mu_{up} = \min\{L, \rho(\mu + 1)\}$. Hence, μ_t can only be increased in finite, in fact at most $\lceil \mu_{up}/(\rho - 1) \rceil$, number of times. So, there exist $\overline{\mu} \geq 0$ and an integer $\overline{t} \geq 0$ such that $\mu_t = \overline{\mu}$ for all $t \geq \overline{t}$.

Since $\beta_t = \max\{\overline{\beta}_t, \tau_t\}$ and $\mu_t = \overline{\mu}$ for all $t \geq \overline{t}$, defining $\kappa = \overline{\mu}/L \in [0, 1]$, we have from Algorithm 3.1 that

$$\beta_t = \max\{\overline{\beta}_t, \overline{\tau}\}, \qquad (3.27)$$

for all $t \geq \overline{t}$, where $\overline{\beta}_t = 2/(t + 1)$, and

$$\overline{\tau} := \frac{\lambda}{2}\left(1 - \sqrt{\frac{1 - \kappa}{1 + \kappa}}\right) + \frac{(1 - \lambda)\kappa}{1 + \kappa} \in \left[0, \frac{1}{2}\right]. \qquad (3.28)$$

Hence, for all $t \geq \bar{t}$, we have from (3.27) that $\beta_{t+1} \leq \beta_t$, which gives

$$\eta_t = \gamma_t/\beta_t = 2L/(2-\beta_t) \geq 2L/(2-\beta_{t+1}) = \eta_{t+1} > L. \tag{3.29}$$

For all $t \geq \bar{t}+1$, it follows from (3.26) that

$$\begin{aligned} & F(\mathbf{x}_{t+1}) + \frac{\eta_t}{2}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 + \frac{\eta_t + \nu}{2}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 \\ \leq \; & F(\mathbf{x}_t) + \frac{\eta_{t-1}}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{\eta_{t-1}+\nu}{2}\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \\ & -\frac{L-\mathcal{L}}{8L^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\beta_t \nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 - R_t, \end{aligned} \tag{3.30}$$

where

$$\begin{aligned} R_t \; = \; & \frac{\eta_{t-1}}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{\eta_{t-1}+\nu}{2}\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \\ & -\frac{\eta_t + \overline{\mu}}{2}\|\theta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}) + \beta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_t)\|^2 \\ \geq \; & \frac{\eta_t}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{\eta_t}{2}\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \\ & -\frac{\eta_t + \overline{\mu}}{2}\|\theta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}) + \beta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_t)\|^2 \\ \geq \; & \frac{\eta_t - (\eta_t + \overline{\mu})\theta_t^2}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{\eta_t - (\eta_t+\overline{\mu})\beta_t^2}{2}\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \\ & -(\eta_t + \overline{\mu})\theta_t\beta_t\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|\,\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\| \end{aligned} \tag{3.31}$$

and the above second inequality follows from (3.29) and $\nu \geq 0$. We first show that when $t = 1$ or $t = 2$, it holds that

$$R_t \geq c\eta_t\left(\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2\right) \tag{3.32}$$

for $c = 1/2$. When $t = 1$, (3.32) holds for any $c > 0$ simply because our definition of $\mathbf{x}_0 = \widetilde{\mathbf{x}}_1 = \mathbf{x}_1$. When $t = 2$, (3.32) holds with $c = 1/2$ since $\widetilde{\mathbf{x}}_2 = \beta_1\breve{\mathbf{x}}_2 + (1-\beta_1)\mathbf{x}_1 = \mathbf{x}_2$ and $\theta_2 = \beta_2(1-\beta_1)/\beta_1 = 0$.

In the following, we divide our analysis into two cases on whether $\overline{\mu} > 0$ or whether $\overline{\mu} = 0$.

Case 1: $\overline{\mu} > 0$. Then, for all $t \geq \overline{t}$, we have from $\kappa = \overline{\mu}/L > 0$ and $\beta_t \geq \overline{\tau} > 0$ by (3.27) that

$$\kappa_t := \frac{\overline{\mu}}{\eta_t} = \frac{\overline{\mu}}{L}\frac{2-\beta_t}{2} \leq \kappa\frac{2-\overline{\tau}}{2} = \kappa - \frac{\kappa\overline{\tau}}{2}, \tag{3.33}$$

where $\overline{\tau}$ is defined in (3.28). In addition, for all $t \geq \widetilde{t} := \max\{\overline{t}+1, 3\}$, by (3.27), we have $\beta_t \leq 1/2$ and $\beta_t/\beta_{t-1} \geq t/(t+1) \geq 3/4$, which give

$$\theta_t = \frac{\beta_t}{\beta_{t-1}}(1-\beta_{t-1}) \geq \frac{3}{8}. \tag{3.34}$$

So, it follows from (3.31), (3.33) and (3.34) that

$$\begin{aligned}\frac{R_t}{\eta_t} \quad &\geq \quad \frac{1-(1+\kappa_t)\theta_t^2}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{1-(1+\kappa_t)\beta_t^2}{2}\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \\ &\quad\quad -(1+\kappa_t)\theta_t\beta_t\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|\,\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\| \\ &\geq \quad h_t + \frac{\kappa\overline{\tau}}{4}\left(\theta_t^2\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \beta_t^2\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2\right) \\ &\geq \quad h_t + c_1\left(\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2\right),\end{aligned} \tag{3.35}$$

for all $t \geq \widetilde{t}$, where

$$c_1 = \frac{\kappa\overline{\tau}}{4}\min\left\{\frac{9}{64}, \overline{\tau}^2\right\} > 0 \tag{3.36}$$

and

$$\begin{aligned}h_t \quad &= \quad \frac{1-(1+\kappa)\theta_t^2}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{1-(1+\kappa)\beta_t^2}{2}\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \\ &\quad\quad -(1+\kappa)\theta_t\beta_t\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|\,\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|.\end{aligned} \tag{3.37}$$

We now show $h_t \geq 0$ for all $t \geq \widetilde{t}$. By Cauchy-Schwarz inequality and (3.37), to show $h_t \geq \widetilde{t}$, it is sufficient to show

$$\left[1-(1+\kappa)\theta_t^2\right]\left[1-(1+\kappa)\beta_t^2\right] \geq (1+\kappa)^2\theta_t^2\beta_t^2, \tag{3.38}$$

which is equivalent to

$$1-(1+\kappa)\left(\theta_t^2+\beta_t^2\right)\geq 0. \tag{3.39}$$

Notice that for all $t\geq \widetilde{t}$, we have $\beta_t\leq\beta_{t-1}$. Hence, for all $t\geq\widetilde{t}$, we have

$$\theta_t=\frac{\beta_t}{\beta_{t-1}}(1-\beta_{t-1})\leq 1-\beta_t, \tag{3.40}$$

which gives

$$1-(1+\kappa)\left(\theta_t^2+\beta_t^2\right)\geq 1-(1+\kappa)\left((1-\beta_t)^2+\beta_t^2\right). \tag{3.41}$$

By the choice of β_t, we have $\frac{1}{2}\geq\beta_t\geq\overline{\tau}\geq\widetilde{\tau}>0$ for all $t\geq\widetilde{t}\geq 3$, where $\overline{\tau}$ is defined in (3.28) and $\widetilde{\tau}=\frac{1}{2}\left(1-\sqrt{(1-\kappa)/(1+\kappa)}\right)$, which implies

$$(1-\beta_t)^2+\beta_t^2\leq(1-\widetilde{\tau})^2+\widetilde{\tau}^2$$

for all $t\geq\widetilde{t}$. So, for all $t\geq\widetilde{t}$, we have from (3.41) and (3.28) that

$$1-(1+\kappa)\left(\theta_t^2+\beta_t^2\right)\geq 1-(1+\kappa)\left((1-\widetilde{\tau})^2+\widetilde{\tau}^2\right)=0.$$

Hence, (3.39) holds, which shows $h_t\geq 0$ and therefore (3.32) holds for all $t\geq\widetilde{t}$ with $c=c_1$ defined in (3.36). Since $c_1<1/2$, by (3.32), we have in fact (3.32) holds for all $t\geq\overline{t}+1$ with $c=c_1$. Then, (3.30) implies (3.22) holds with $c=c_1$ for all $t\geq\overline{t}+1$.

Case 2: $\overline{\mu}=0$. Then, we have $\mu_t=0$ and $\tau_t=0$ for all $t\geq 1$. So, $\overline{t}=1$ and for all $t\geq 1$, we have $\beta_t=\overline{\beta}_t=2/(t+1)$, $\gamma_t/\beta_t=\eta_t$ and $\eta_t=2L/(2-\beta_t)=L(t+1)/t$. In addition, we have $\theta_t=\frac{\beta_t}{\beta_{t-1}}(1-\beta_{t-1})=\frac{t-2}{t+1}<1-\beta_t$. So, it follows that

$$\begin{aligned}
&\|\theta_t\left(\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\right)+\beta_t\left(\widetilde{\mathbf{x}}_t-\mathbf{x}_t\right)\|^2\\
\leq\;&\left(\theta_t\|\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\|+\beta_t\|\widetilde{\mathbf{x}}_t-\mathbf{x}_t\|\right)^2\\
\leq\;&\left((1-\beta_t)\|\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\|+\beta_t\|\widetilde{\mathbf{x}}_t-\mathbf{x}_t\|\right)^2\\
\leq\;&(1-\beta_t)\|\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\|^2+\beta_t\|\widetilde{\mathbf{x}}_t-\mathbf{x}_t\|^2.
\end{aligned}$$

Hence, for all $t \geq 2$, we have from (3.31) and $\nu \geq 0$ that

$$\begin{aligned}
\frac{2R_t}{\eta_t} &= \frac{t^2}{t^2-1}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{t^2}{t^2-1}\left\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\right\|^2 - \left\|\theta_t\left(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\right) + \beta_t\left(\widetilde{\mathbf{x}}_t - \mathbf{x}_t\right)\right\|^2 \\
&\geq \left(\frac{t^2}{t^2-1} - \frac{t-1}{t+1}\right)\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \left(\frac{t^2}{t^2-1} - \frac{2}{t+1}\right)\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \\
&= \frac{2t-1}{t^2-1}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{t-1}{t+1}\left\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\right\|^2,
\end{aligned}$$

which implies for all $t \geq 3$,

$$R_t \geq \frac{L}{t}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{\eta_t}{4}\left\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\right\|^2. \tag{3.42}$$

Then, we have from (3.30), (3.32) and (3.42) that (3.22) holds with $c = 1/4$ for all $t \geq 1$.

Combing the above two cases, Case 1 and Case 2, we have (3.22) holds with $c = \min\{c_1, 1/4\} = c_1$ for all $t \geq t_0 := \bar{t}+1$, where c_1 is defined in (3.36). Finally, (3.24) follows from (3.22) and Assumption 3.3.3. □

There are a few remarks we would like to add here. In the results of Theorem 3.3.2, the inequality (3.22) provides a descent property of the potential function E_t, i.e., the function E_t is monotonically decreasing, which is a key relation for the global convergence of Algorithm 3.1. In addition, (3.24) gives the convergence rate $\mathcal{O}(1/T)$ of the square of the gradient mapping. Therefore, in at most $\mathcal{O}(1/\epsilon)$ iterations, we can find an ϵ-solution of the problem (3.1).

We say $\mathbf{x}^*$ is a stationary point of problem (3.1) if there exists a constant $\eta > 0$ such that

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{\langle\nabla f(\mathbf{x}^*), \mathbf{x}\rangle + \frac{\eta}{2}\left\|\mathbf{x} - \mathbf{x}^*\right\|^2 + p(\mathbf{x})\right\}. \tag{3.43}$$

Note that if (3.43) holds for some $\eta > 0$, then (3.43) holds for all $\eta > 0$. The following corollary follows directly from Theorem 3.3.2.

Corollary 3.3.1. *Suppose the Assumptions 3.3.1, 3.3.2 and 3.3.3 hold. Then, for the iterates generated by Algorithm 3.1, we have (i)*

$$\lim_{t\to\infty} \|\mathbf{g}(\widehat{\mathbf{x}}_t)\| = 0, \tag{3.44}$$

where $\mathbf{g}(\widehat{\mathbf{x}}_t) = \eta_t(\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1})$ *defined in (3.18) and (ii) the sequences* $\{\widehat{\mathbf{x}}_t\}$, $\{\mathbf{x}_t\}$ *and* $\{\widetilde{\mathbf{x}}_t\}$ *have the same set of cluster points, which are all stationary points of problem (3.1).*

Proof. By (3.27), we have $\lim_{t\to\infty} = \beta_t = \overline{\tau}$, which implies

$$\lim_{t\to\infty} \eta_t = \lim_{t\to\infty} 2L/(2-\beta_t) = 2L/(2-\overline{\tau}) =: \overline{\eta} \geq L. \tag{3.45}$$

Hence, by Theorem 3.3.2 and Assumption 3.3.3, we have

$$\sum_{t=t_0}^{\infty} (\|\mathbf{g}(\widehat{\mathbf{x}}_t)\| + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|) < \infty,$$

where $\mathbf{g}(\widehat{\mathbf{x}}_t)$ and $\widetilde{\mathbf{x}}_t$ are defined in (3.18) and (3.13), respectively. So, (3.44) holds and it follows from (3.45) and $\mathbf{g}(\widehat{\mathbf{x}}_t) = \eta_t(\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1})$ that

$$\lim_{t\to\infty} (\|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\| + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|) = 0. \tag{3.46}$$

Thus, the sequences $\{\widehat{\mathbf{x}}_t\}$, $\{\mathbf{x}_t\}$ and $\{\widetilde{\mathbf{x}}_t\}$ have the same set of cluster points. Now, given any cluster point $\widehat{\mathbf{x}}^*$ of $\{\widehat{\mathbf{x}}_t\}$, we can have from (3.15) (3.45), (3.46) and the closedness of p that

$$\widehat{\mathbf{x}}^* = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \nabla f(\widehat{\mathbf{x}}^*), \mathbf{x} \rangle + \frac{\overline{\eta}}{2} \|\mathbf{x} - \widehat{\mathbf{x}}^*\|^2 + p(\mathbf{x}) \right\}, \tag{3.47}$$

which by (3.43) implies $\widehat{\mathbf{x}}^*$ is a stationary point of problem (3.1). Hence, the statement (ii) holds. □

In terms of the convergence of the objective function value of problem (3.1), we have the following corollary.

Corollary 3.3.2. *Suppose the Assumptions 3.3.1, 3.3.2 and 3.3.3 hold.*

(i) If any of the following conditions holds:

(a) f is a convex function;

(b) $\overline{\mu} > 0$, where $\overline{\mu} = \lim_{t\to\infty} \mu_t$;

(c) $\nu > 0$, where ν is defined in (3.4);

(d) $\{\mathbf{x}_t\}$ and $\{\breve{\mathbf{x}}_t\}$ are bounded, e.g., when $\mathcal{X}$ is a bounded set,

we have $\lim_{t\to\infty} F(\mathbf{x}_t)$ exists.

(ii) If $F^* = \lim_{t\to\infty} F(\mathbf{x}_t)$, then for any cluster point $\overline{\mathbf{x}}$ of $\{\mathbf{x}_t\}$, we have $F(\overline{\mathbf{x}}) = F^*$.

Proof. We first show (i) by considering the cases (a), (b), (c) and (d) separately.

Case (a): By Assumption 3.3.3, there exists an F^* such that $F^* = \liminf_{t\to\infty} F(\mathbf{x}_t)$. Then, when f is a convex function, it follows from (3.11) that for any $\epsilon > 0$ and $\mathbf{x}_\epsilon$ such that $F(\mathbf{x}_\epsilon) \leq F^* + \epsilon$, we have $\lim_{t\to\infty} F(\mathbf{x}_t) \leq F^* + \epsilon$, which implies $\lim_{t\to\infty} F(\mathbf{x}_t) \leq F^*$. Hence, $\lim_{t\to\infty} F(\mathbf{x}_t) = F^*$.

Case (b): Since μ_t is monotonically nondescreasing and has upper bound L, we have $\overline{\mu} = \lim_{t\to\infty} \mu_t$ exists. If $\overline{\mu} > 0$, we have from (3.30), (3.32) and (3.35) that

$$\begin{aligned} E_{t+1} \quad \leq \quad & E_t - \frac{L-\mathcal{L}}{8L^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - c_1\eta_t\left(\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2\right) \\ & -\frac{\beta_t\nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 \end{aligned} \tag{3.48}$$

for all $t \geq \bar{t} + 1$, where E_t is defined in (3.23) and $c_1 > 0$ is a constant given in (3.36). Then, by (3.48), Assumption 3.3.3 and $\eta_t \in [L, 2L]$, we have

$$\lim_{t\to\infty} \left(\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\| + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|\right) = 0 \tag{3.49}$$

and there exists an F^* such that

$$\lim_{t\to\infty} F(\mathbf{x}_t) = \lim_{t\to\infty} E_t = F^*. \tag{3.50}$$

Case (c): If $\nu > 0$, it follows from Theorem 3.3.2 and Assumption 3.3.3 that

$$\lim_{t\to\infty} (\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\| + \|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|) = 0.$$

Then, by (3.21), (3.49) also holds and hence (3.50) holds by Theorem 3.3.2.

Case (d): Suppose the sequences $\{\mathbf{x}_t\}$ and $\{\breve{\mathbf{x}}_t\}$ are bounded. If $\overline{\mu} > 0$, the result follows from Case (b). So, we only consider the case when $\overline{\mu} = 0$, which gives $\beta_t = 2/(t+1)$ and therefore $\lim_{t\to\infty} \beta_t = 0$. By (3.13), $\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t = \beta_t(\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t)$. Hence, we have $\lim_{t\to\infty} \|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t\| = 0$ from the boundedness of $\{\mathbf{x}_t\}$ and $\{\breve{\mathbf{x}}_t\}$, which together with (3.46) implies (3.49) holds. Hence, (3.50) also holds.

Now, we show (ii) holds. By (3.15), for any $\mathbf{z} \in \mathcal{X}$, we have

$$\begin{aligned} &\langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \mathbf{z}\rangle + p(\mathbf{x}_{t+1}) \\ \leq\ & \frac{\eta_t}{2}\left(\|\mathbf{z} - \widehat{\mathbf{x}}_t\|^2 - \|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|^2\right) + p(\mathbf{z}) - \frac{\eta_t + \nu}{2}\|\mathbf{x}_{t+1} - \mathbf{z}\|^2, \end{aligned}$$

which by $\nu \geq 0$ gives

$$\begin{aligned} &f(\widehat{\mathbf{x}}_t) + \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\rangle + \frac{\eta_t}{2}\|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|^2 + p(\mathbf{x}_{t+1}) \\ \leq\ & \frac{\eta_t}{2}\|\mathbf{z} - \widehat{\mathbf{x}}_t\|^2 + f(\widehat{\mathbf{x}}_t) + \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{z} - \widehat{\mathbf{x}}_t\rangle + p(\mathbf{z}). \end{aligned} \tag{3.51}$$

For any $\mathbf{z} \in \mathcal{X}$, it follows from Assumption 3.3.1 that

$$|f(\mathbf{z}) - f(\widehat{\mathbf{x}}_t) - \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{z} - \widehat{\mathbf{x}}_t\rangle| \leq \frac{\mathcal{L}}{2}\|\mathbf{z} - \widehat{\mathbf{x}}_t\|^2.$$

Hence, by (3.51), $\eta_t \in [L, 2L]$ and $L > \mathcal{L}$, for any $\mathbf{z} \in \mathcal{X}$, we have

$$\begin{aligned} F(\mathbf{x}_{t+1}) &= f(\mathbf{x}_{t+1}) + p(\mathbf{x}_{t+1}) \le F(\mathbf{z}) + \frac{3L}{2} \left\| \mathbf{z} - \widehat{\mathbf{x}}_t \right\|^2 \\ &\le F(\mathbf{z}) + 3L \left\| \mathbf{z} - \mathbf{x}_{t+1} \right\|^2 + 3L \left\| \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t \right\|^2 . \end{aligned} \tag{3.52}$$

Then, for any subsequence $\{\mathbf{x}_{t_i+1}\}$ of $\{\mathbf{x}_t\}$ converging to $\overline{\mathbf{x}} \in \mathcal{X}$, we have from (3.52) that

$$F(\mathbf{x}_{t_i+1}) \le F(\overline{\mathbf{x}}) + 3L \left\| \overline{\mathbf{x}} - \mathbf{x}_{t_i+1} \right\|^2 + 3L \left\| \mathbf{x}_{t_i+1} - \widehat{\mathbf{x}}_{t_i} \right\|^2 .$$

Taking i to infinity in the above inequality, we have from $\lim_{i\to\infty} \mathbf{x}_{t_i+1} = \overline{\mathbf{x}}$, (3.46) and part (i) that $F^* = \lim_{i\to\infty} F(\mathbf{x}_{t_i+1}) \le F(\overline{\mathbf{x}})$. In addition, by the lower semicontinuity of F, we have $F(\overline{\mathbf{x}}) \le \lim_{i\to\infty} F(\mathbf{x}_{t_i+1}) = F^*$. Hence, we have $F(\overline{\mathbf{x}}) = F^*$. □

3.4. Linear Convergence

In this section, we discuss the linear convergence of $\{\mathbf{x}_t\}$ and $\{F(\mathbf{x}_t)\}$. Let us define $h(\mathbf{x}) = p(\mathbf{x}) + \delta_{\mathcal{X}}(\mathbf{x})$, where $\delta_{\mathcal{X}}(\mathbf{x})$ is the indicator function on the set $\mathcal{X}$. Let Ω be the set of all stationary points of problem (3.1), i.e.,

$$\Omega = \{\mathbf{x}^* \in \mathcal{X} : -\nabla f(\mathbf{x}^*) \in \partial h(\mathbf{x}^*)\} = \{\mathbf{x}^* \in \mathcal{X} : \mathbf{x}^* \text{ satisfies (3.43)}\}. \tag{3.53}$$

Note that Ω is a closed set. Recall the definition of proximal operator in Definition 2.1.23 in Chapter 2. Then we have from Algorithm 3.1 that

$$\mathbf{x}_{t+1} = \mathrm{prox}_{\frac{1}{\eta_t},h} \left(\widehat{\mathbf{x}}_t - \frac{1}{\eta_t} \nabla f(\widehat{\mathbf{x}}_t) \right). \tag{3.54}$$

For studying linear convergence, we need the following further assumptions in this section.

Assumption 3.4.1. *(a) For any $\xi \ge \inf_{x\in\mathcal{X}} F(\mathbf{x})$, there exist $\epsilon > 0$ and $\sigma > 0$ such that*

$$dist(\mathbf{x}, \Omega) \le \sigma \left\| prox_{\frac{1}{\eta},h} \left(\mathbf{x} - \frac{1}{\eta} \nabla f(\mathbf{x}) \right) - \mathbf{x} \right\|, \tag{3.55}$$

whenever $\left\| prox_{\frac{1}{\eta},h}\left(\mathbf{x}-\frac{1}{\eta}\nabla f(\mathbf{x})\right)-\mathbf{x}\right\| < \epsilon$*,* $F(\mathbf{x}) < \xi$ *and* $\eta \in [L, 2L]$*.*

(b) Ω *is nonempty and there exists* $\omega > 0$ *such that* $\|\mathbf{x}-\mathbf{y}\| \geq \omega$ *whenever* $\mathbf{x}$*,* $\mathbf{y} \in \Omega$ *and* $F(\mathbf{x}) \neq F(\mathbf{y})$*.*

Assumption 3.4.1 (a) provides a local error bound condition. The distance from a point $\mathbf{x}$, which is in the neighborhood of the solutions of the proximal minimization problem, to the solution set Ω can be bounded by its gradient mapping up to a multiplicative factor. Assumption 3.4.1 can be satisfied when (i) $f(\mathbf{x}) = \phi(A\mathbf{x})$ with strongly convex ϕ or f is quadratic (possibly nonconvex); (ii) h is a polyhedral function.

According to the assumption, we have the following lemma.

Lemma 3.4.1. *Suppose the Assumptions 3.3.1, 3.3.2, 3.3.3 and 3.4.1 hold. We have*

(i) $\lim_{t\to\infty} dist(\mathbf{x}_t, \Omega) = 0$*;*

(ii) in addition, if there exists a constant $\widetilde{c} > 0$ *such that for all* t *sufficiently large, we have*

$$\widetilde{E}_{t+1} \leq \widetilde{E}_t - \widetilde{c}d_t, \tag{3.56}$$

where

$$\begin{aligned}\widetilde{E}_t &= F(\mathbf{x}_t) + \frac{\eta_{t-1}}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{\eta_{t-1}+\nu}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_t\|^2 \\ &\quad + \frac{L-\mathcal{L}}{8}\|\mathbf{x}_t - \widehat{\mathbf{x}}_{t-1}\|^2 + \frac{\beta_{t-1}\nu}{2}\|\breve{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 \end{aligned} \tag{3.57}$$

and

$$\begin{aligned} d_t &= \|\mathbf{x}_t - \widehat{\mathbf{x}}_{t-1}\|^2 + \|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \\ &\quad + \beta_{t-1}\nu\|\breve{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2, \end{aligned} \tag{3.58}$$

then for t sufficiently large, we have

$$|F(\mathbf{x}_t) - F^*| \leq \overline{\theta}\|\mathbf{x}_t - \widehat{\mathbf{x}}_{t-1}\|^2 \tag{3.59}$$

and

$$0 \leq \widetilde{E}_{t+1} - F^* \leq \theta(\widetilde{E}_t - F^*), \tag{3.60}$$

where $\overline{\theta} > 0$ and $\theta \in (0,1)$ are constants and $F^* = \lim_{t\to\infty} F(\mathbf{x}_t) = \lim_{t\to\infty} \widetilde{E}_t$.

Proof. By Theorem 3.3.2, there exists a $\xi > 0$ such that $E_t \leq \xi$ for all $t \geq 1$, which implies $F(\mathbf{x}_t) \leq \xi$ for all $t \geq 1$. In addition, by (3.44) and (3.54), we have

$$\begin{aligned} 0 &= \lim_{t\to\infty} \|\mathbf{g}(\widehat{\mathbf{x}}_t)\| = \lim_{t\to\infty} \|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\| \\ &= \lim_{t\to\infty} \left\|\operatorname{prox}_{\frac{1}{\eta_t},h}\left(\widehat{\mathbf{x}}_t - \frac{1}{\eta_t}\nabla f(\widehat{\mathbf{x}}_t)\right) - \widehat{\mathbf{x}}_t\right\|. \end{aligned} \tag{3.61}$$

By the nonexpansion property of the proximal operator, (3.54), $\eta_t > L > \mathcal{L}$ and Assumption 3.3.1, we have

$$\begin{aligned} &\left\|\operatorname{prox}_{\frac{1}{\eta_t},h}\left(\mathbf{x}_{t+1} - \frac{1}{\eta_t}\nabla f(\mathbf{x}_{t+1})\right) - \mathbf{x}_{t+1}\right\| \\ = &\left\|\operatorname{prox}_{\frac{1}{\eta_t},h}\left(\mathbf{x}_{t+1} - \frac{1}{\eta_t}\nabla f(\mathbf{x}_{t+1})\right) - \operatorname{prox}_{\frac{1}{\eta_t},h}\left(\widehat{\mathbf{x}}_t - \frac{1}{\eta_t}\nabla f(\widehat{\mathbf{x}}_t)\right)\right\| \\ \leq &\left\|\left(\mathbf{x}_{t+1} - \frac{1}{\eta_t}\nabla f(\mathbf{x}_{t+1})\right) - \left(\widehat{\mathbf{x}}_t - \frac{1}{\eta_t}\nabla f(\widehat{\mathbf{x}}_t)\right)\right\| \\ \leq &\left(1 + \frac{\mathcal{L}}{\eta_t}\right)\|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\| \leq 2\|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|. \end{aligned}$$

Hence, it follows from $\eta_t \in [L, 2L]$ and Assumption 3.4.1 (a) and (3.61) that

$$\begin{aligned} \operatorname{dist}(\mathbf{x}_{t+1}, \Omega) &\leq \sigma\left\|\operatorname{prox}_{\frac{1}{\eta_t},h}\left(\mathbf{x}_{t+1} - \frac{1}{\eta_t}\nabla f(\mathbf{x}_{t+1})\right) - \mathbf{x}_{t+1}\right\| \\ &\leq 2\sigma\|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|, \end{aligned} \tag{3.62}$$

for t sufficiently large. So, we have (i) holds by (3.61).

Now, we prove (ii). Let us define $\overline{\mathbf{x}}_t \in \Omega$ such that $\mathrm{dist}(\mathbf{x}_t, \Omega) = \|\mathbf{x}_t - \overline{\mathbf{x}}_t\|$. By (3.56) and Assumption 3.3.3, we have $\lim_{t\to\infty} d_t = 0$, which gives

$$\lim_{t\to\infty} \|\mathbf{x}_t - \mathbf{x}_{t-1}\| \leq \lim_{t\to\infty} \left(\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\| + \|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|\right) \leq \lim_{t\to\infty} \sqrt{2d_t} = 0.$$

Hence, we have from property (i) that

$$\lim_{t\to\infty} \|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_{t-1}\| \leq \lim_{t\to\infty} \|\overline{\mathbf{x}}_t - \mathbf{x}_t\| + \|\mathbf{x}_t - \mathbf{x}_{t-1}\| + \|\mathbf{x}_{t-1} - \overline{\mathbf{x}}_{t-1}\| = 0.$$

This together with the Assumption 3.4.1 (b) implies that $F(\overline{\mathbf{x}}_t) = F^*$ for all t sufficiently large, where F^* is some constant. Hence, for t sufficiently large, replacing $t+1$ by t and taking $\mathbf{z} = \overline{\mathbf{x}}_t$ in (3.52), we have

$$\begin{aligned} F(\mathbf{x}_t) - F^* &\leq 3L\left\|\overline{\mathbf{x}}_t - \mathbf{x}_t\right\|^2 + 3L\left\|\mathbf{x}_t - \widehat{\mathbf{x}}_{t-1}\right\|^2 \\ &= 3L\mathrm{dist}(\mathbf{x}_t, \Omega)^2 + 3L\left\|\mathbf{x}_t - \widehat{\mathbf{x}}_{t-1}\right\|^2 \\ &\leq (12\sigma^2 + 3)L\left\|\mathbf{x}_t - \widehat{\mathbf{x}}_{t-1}\right\|^2, \end{aligned} \tag{3.63}$$

where the last inequality follows from (3.62). On the other hand, since $\overline{\mathbf{x}}_t \in \Omega$, we have from (3.43) that

$$\langle \nabla f(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_t \rangle + p(\overline{\mathbf{x}}_t) \leq \langle \nabla f(\overline{\mathbf{x}}_t), \mathbf{x}_t \rangle + \frac{\eta}{2}\left\|\mathbf{x}_t - \overline{\mathbf{x}}_t\right\|^2 + p(\mathbf{x}_t)$$

for some $\eta > 0$, which by Assumption 3.3.1 and (3.62) gives

$$\begin{aligned} F^* &= F(\overline{\mathbf{x}}_t) = f(\overline{\mathbf{x}}_t) + p(\overline{\mathbf{x}}_t) \\ &\leq f(\overline{\mathbf{x}}_t) + \langle \nabla f(\overline{\mathbf{x}}_t), \mathbf{x}_t - \overline{\mathbf{x}}_t \rangle + \frac{\eta}{2}\left\|\mathbf{x}_t - \overline{\mathbf{x}}_t\right\|^2 + p(\mathbf{x}_t) \\ &\leq f(\mathbf{x}_t) + p(\mathbf{x}_t) + \frac{\mathcal{L} + \eta}{2}\left\|\mathbf{x}_t - \overline{\mathbf{x}}_t\right\|^2 \\ &= F(\mathbf{x}_t) + \frac{\mathcal{L} + \eta}{2}\mathrm{dist}(\mathbf{x}_t, \Omega)^2 \\ &\leq F(\mathbf{x}_t) + 2(\mathcal{L} + \eta)\sigma^2\left\|\mathbf{x}_t - \widehat{\mathbf{x}}_{t-1}\right\|^2. \end{aligned} \tag{3.64}$$

So, by (3.63) and (3.64), we have (3.59) holds. In addition, it follows from (3.56), $\lim_{t\to\infty} d_t = 0$, (3.61) and (3.59) that

$$\lim_{t\to\infty} \widetilde{E}_t = \lim_{t\to\infty} F(\mathbf{x}_t) = F^*$$

and $\widetilde{E}_t \geq F^*$ for all t sufficiently large. So, by (3.63) and the definitions of $\widetilde{E}_t$ and d_t in (3.57) and (3.58), there exists a constant $c > 0$ such that $0 \leq (\widetilde{E}_t - F^*) \leq cd_t$ for t sufficiently large. Therefore, by (3.56) we have (3.60) holds with $\theta = (c-1)/c \in (0,1)$. □

From the above lemma, we see that (3.56) shows that the new energy function $\widetilde{E}_t$ is monotonically decreasing, and by (3.60), the sequence $\{\widetilde{E}_t\}$ converges Q-linearly to F^*. Furthermore, (3.59) gives a bound on the objective function value gap.

Based on the Lemma 3.4.1, we can have the following linear convergence result.

Theorem 3.4.1. *Suppose the Assumptions 3.3.1, 3.3.2, 3.3.3 and 3.4.1 hold. If any of the following conditions holds:*

(a) $\overline{\mu} > 0$, where $\overline{\mu} = \lim \mu_t$;

(b) $\nu > 0$, where ν is defined in (3.5);

(c) restart Algorithm 3.1 after every $k_0 > 0$ iterations,

we have

(i) the sequence $\{F(\mathbf{x}_t)\}$ converges R-linearly;

(ii) the sequence $\{\mathbf{x}_t\}$ converges R-linearly to a stationary point of problem (3.1).

Proof. If $\overline{\mu} > 0$, we have (3.48) holds, which together with $\mathbf{g}(\widehat{\mathbf{x}}_t) = \eta_t(\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1})$ and $\eta_t \geq L$

gives

$$\begin{aligned} E_{t+1} &\leq E_t - \frac{L-\mathcal{L}}{8}\|\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2 - c_1 L\left(\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2\right) \\ &\quad -\frac{\beta_t \nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 \end{aligned}$$

for all $t \geq t_0 = \bar{t} + 1$, where c_1 is defined in (3.36). By rearranging the above inequality with the definition of $\widetilde{E}_t$ in (3.57), we have

$$\begin{aligned} \widetilde{E}_{t+1} &\leq \widetilde{E}_t - \frac{L-\mathcal{L}}{8}\|\mathbf{x}_t - \widehat{\mathbf{x}}_{t-1}\|^2 - c_1 L\left(\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2\right) \\ &\quad -\frac{\beta_{t-1}\nu}{2}\|\breve{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2, \end{aligned}$$

which implies (3.56) holds with $\widetilde{c} = \min\{(L-\mathcal{L})/8, c_1 L, 1/2\}$ when $t \geq \bar{t}+1$.

Similarly, if $\nu > 0$, it follows from (3.22) that

$$E_{t+1} \leq E_t - \frac{L-\mathcal{L}}{8}\|\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2 - c_1 L\,\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 - \frac{\beta_t\nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2$$

for all $t \geq t_0 = \bar{t}+1$. By rearranging this inequality with the definition of $\widetilde{E}_t$ in (3.57), we have

$$\widetilde{E}_{t+1} \leq \widetilde{E}_t - \frac{L-\mathcal{L}}{8}\|\mathbf{x}_t - \widehat{\mathbf{x}}_{t-1}\|^2 - c_1 L\,\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 - \frac{\beta_{t-1}\nu}{2}\|\breve{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2,$$

which together with $\beta_t \in (0,1]$ and $\breve{\mathbf{x}}_t - \mathbf{x}_{t-1} = 1/\beta_{t-1}(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1})$ implies

$$\begin{aligned} \widetilde{E}_{t+1} &\leq \widetilde{E}_t - \frac{L-\mathcal{L}}{8}\|\mathbf{x}_t - \widehat{\mathbf{x}}_{t-1}\|^2 - c_1 L\,\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 - \frac{\beta_{t-1}\nu}{4}\|\breve{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 \\ &\quad -\frac{\nu}{4}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2. \end{aligned}$$

Hence, we have from $\nu > 0$ that (3.56) holds with $\widetilde{c} = \min\{(L-\mathcal{L})/8, c_1 L, 1/4, \nu/4\} > 0$ when $t \geq \bar{t}+1$.

By the previous analysis, under condition (a), we have (3.56) holds for sufficiently large t. So, by Lemma 3.4.1, we have (3.59) and (3.60) hold for t sufficiently large. Hence, by (3.56), (3.59), (3.60) and the definition of d_t, for t sufficiently large, we have

$$|F(\mathbf{x}_t) - F^*| \leq \overline{\theta}\|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|^2 \leq \overline{\theta} d_t \leq \frac{\overline{\theta}}{\widetilde{c}}(\widetilde{E}_t - \widetilde{E}_{t+1}) \leq \frac{\overline{\theta}}{\widetilde{c}}(\widetilde{E}_t - F^*),$$

which together with (3.60) implies the R-linear convergence of $F(\mathbf{x}_t)$ to F^*, i.e., conclusion (i) holds. By (3.56), (3.60) and the definition of d_t, we also have

$$\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \leq 2(\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2) \leq 2d_t \leq \frac{2}{\widetilde{c}}(\widetilde{E}_t - F^*),$$

for t sufficiently large. This inequality and (3.60) show R-linear convergence of $\|\mathbf{x}_t - \mathbf{x}_{t-1}\|$, which implies there exists an $\mathbf{x}^*$ such that the sequence $\{\mathbf{x}_t\}$ converges to $\mathbf{x}^*$ R-linearly. Finally, the conclusion (i) of Lemma 3.4.1 shows $\mathbf{x}^*$ is a stationary point of problem (3.1). Hence, conclusion (ii) holds. Under condition (c), we have $\overline{\mu} > 0$, then the conclusion (i) and (ii) follow from the previous analysis. □

3.5. Numerical Experiments

In this section, we evaluate the performance of the unified proximal gradient method (UPG) in Algorithm 3.1 by solving two nonconvex optimization problems: the smoothly clipped absolute deviation (SCAD) penalty problems and nonconvex quadratic programming with simplex constraints. Note that, depending on the selection of the matrix G, the latter problem can be convex if G is chosen to be positive semidefinite. We compare UPG with three other algorithms: proximal gradient method (PG), FISTA and proximal gradient algorithm with extrapolation (PGE), where the proximal gradient method comes from the proximal gradient algorithm with extrapolation by setting the parameter $\beta_k = 0$ in Algorithm 1 in [123]. The complete description of these algorithms can be found in the literature as well.

3.5.1. The Smoothly Clipped Absolute Deviation Penalty Problem

In this subsection, we apply Algorithm 3.1 to solve the smoothly clipped absolute deviation (SCAD) penalty problem, which is defined as

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|^2 + \sum_{i=1}^{n} g_\kappa(|\mathbf{x}_i|), \tag{3.65}$$

where $A \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$ and g_κ is the SCAD penalty given by

$$g_\kappa(\theta) = \begin{cases} \kappa\theta, & \theta \le \kappa, \\ \frac{-\theta^2+2c\kappa\theta-\kappa^2}{2(c-1)}, & \kappa < \theta \le c\kappa, \\ \frac{(c+1)\kappa^2}{2}, & \theta > c\kappa, \end{cases} \tag{3.66}$$

with parameters $c > 2$ and $\kappa > 0$. The SCAD penalty corresponds to a quadratic spline function with knots at κ and $c\kappa$, and combines the benefits of using l_1 penalty and hard thresholding penalty. We refer to [41] for more details about SCAD penalty. The SCAD penalty is used in statistics and applications involving regularization, especially for penalization when the noise level in data is low [41] and doing variable selection [138, 127].

The SCAD problem (3.65) is nonconvex as the SCAD penalty is a nonconvex function. However, it is required in the problem (3.1) that the function p is convex. Fortunately, it is proved in [64] that $g_\kappa(\cdot) + \frac{\omega}{2}|\cdot|^2$ with $\omega \ge \frac{1}{c-1}$ is convex. Therefore, we can rewrite problem (3.65) into the form of (3.1) with

$$f(\mathbf{x}) := \frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|^2 - \frac{1}{2(c-1)}\|\mathbf{x}\|^2 \text{ and } p(\mathbf{x}) := \sum_{i=1}^{n} g_\kappa(|\mathbf{x}_i|) + \frac{1}{2(c-1)}\|\mathbf{x}\|^2.$$

Hence, f is Lipschitz continuously differentiable but possibly nonconvex and p is a convex function, which satisfy the problem settings of (3.1). Then, we can apply UPG, PG, PGE and FISTA to solve this reformulated problem. We have to point out here FISTA does not guarantee convergence since the objective function is nonconvex. We apply FISTA here simply for practical numerical comparison purpose.

When applying Algorithm 3.1, the two subproblems in Step 4 and 6 can be rewritten as

$$\breve{\mathbf{x}}_{t+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2\nu_1} \left\| \mathbf{x} - \nu_1 \left(\gamma_t \breve{\mathbf{x}}_t - \nabla f(\widehat{\mathbf{x}}_t) \right) \right\|^2 + \sum_{i=1}^{n} g_\kappa(|\mathbf{x}_i|) \right\},$$

and

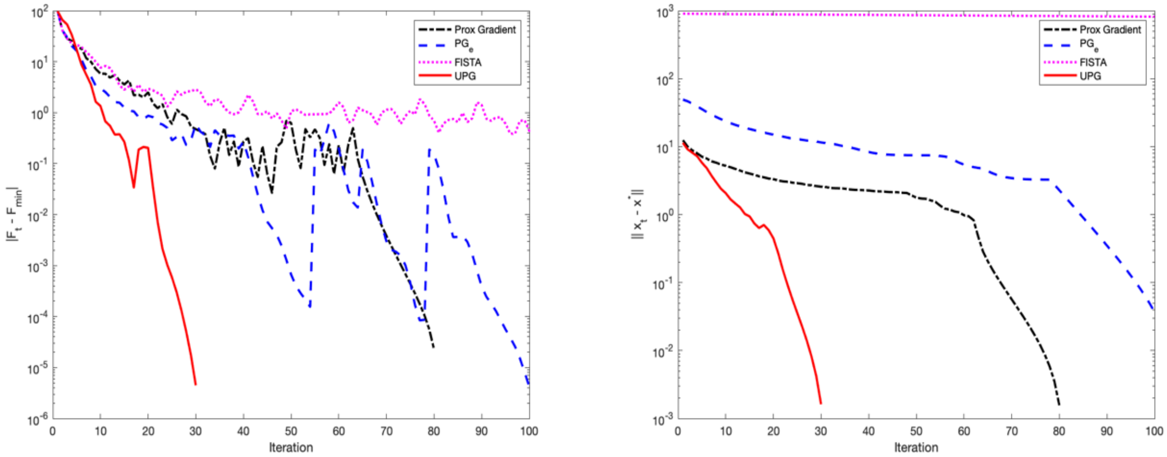
$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2\nu_2} \left\| \mathbf{x} - \nu_2 \left(\eta_t \widehat{\mathbf{x}}_t - \nabla f(\widehat{\mathbf{x}}_t) \right) \right\|^2 + \sum_{i=1}^{n} g_\kappa(|\mathbf{x}_i|) \right\},$$

where $\nu_1 = \frac{c-1}{\gamma_t(c-1)+1}$ and $\nu_2 = \frac{c-1}{\eta_t(c-1)+1}$. In addition, it can be easily verified that $1+\nu_i \le c$ holds for $i = 1, 2$ and it is given in [125] that the solution of the following minimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2\nu} \|\mathbf{x} - \mathbf{q}\|^2 + \sum_{i=1}^{n} g_\kappa(|\mathbf{x}_i|) \tag{3.67}$$

with $1 + \nu \le c$ and known $\mathbf{q}$ has closed form. Hence, the subproblems for finding $\breve{\mathbf{x}}_{t+1}$ and $\mathbf{x}_{t+1}$ in Algorithm 3.1 can be solved trivially. Also note that the subproblems in PG, PGE and FISTA can be also converted into the form of (3.67) and therefore, closed form solutions are guaranteed as well.

In our experiment, the dimension is set to be $m = 100$ and $n = 500$. We randomly generate $A \in \mathbb{R}^{m\times n}$ with entries from standard normal distribution and then normalize its columns. Then, the vector $\mathbf{b}$ is obtained as $\mathbf{b} = A\mathbf{b}^* + \epsilon$ where $\mathbf{b}^*$ is a sparse uniformly distributed random vector in $\mathbb{R}^n$ being generated with density of 0.02 and ϵ is a noise vector in $\mathbb{R}^m$ with entries from $\mathcal{N}(0, 0.01)$. The parameters c and κ can be chosen by cross-validation in practice. Here, we simply choose $c = 3.7$ and $\kappa = 0.1$ since the purpose of this numerical example is to show the efficiency of the unified proximal gradient



(a) Function value gap against iterations

(b) Minimizer gap against iterations

Figure 3.1. Comparison of UPG, PG, PGE and FISTA for the SCAD problem

algorithm. We also specify $L = \max(|\lambda_H|)$, where λ_H corresponds to all the eigenvalues of matrix H and H is the Hessian of $f(\mathbf{x})$, for PG, FISTA, PGE and UPG. As suggested in [123], we set $l = |\min(\lambda_H)|$ and $\beta_t = 0.85\sqrt{\frac{L}{L+l}}$ for PGE. Additionally, we choose $\rho = 1.5$ and $\lambda = 0.5$ for UPG. The initialization of $\mathbf{x} \in \mathbb{R}^n$ for all the algorithms is randomly selected from standard uniform distribution in $(0, 1)$. The stopping condition for all four algorithms is

$$\frac{\|\mathbf{x}_{t+1} - \mathbf{x}_t\|}{\max(\|\mathbf{x}_{t+1}\|, 1)} \leq 10^{-4}, \tag{3.68}$$

and the maximal iteration number is set to 1000.

The results are presented in Figure 3.1. The plot on the left is the gap of the objective function values $|F_t - F_{\min}|$ versus iterations, where $F_{\min}$ is the minimum of the sequence $\{F_t\}$ generated by each algorithm and F_t is the abbreviation of $F(\mathbf{x}_t)$. We also plot $\|\mathbf{x}_t - \mathbf{x}^*\|$ against iterations in Figure 3.1 (b), where $\mathbf{x}^*$ is the solution corresponding to the minimum $F_{\min}$ of each algorithm. We can see from Figure 3.1 that UPG greatly outperforms the other algorithms. UPG converges in much fewer iterations and approaches the desired solution rapidly. One can also find that FISTA does not converge as we mentioned previously FISTA is not proved to be convergent for nonconvex problems. Additionally, we can observe the R-linearly convergence of $\{\mathbf{x}_t\}$ in Figure 3.1 (b), which matches the theoretical result in Theorem 3.4.1.

3.5.2. Nonconvex Quadratic Programming with Simplex Constraints

In this subsection, we consider the following possibly nonconvex problem

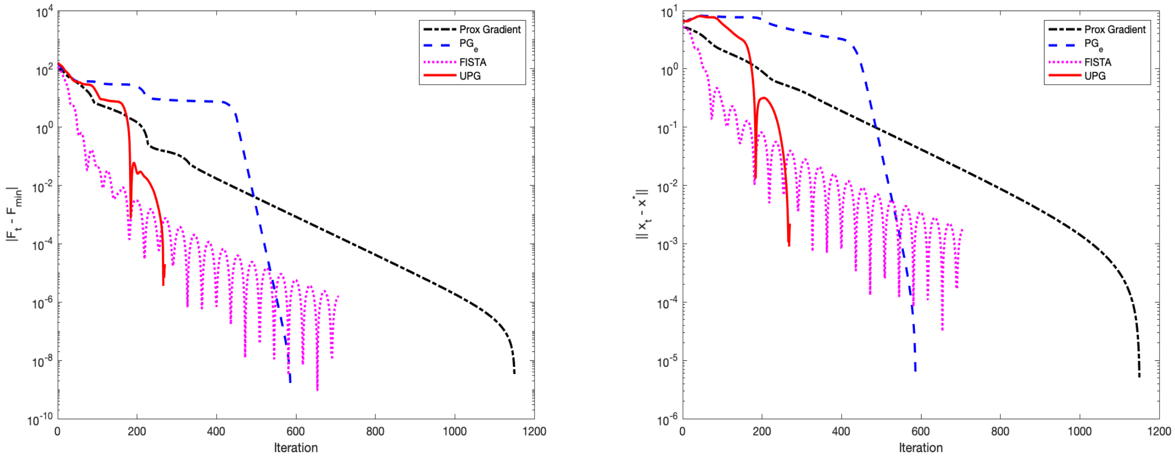
$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \ & \frac{1}{2}\mathbf{x}^\mathsf{T} G\mathbf{x} - \mathbf{g}^\mathsf{T}\mathbf{x} \\ \text{s.t.} \quad & \mathbf{e}^\mathsf{T}\mathbf{x} = c, \quad \mathbf{x} \geq \mathbf{0}, \end{aligned} \tag{3.69}$$

where $G \in \mathbb{R}^{n\times n}$ is not necessarily positive semidefinite, $\mathbf{g} \in \mathbb{R}^n$, $\mathbf{e}$ is a vector of ones in $\mathbb{R}^n$ and c is a positive number. We can easily rewrite (3.69) in the form of (3.1) with

$$f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\mathsf{T} G\mathbf{x} - \mathbf{g}^\mathsf{T}\mathbf{x} \qquad \text{and} \qquad p(\mathbf{x}) := \delta_{\mathcal{C}}(\mathbf{x}),$$

where $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{e}^\mathsf{T}\mathbf{x} = c,\ \mathbf{x} \geq \mathbf{0}\}$ and $\delta_{\mathcal{C}}(\cdot)$ is the indicator function of the simplex $\mathcal{C}$. Note that p is convex as $\mathcal{C}$ is a convex set. Nonconvex quadratic programming problems (NQP) appear in many practical applications, e.g., resource allocation [77], portfolio selection [95] and the maximal clique problem [52]. However, problem (3.69) is not easy to solve since it involves projections onto the simplex. Furthermore, when $\mathbf{g}$ is $\mathbf{0}$, the problem is NP-hard (see [98]).

In the initialization of the experiment, we first generate the matrix G. We randomly generate entries of a matrix $D \in \mathbb{R}^{n\times n}$ from i.i.d. standard Gaussian distribution and then let $G = D + D^\mathsf{T}$ so that G is symmetric. The vector $\mathbf{g}$ is also generated with i.i.d. standard Gaussian entries. The constant c is selected as $\max\{1, 10 * t\}$, where t is



(a) Function value gap against iterations

(b) Minimizer gap against iterations

Figure 3.2. Comparison of UPG, PG, PGE and FISTA for the NQP problem

a random scalar uniformly generated on $[0, 1]$. For the parameters required in PG, PGE and FISTA, we set $L = \max\{\sigma_{\max}(G), |\sigma_{\min}(G)|\}$ where $\sigma_{\max}(G)$ and $\sigma_{\min}(G)$ denote the largest eigenvalue and the smallest eigenvalue of the matrix G, respectively. We choose $l = |\sigma_{\min}(G)|$. The selection of L is used for UPG as well. Additionally, we let $n = 2000$ and terminate the algorithm when the maximal iteration reaches 5000 or the condition in (3.68) is satisfied.

We plot $|F_t - F_{\min}|$ against iterations and $\|\mathbf{x}_t - \mathbf{x}^*\|$ versus iterations in Figure 3.2, where $F_{\min}$ is the minimum function value of each algorithm and $\mathbf{x}^*$ is the corresponding solution of the $F_{\min}$ for every algorithm. We can see from Figure 3.2 UPG outperforms all other methods in terms of the convergence of the sequences $\{F_t\}$ and $\{\mathbf{x}_t\}$. Specifically, UPG uses much less iterations to achieve the required accuracy than PG, PGE and FISTA.

3.6. Stochastic Unified Proximal Gradient Method

In this section, we extend the proposed unified gradient method in Algorithm 3.1 in Section 3.2 to solve stochastic optimization problems. More specifically, the stochastic unified proximal gradient method (SUPG) inherits the advantages of UPG that it solves both convex and nonconvex composite problems by unified treatment. The main difference between UPG and SUPG in terms of the algorithmic form is that UPG exploits deterministic gradients while SUPG uses stochastic gradients and a variance reduction technique is applied. The variance reduction technique in SUPG is to reduce the variance effect of the stochastic gradient, that may slow down the convergence of the method [104].

3.6.1. Stochastic Composite Optimization Problem

Let us consider the following composite optimization problem

$$\min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x}) := f(\mathbf{x}) + p(\mathbf{x}), \tag{3.70}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is a closed convex set, f is the average of the Lipschitz continuously differentiable functions f_1, …, f_n on an open set containing $\mathcal{X}$, i.e., $f = \frac{1}{n}\sum_{i=1}^{n} f_i$, but possibly nonconvex and $p : \mathcal{X} \to \mathbb{R}$ is a proper closed convex, but possibly nonsmooth, function.

Problems in the form of (3.70) arise in many fields, such as machine learning, statistics and operation research. This kind of problems are also known as regularized empirical minimization problems (e.g., [104, 1, 121]). For example, given n samples in a dataset, we can fit a general ridge regression model if each f_i is the divergence of a linear combination of features and the desired dependent variable, and p is an l_2 regularization. When the penalty term is l_1 norm, then the problem becomes LASSO. In the case f_i is a logistic loss function and p is l_1 penalty, we can obtain regularized logistic regression.

```
Initialization: Given iteration numbers $K$, $m$, $\mathbf{x}_m^1 \in \mathcal{X}$, $\lambda \in [0,1]$ and $L > \mathcal{L}$;
                Set $\breve{\mathbf{x}}_m^1 = \mathbf{x}_m^1$.
For $k = 1, 2, 3, \ldots, K$
1.  $\nabla f(\mathbf{x}_m^k) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_m^k)$.
    For $t = 1, 2, \ldots, m$
2.      Set $\overline{\beta}_t = 2/(t+1)$.
3.      Randomly choose subset $I_t \subset \{1, 2, \ldots, n\}$ of size $b$, such that the
        probability of each index being selected is $\frac{b}{n}$.
4.      Choose $\mu_t \in [0, \mathcal{L}]$ and $\widehat{\mathbf{x}}_t^k = \beta_t \breve{\mathbf{x}}_t^k + (1-\beta_t)\mathbf{x}_t^k$ satisfy
            $f_{I_t}(\mathbf{x}_t^k) - f_{I_t}(\widehat{\mathbf{x}}_t^k) - \langle \nabla f_{I_t}(\widehat{\mathbf{x}}_t^k), \mathbf{x}_t^k - \widehat{\mathbf{x}}_t^k \rangle \geq -\frac{\mu_t}{2}\|\mathbf{x}_t^k - \widehat{\mathbf{x}}_t^k\|^2$, where
        $\beta_t = \max\{\overline{\beta}_t, \tau_t\}$ with $\tau_t = \lambda \underline{\tau}_t + (1-\lambda)\overline{\tau}_t$, $\underline{\tau}_t = \frac{1}{2}\left(1 - \sqrt{\frac{L-\mu_t}{L+\mu_t}}\right)$, $\overline{\tau}_t = \frac{\mu_t}{L+\mu_t}$.
5.      Set $\gamma_t = \beta_t \eta_t$, where $\eta_t = 2L/(2-\beta_t)$.
6.      $\mathbf{v}_t^k = \nabla f_{I_t}(\widehat{\mathbf{x}}_t^k) - \nabla f_{I_t}(\mathbf{x}_m^k) + \nabla f(\mathbf{x}_m^k)$.
7.      $\breve{\mathbf{x}}_{t+1}^k = \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{v}_t^k, \mathbf{x} \rangle + \frac{\gamma_t}{2} \left\| \mathbf{x} - \breve{\mathbf{x}}_t^k \right\|^2 + p(\mathbf{x}) \right\}$.
8.      If $\beta_t = 1$, let $\mathbf{x}_{t+1}^k = \breve{\mathbf{x}}_{t+1}^k$;
9.      Else    $\mathbf{x}_{t+1}^k = \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{v}_t^k, \mathbf{x} \rangle + \frac{\eta_t}{2} \left\| \mathbf{x} - \widehat{\mathbf{x}}_t^k \right\|^2 + p(\mathbf{x}) \right\}$.
    end
10. $\breve{\mathbf{x}}_1^{k+1} = \breve{\mathbf{x}}_m^k$ and $\mathbf{x}_1^{k+1} = \mathbf{x}_m^k$.
end
```

Algorithm 3.2. A stochastic unified proximal algorithm for nonconvex composite optimization

3.6.2. Algorithm Description

We propose Algorithm 3.2 to solve problem (3.70). Algorithm 3.2 is a stochastic mini-batch accelerated gradient method with variance reduction technique for solving composite optimization where f is not necessarily convex. At each stage, the algorithm performs m iterations with directions of mini-batch gradient

$$\mathbf{v}_t = \nabla f_{I_t}(\widehat{\mathbf{x}}_t) - \nabla f_{I_t}(\mathbf{x}_m) + \nabla f(\mathbf{x}_m)$$

where I_t is a randomly picked subset of size b from $\{1, 2, \ldots, n\}$ and $\nabla f_{I_t}(\mathbf{x}) = \frac{1}{b}\sum_{i=1}^{b} \nabla f_i(\mathbf{x})$ for $i \in I_t$.

There are two loops in Algorithm 3.2. At every iteration of the outer loop, we evaluate the full gradient of f at the most recent point $\mathbf{x}_m$ obtained from the inner loop. This full gradient is used to calculate the stochastic gradient $\mathbf{v}_t$ in Step 6. The steps in the inner loop are almost the same as that in UPG, except the choice of μ_t and the gradients in the subproblems. In Algorithm 3.2, we assume μ_t is given such that the conditions in Step 4 are satisfied. The details of how we choose μ_t will be discussed in Section 3.6.3 when presenting the numerical experiments. Furthermore, we compute the stochastic gradient $\mathbf{v}_t$ in Step 6, which is equipped with the variance reduction technique. The variance bound is given in Lemma 3.6.1. In addition, the stochastic gradient $\mathbf{v}_t$ is used rather than the gradient $\nabla f(\widehat{\mathbf{x}}_t)$, since it is easy to see $\mathbb{E}[\mathbf{v}_t] = \nabla f(\widehat{\mathbf{x}}_t)$.

We give the bound of the variance in the following lemma. Lemma 3.6.1 shows the mini-batch size b can help reduce the variance. We can observe if $b = n$, the bound given in (3.71) becomes 0, then the algorithm reduces to one for the deterministic case.

Lemma 3.6.1. *Suppose the gradient of f is Lipschitz continuous with constant $\mathcal{L}$. For the iterates generated by Algorithm 3.2, we have*

$$\mathbb{E}_{I_t}\left[\|\mathbf{v}_t - \nabla f(\widehat{\mathbf{x}}_t)\|^2\right] \;\; \leq \;\; \frac{\mathcal{L}^2(n-b)}{b(n-1)}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2. \tag{3.71}$$

Proof. By the definition of $\mathbf{v}_t$, we have

$$\begin{aligned}
& \mathbb{E}_{I_t}\left[\|\mathbf{v}_t - \nabla f(\widehat{\mathbf{x}}_t)\|^2\right] \\
= \; & \mathbb{E}_{I_t}\left[\|\nabla f_{I_t}(\widehat{\mathbf{x}}_t) - \nabla f_{I_t}(\mathbf{x}_t) + \nabla f(\mathbf{x}_t) - \nabla f(\widehat{\mathbf{x}}_t)\|^2\right] \\
= \; & \mathbb{E}_{I_t}\left[\|\nabla f_{I_t}(\widehat{\mathbf{x}}_t) - \nabla f_{I_t}(\mathbf{x}_t) - (\nabla f(\widehat{\mathbf{x}}_t) - \nabla f(\mathbf{x}_t))\|^2\right] \\
= \; & \frac{1}{b^2}\mathbb{E}_{I_t}\left[\left\|\sum_{i\in I_t}(\nabla f_i(\mathbf{x}_t) - \nabla f_i(\widehat{\mathbf{x}}_t) - (\nabla f(\mathbf{x}_t) - \nabla f(\widehat{\mathbf{x}}_t)))\right\|^2\right] \\
= \; & \frac{n-b}{b(n-1)}\mathbb{E}\left[\|\nabla f_i(\mathbf{x}_t) - \nabla f_i(\widehat{\mathbf{x}}_t) - (\nabla f(\mathbf{x}_t) - \nabla f(\widehat{\mathbf{x}}_t))\|^2\right] \\
\leq \; & \frac{n-b}{b(n-1)}\mathbb{E}\left[\|\nabla f_i(\mathbf{x}_t) - \nabla f_i(\widehat{\mathbf{x}}_t)\|^2\right] \\
\leq \; & \frac{\mathcal{L}^2(n-b)}{b(n-1)}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2, \qquad (3.72)
\end{aligned}$$

where the fourth equality follows from Lemma 4 in [80], the first inequality comes from $\mathbb{E}\|\xi - \mathbb{E}\xi\|^2 \leq \mathbb{E}\|\xi\|^2$, and the last inequality is obtained from the Lipschitz continuity of ∇f_i. □

3.6.3. Numerical Experiments

Though the theoretical results are still a work in progress, we would like to show the potential of the proposed stochastic proximal gradient method in practical applications. In this subsection, we examine the empirical performance of SUPG on the regularized empirical risk minimization (ERM) on regression and classification.

Given a dataset of n samples, $\{(\mathbf{a}_1, b_1), (\mathbf{a}_2, b_2), \cdots, (\mathbf{a}_n, b_n)\}$, where $\mathbf{a}_i \in \mathbb{R}^d$ for $i = 1, \cdots, n$, and b is a constant which is the value of dependent variable for regression problems or label for classification. We consider the following composite optimization problem in [136]

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) + \lambda_1\|\mathbf{x}\|_1 + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2, \qquad (3.73)$$

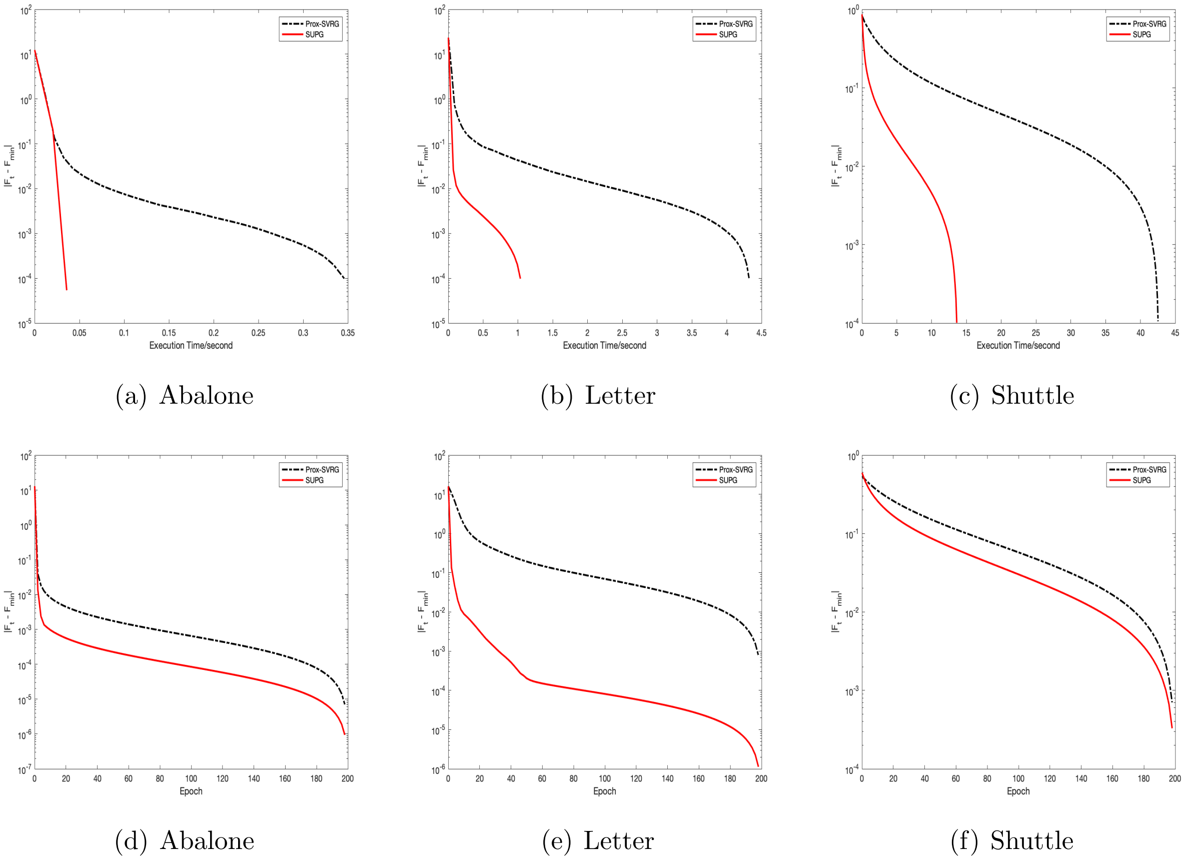
where $\lambda_1, \lambda_2 > 0$ are penalty parameters. Different models in machine learning can be obtained by choosing f_i to be various loss functions. Here, we consider f_i being sigmoid loss $\frac{1}{1+\exp(b_i \mathbf{a}_i^\mathsf{T} \mathbf{x})}$ and logistic loss $\log(1 + \exp(-b_i \mathbf{a}_i^\mathsf{T} \mathbf{x}))$.

In this experiment, we compare SUPG with proximal stochastic variance-reduced gradient (prox-SVRG) for nonconvex composite objectives in Algorithm 1 in [136] on three publicly available datasets: abalone from [37], letter and shuttle in [21]. Abalone dataset is used to find a regression model to predict the age of abalone from physical measurements while datasets shuttle and letter are for classification problems that classify classes of shuttles by attributes and identify black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet, respectively.

On the other hand, the initialization of $\mathbf{x} \in \mathbb{R}^d$ for all the algorithms is randomly selected from standard uniform distribution in $(0, 1)$ and the maximum iteration number for inner loop is $m = 150$. The termination iteration for the outer loop K is set to be 100 for plotting with respect to epoch and the stopping condition of the outer loop for figures with respect to time is

$$\frac{|F_t - F_{\min}^c|}{|F_{\min}^c + 1|} \leq 10^{-4},$$

where $F_{\min}^c$ denotes the minimum of current sequence $\{F_t\}$ at the t-th iteration. Additionally, the penalty parameters λ_1 and λ_2 are set to 10^{-5}. Note that they are usually selected by cross-validation in practice. Moreover, as we mentioned in previous subsection, instead of estimating μ_t in the way of Step 2 in Algorithm 3.1, here we monotonically increase μ_t by $0.01L$ in the range $[0, L]$ until the requirement in Step 4 is satisfied. Then, we have the experimental results in the following.

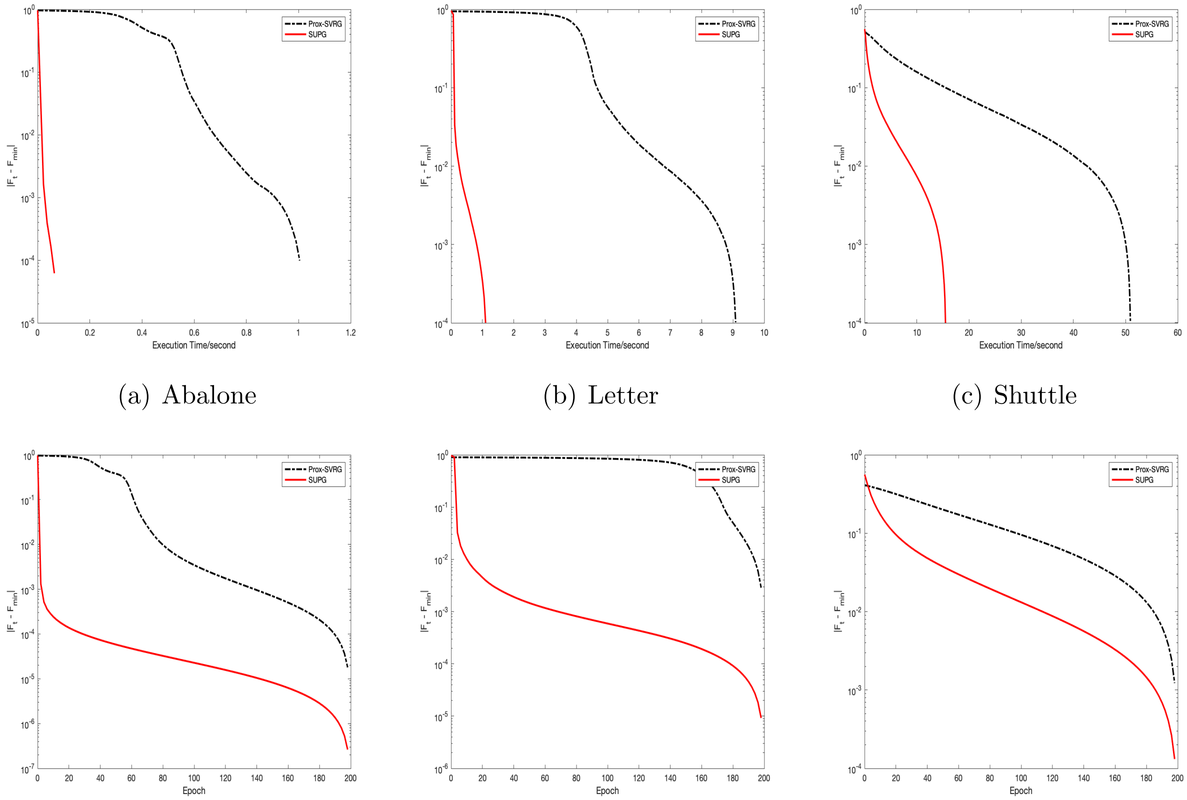


(a) Abalone (b) Letter (c) Shuttle

(d) Abalone (e) Letter (f) Shuttle

Figure 3.3. Comparison on execution time and epoch of SUPG and prox-SVRG for logistic loss

Figure 3.3 and Figure 3.4 show the results of applying SUPG and prox-SVRG on solving problem (3.70) when f_i are logistic loss and sigmoid loss. We plot $|F_t - F_{\min}|$ against execution time in (a), (b) and (c) and $|F_t - F_{\min}|$ versus epoch in (d), (e) and (f) for both Figure 3.3 and Figure 3.4, where $F_{\min}$ is the minimum of the sequence $\{F_t\}$ generated by each algorithm. We can observe from Figure 3.3 (a), (b) and (c) that both methods converge shortly to the preset accuracy. However, SUPG terminates much faster than prox-SVRG. In other words, prox-SVRG takes more than three times of the time that SUPG needs to reach a solution. On the other hand, we can see from Figure 3.3 (d), (e) and (f) that, when using the same amount of epochs, SUPG obtains an objective func-



(a) Abalone

(b) Letter

(c) Shuttle

(d) Abalone

(e) Letter

(f) Shuttle

Figure 3.4. Comparison on execution time and epoch of SUPG and prox-SVRG for sigmoid loss

tion value with smaller function value gap than that of prox-SVRG. Overall, SUPG outperforms prox-SVRG from the aspects of less running time and higher accuracy with same epochs. Similar conclusions can be obtained from Figure 3.4. SUPG uses much less time to reach a desired accuracy than prox-SVRG and SUPG can find an **x** such that the gap between current function value and the minimum of the generated sequence of objective function values is smaller than that of prox-SVRG.

Chapter 4. An Inexact ADMM for Separable Nonconvex and Nonsmooth Optimization

In this chapter, we present an algorithm that is a variant of the alternating direction proximal method of multiplier and is equipped with linesearch. The method solves a family of separable minimization optimization problems with linear equality constraints, where the objective function is the sum of a smooth but possibly nonconvex function and a possibly nonsmooth nonconvex function. Since most of the computational effort is spent on finding the minimizers to the subproblems in such methods, inexact solutions are used in the proposed method. An expansion step is applied to $\mathbf{x}$-iterates to further improve the performance of the algorithm.

4.1. Separable Nonconvex and Nonsmooth Optimization Problems

Throughout the chapter, we consider the following separable nonconvex and nonsmooth linearly constrained optimization problem

$$\begin{aligned} \min_{(\mathbf{x},\mathbf{y})\in\mathbb{R}^{n_x}\times\mathbb{R}^{n_y}} \; & F(\mathbf{x},\mathbf{y}) := f(\mathbf{x}) + g(\mathbf{y}) \\ \text{s.t.} \quad & A\mathbf{x} + B\mathbf{y} = \mathbf{b}, \end{aligned} \tag{4.1}$$

where $f : \mathbb{R}^{n_x} \to \mathbb{R}$ is Lipschitz continuously differentiable, but possibly nonconvex, $g : \mathbb{R}^{n_y} \to \mathbb{R}$ is a proper, lower semicontinuous, possibly nonconvex and nonsmooth function and $A \in \mathbb{R}^{m\times n_x}$, $B \in \mathbb{R}^{m\times n_y}$ and $\mathbf{b} \in \mathbb{R}^m$ are given. Note that constraints of the form $\mathbf{y} \in \mathcal{Y}$ for a closed set $\mathcal{Y} \subset \mathbb{R}^{n_y}$ can be incorporated in the objective by using $g(\mathbf{y})$ as an indicator function of $\mathcal{Y}$.

As we have mentioned in Section 2.3 in Chapter 2, ADMMs have obtained great success in both theory and numerical efficiency for solving linearly constrained separable convex optimization. The original ADMM [46, 53] and its variants for solving convex problems have been further extended to solve the nonconvex structured optimization problem (4.1). With proper choice of β, the excellent performance of ADMM on nonconvex cases has been observed in recent applications [122]. Note that the dominant computation in each iteration of ADMM is to solve its subproblems. Hence, how to solve these subproblems inexactly while still maintaining nice convergence properties will be critical for the overall success of ADMM, especially when no closed form solution of the subproblem exists [65, 66, 132].

Motivated by the recent surged interests of the applications of ADMM on nonconvex cases and the adaptive relative error strategy used in ALM and convex ADMM (e.g., [65]), we propose an inexact ADMM (I-ADMM) framework with an expansion linesearch step (see Algorithm 4.1) to solve the nonconvex problem (4.1). Our contribution mainly lies in the following aspects.

First, the proposed I-ADMM solves the subproblems inexactly to adaptive accuracy while global convergence and a linear convergence rate are guaranteed under proper conditions. Solving subproblems in ADMMs inexactly has been widely used for convex optimization problems. One common way is to solve the subproblems to the accuracy based on some absolute summable error criteria, but the guidance on how to adaptively select the error tolerance is absent except requiring it to be summable. Moreover, ADMM is a splitting version of the augmented Lagrangian method, for which nice theoretical convergence results and numerical experiments have been obtained [39, 114] using adaptive relative subproblem stopping criteria. Hence, ideally we should also be able to solve the subproblems of I-ADMM to an adaptive accuracy while maintaining desirable convergence properties. Here we establish global convergence and a linear convergence rate of I-ADMM under a local error bound condition and a weakly convex property of g.

Second, the proposed I-ADMM allows a more flexible range of the stepsize, $s \in (0,2)$, in the update step of dual variable and applies an expansion linesearch step to accelerate the convergence. The commonly known dual stepsize s of ADMM for solving convex optimization can be arbitrarily chosen from the interval $(0, (\sqrt{5}+1)/2)$ (see [44, 4]). However, only the unit stepsize was discussed in almost all current ADMMs for nonconvex problems [5, 75, 86, 122], except the methods in [133, 135] that allow $s \in (0, (\sqrt{5}+1)/2)$ for an image recovery problem as the original ADMM and $s \in (0,2)$ for a linearized ADMM. On the other hand, both methods find exact solutions of the subproblems or the linearized subproblems. Therefore, by constructing different potential energy functions, we extend the dual stepsize interval to $(0,2)$ even with inexact subproblem solutions. In addition, an expansion linesearch step (see step 6 of Algorithm 4.1) is applied in the proposed I-ADMM, which improves the numerical performance and reduces the sensitivity of algorithm parameters as well.

Third, we propose a generalized accelerated gradient method (G-AGM) with momentum acceleration to solve the nonconvex smooth $\mathbf{x}$-subproblem. Our G-AGM method is motivated by the extrapolation technique for solving both convex and nonconvex optimization [9, 123]. G-AGM is particularly designed for solving $\mathbf{x}$-subproblem arising in our I-ADMM. It can be viewed as a special case of Algorithm 3.1 in Chapter 3. This method guarantees global convergence for solving the smooth possibly nonconvex subproblem and will automatically reduce to an optimal gradient method when the function f in the objective is convex.

Additionally, the framework of I-ADMM is more general and flexible than most of existing ADMMs. When no expansion step (Step 6 of Algorithm 4.1) is used, this I-ADMM will just reduce to a particular inexact version of nonconvex ADMM. But our linesearch expansion step often allows a much larger stepsize than the fixed relaxation stepsize used in [38, 68, 72]. We also have more general problem settings and different assumptions for establishing global convergence and the linear convergence rate than that in [14, 86] which require $B = \mathbf{I}$, $\mathbf{b} = \mathbf{0}$ and the Kurdyka-Łojasiewicz property. Although the over-relaxation step was adopted in [57], the involved subproblems were also solved exactly. Moreover, our numerical experiments show that our proposed I-ADMM is very effective compared with other state-of-the-art ADMM algorithms in the literature and can obtain more accurate solution.

4.2. Algorithm Description

We propose an inexact ADMM (I-ADMM, i.e., Algorithm 4.1) with an expansion linesearch step to solve the possibly nonsmooth and nonconvex problem (4.1). At each iteration, both the $\mathbf{y}$-subproblem, i.e.,

$$\min_{\mathbf{y}\in\mathbb{R}^{n_y}} \mathcal{L}_y^k(\mathbf{y}) := \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}, \boldsymbol{\lambda}^k) + \frac{\beta}{2}\|\mathbf{y} - \mathbf{y}^k\|^2_{\mathcal{D}_y^k}, \tag{4.2}$$

and the $\mathbf{x}$-subproblem, i.e.,

$$\min_{\mathbf{x}\in\mathbb{R}^{n_x}} \mathcal{L}_x^k(\mathbf{x}) := \mathcal{L}_\beta(\mathbf{x}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{D}_x^k}, \tag{4.3}$$

are allowed to be solved inexactly, where $\mathcal{D}_x^k \succeq \mathbf{0}$ and $\mathcal{D}_y^k \succeq \mathbf{0}$ can be two uniformly upper bounded positive semidefinite matrices that are adaptively chosen. More precisely, in Algorithm 4.1, it requires the $\mathbf{y}^{k+1}$ generated at the k-th iteration satisfies

$$\frac{\beta}{2}\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2_{\mathcal{D}_y} + \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) \leq \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k) \tag{4.4}$$

for some positive definite matrix $\mathcal{D}_y \succ \mathbf{0}$, and there exist a positive constant c_y and some $\xi_y^{k+1} \in \partial_y \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k)$ such that

$$\|\xi_y^{k+1}\| \leq c_y \beta \|\mathbf{y}^{k+1} - \mathbf{y}^k\|. \tag{4.5}$$

For the inexact solution of $\mathbf{x}$-subproblem, it requires the $\widehat{\mathbf{x}}^k$ generated at the k-th iteration of Algorithm 4.1 satisfies

$$\frac{\beta}{2}\|\widehat{\mathbf{x}}^k - \mathbf{x}^k\|_{\mathcal{D}_x}^2 + \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) \leq \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) \tag{4.6}$$

for some positive definite matrix $\mathcal{D}_x \succ \mathbf{0}$, and there exists a positive constant $c_x > 0$ such that $\xi_x^{k+1} = \nabla_x \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k)$ satisfies

$$\|\xi_x^{k+1}\| \leq c_x \beta \left(\|\widehat{\mathbf{x}}^k - \mathbf{x}^k\| + \|\mathbf{y}^{k+1} - \mathbf{y}^k\|\right). \tag{4.7}$$

The algorithm stops when R^{k+1} is sufficiently small, where

$$R^{k+1} = \|\widehat{\mathbf{x}}^k - \mathbf{x}^k\| + \|\mathbf{y}^{k+1} - \mathbf{y}^k\| + \|\widehat{\mathbf{r}}^{k+1}\|, \tag{4.8}$$

and $\widehat{\mathbf{r}}^{k+1} = A\widehat{\mathbf{x}}^k + B\mathbf{y}^{k+1} - \mathbf{b}$.

Furthermore, we see that an expansion linesearch step for $\mathbf{x}$-iterates is applied in Step 6 of Algorithm 4.1. From this expansion step, we have $\phi(\alpha_k) = \mathcal{L}_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1})$, $\phi(1) = \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1})$ and the stepsize $\alpha_k \geq 1$ is chosen such that $\bar{\xi}_x^{k+1} = \nabla_x \mathcal{L}_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k)$ satisfies

$$\|\bar{\xi}_x^{k+1}\| \leq c_x \beta \left(\|\widehat{\mathbf{x}}^k - \mathbf{x}^k\| + \|\mathbf{y}^{k+1} - \mathbf{y}^k\|\right), \tag{4.9}$$

and

$$\mathcal{L}_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1}) \leq \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1}) - \delta\beta\|\mathbf{x}^{k+1} - \widehat{\mathbf{x}}^k\|^2, \tag{4.10}$$

```
Initialization:  parameters $\beta > 0$, $s \in (0,2)$, $\delta \in (0,1)$ and $\eta > 1$,
                 starting point $\mathbf{w}^0 = (\mathbf{x}^0, \mathbf{y}^0, \boldsymbol{\lambda}^0)$;
For $k = 0, 1, 2, \ldots$
1.  Choose uniformly upper bounded matrices $\mathcal{D}_y^k \succeq \mathbf{0}$ and $\mathcal{D}_x^k \succeq \mathbf{0}$.
2.  Solve $\mathbf{y}^{k+1} \approx \arg\min_{\mathbf{y} \in \mathbb{R}^{n_y}} \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}, \boldsymbol{\lambda}^k) + \frac{\beta}{2}\|\mathbf{y} - \mathbf{y}^k\|_{\mathcal{D}_y^k}^2$ inexactly such
    that (4.4) and (4.5) are satisfied.
3.  Solve $\widehat{\mathbf{x}}^k \approx \arg\min_{\mathbf{x} \in \mathbb{R}^{n_x}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^k\|_{\mathcal{D}_x^k}^2$ inexactly such
    that (4.6) and (4.7) are satisfied.
4.  If $R^{k+1}$ defined in (4.8) is sufficiently small, stop.
5.  Update the Lagrange Multiplier:
       $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k - s\beta(A\widehat{\mathbf{x}}^k + B\mathbf{y}^{k+1} - \mathbf{b})$.
6.  Expansion step for the x-iterate:
       $\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \widehat{\mathbf{d}}_x^k$, where $\widehat{\mathbf{d}}^k = \widehat{\mathbf{x}}^k - \mathbf{x}^k$ and $\alpha_k = \eta^j$ with $j \geq 0$
    being the largest integer such that
       $\phi(\alpha_k) \leq \phi(1) - \delta\beta\|\mathbf{x}^{k+1} - \widehat{\mathbf{x}}^k\|^2$ and (4.9) are satisfied,
    where $\phi(\alpha) = \mathcal{L}_\beta(\mathbf{x}^k + \alpha\widehat{\mathbf{d}}_x^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1})$.
end
```

Algorithm 4.1. An inexact ADMM (I-ADMM) for separable nonconvex optimization

where $\delta \in (0,1)$ is a preset parameter. As standard linesearch techniques in optimization, this Armijo-type linesearch step can significantly improve the performance of the algorithm and reduce the sensitivity of the choices of the parameters in the algorithm.

We now have the following comments regarding the conditions (4.4), (4.5), (4.6) and (4.7) for the subproblem solutions. First, since $\{\mathcal{D}_x^k\}$ and $\{\mathcal{D}_y^k\}$ are chosen uniformly upper bounded, supposing functions $\mathcal{L}_x^k(\cdot)$ and $\mathcal{L}_y^k(\cdot)$ are bounded from below, we can find $\mathbf{y}^{k+1}$ and $\widehat{\mathbf{x}}^k$ such that conditions (4.5) and (4.7) are satisfied. In addition, if $R^{k+1} = 0$, we can derive $\mathbf{w}^k := (\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k)$ is a stationary point of problem (4.1) (see definition (4.35)). On the other hand, if $\{\mathcal{D}_x^k\}$ and $\{\mathcal{D}_y^k\}$ are chosen such that

$$\|\widehat{\mathbf{x}}^k - \mathbf{x}^k\|_{\mathcal{D}_x^k}^2 \geq \eta_x \|\widehat{\mathbf{x}}^k - \mathbf{x}^k\|^2 \quad \text{and} \quad \|\mathbf{y}^{k+1} - \mathbf{y}^k\|_{\mathcal{D}_y^k}^2 \geq \eta_y \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2 \tag{4.11}$$

for some constants $\eta_x > 0$ and $\eta_y > 0$, then for any $\widehat{\mathbf{x}}^k$ satisfying $\mathcal{L}_x^k(\widehat{\mathbf{x}}^k) \leq \mathcal{L}_x^k(\mathbf{x}^k)$ and any

$\mathbf{y}^{k+1}$ satisfying $\mathcal{L}_y^k(\mathbf{y}^{k+1}) \leq \mathcal{L}_y^k(\mathbf{y}^k)$, the conditions (4.4) and (4.6) will hold with $\mathcal{D}_x = \eta_x \mathbf{I}$ and $\mathcal{D}_y = \eta_y \mathbf{I}$.

Therefore, obviously one simple choice of matrices $\mathcal{D}_x^k$ and $\mathcal{D}_y^k$ can be $\mathcal{D}_x^k = \eta_x \mathbf{I}$ and $\mathcal{D}_y^k = \eta_y \mathbf{I}$ for all $k \geq 0$. However, under certain circumstances, it is not even necessary to require positive definiteness of $\{\mathcal{D}_x^k\}$ or $\{\mathcal{D}_y^k\}$ in order to satisfy the conditions (4.4) and (4.6). For instance, denoting $L > 0$ as the Lipschitz constant of ∇f, if $A^\top A + \mathcal{D}_x^k \succ \mathbf{0}$ and the parameter β is sufficiently large such that $\beta(A^\top A + \mathcal{D}_x^k) \succeq (L + 2\eta\beta)\mathbf{I}$ for some $\eta > 0$, the objective function $\mathcal{L}_x^k(\cdot)$ of the $\mathbf{x}$-subproblem (4.3) will be uniformly strongly convex with modulus greater than $2\eta\beta > 0$. In this case, all points sufficiently close to the minimizer of the $\mathbf{x}$-subproblem (4.3) will satisfy (4.6) with $\mathcal{D}_x = \eta \mathbf{I}$. Hence, in the rest of the chapter, we assume that we can solve the subproblems (4.2) and (4.3) inexactly to meet conditions (4.4), (4.5), (4.6) and (4.7).

4.3. Convergence Analysis

In this section, we study the convergence properties of Algorithm 4.1. For the convergence analysis, we need the following assumptions throughout the chapter.

Assumption 4.3.1. *The gradient of f is Lipschitz continuous, i.e., there exists a constant $L > 0$ such that*

$$\|\nabla f(\mathbf{z}_1) - \nabla f(\mathbf{z}_2)\| \leq L\|\mathbf{z}_1 - \mathbf{z}_2\| \tag{4.12}$$

for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{n_x}$.

Assumption 4.3.2. $(\mathit{Range}(B) \cup \mathbf{b}) \subseteq \mathit{Range}(A)$.

Based on Assumption 4.3.2, we have $\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k = -s\beta\widehat{\mathbf{r}}^{k+1} \in \mathrm{Range}(A)$, which implies

$$\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\| \leq \sigma_A^{-\frac{1}{2}} \|A^\mathsf{T}(\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k)\|, \tag{4.13}$$

where σ_A is the smallest positive eigenvalue of $A^\mathsf{T}A$ (or equivalently the smallest positive eigenvalue of AA^T). Certainly, Assumption 4.3.2 holds if A is nonsingular or has full column or full row rank.

4.3.1. Preliminary Lemmas

For the convenience of analysis, let us denote

$$\widehat{\mathbf{d}}_x^k = \widehat{\mathbf{x}}^k - \mathbf{x}^k, \quad \mathbf{d}_y^k = \mathbf{y}^{k+1} - \mathbf{y}^k \quad \text{and} \quad \mathbf{d}_\lambda^k = \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k$$

and define

$$\psi_1(s) = \max\left\{1, \frac{s^2}{(2-s)^2}\right\} \quad \text{and} \quad \psi_2(s) = \max\left\{\frac{1-s}{s}, \frac{s-1}{2-s}\right\}. \tag{4.14}$$

It is easy to see that $\psi_1(s) > 0$ and $\psi_2(s) \geq 0$ for any $s \in (0,2)$. Then, we have the following lemma.

Lemma 4.3.1. *Suppose the Assumption 4.3.1 holds and the iterates $\{\mathbf{w}^k\}$ generated by Algorithm 4.1 satisfy the condition (4.7). Then, we have*

$$\begin{aligned}\|A^\mathsf{T}\mathbf{d}_\lambda^k\|^2 \quad \leq \quad & \psi_2(s)\left(\|A^\mathsf{T}\mathbf{d}_\lambda^{k-1}\|^2 - \|A^\mathsf{T}\mathbf{d}_\lambda^k\|^2\right) + 2\psi_1(s)(L + c_x\beta)^2\|\widehat{\mathbf{d}}_x^k\|^2 \\ & +6\psi_1(s)c_x^2\beta^2\left(\|\widehat{\mathbf{d}}_x^{k-1}\|^2 + \|\mathbf{d}_y^k\|^2 + \|\mathbf{d}_y^{k-1}\|^2\right).\end{aligned} \tag{4.15}$$

Proof. By the definition of $\xi_x^{k+1} = \nabla_x \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k)$, we have

$$\xi_x^{k+1} = \nabla f(\widehat{\mathbf{x}}^k) + A^\mathsf{T}\left[-\boldsymbol{\lambda}^k + \beta\widehat{\mathbf{r}}^{k+1}\right],$$

where $\widehat{\mathbf{r}}^{k+1} = A\widehat{\mathbf{x}}^k + B\mathbf{y}^{k+1} - \mathbf{b}$. Hence, we have

$$A^\mathsf{T}\boldsymbol{\lambda}^k = \nabla f(\widehat{\mathbf{x}}^k) - \xi_x^{k+1} + \beta A^\mathsf{T}\widehat{\mathbf{r}}^{k+1},$$

which follows from $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k - s\beta\widehat{\mathbf{r}}^{k+1}$ that

$$sA^\mathsf{T}\boldsymbol{\lambda}^k = s\left(\nabla f(\widehat{\mathbf{x}}^k) - \xi_x^{k+1}\right) + A^\mathsf{T}\left(\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1}\right).$$

So, we have

$$A^\mathsf{T}\boldsymbol{\lambda}^{k+1} = s\left(\nabla f(\widehat{\mathbf{x}}^k) - \xi_x^{k+1}\right) + (1-s)A^\mathsf{T}\boldsymbol{\lambda}^k. \tag{4.16}$$

Similarly, by the definitions of $\bar{\xi}_x^{k+1}, \widehat{\mathbf{r}}^{k+1}$ and $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k - s\beta\widehat{\mathbf{r}}^{k+1}$, we have

$$A^\mathsf{T}\boldsymbol{\lambda}^{k+1} = s\left(\nabla f(\mathbf{x}^{k+1}) - \bar{\xi}_x^{k+1}\right) + (1-s)A^\mathsf{T}\boldsymbol{\lambda}^k. \tag{4.17}$$

Then, by replacing $k+1$ by k in (4.17) and subtracting it from (4.16) and $\mathbf{d}_\lambda^k = \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k$, it gives

$$A^\mathsf{T}\mathbf{d}_\lambda^k = s\boldsymbol{\delta}^k + (1-s)A^\mathsf{T}\mathbf{d}_\lambda^{k-1}, \tag{4.18}$$

where

$$\boldsymbol{\delta}^k = \nabla f(\widehat{\mathbf{x}}^k) - \nabla f(\mathbf{x}^k) - \xi_x^{k+1} + \bar{\xi}_x^k. \tag{4.19}$$

In the following we consider two cases, $s \in (0,1]$ and $s \in (1,2)$.

Case 1: $s \in (0,1]$. Then, it follows from (4.18) and the convexity of $\|\cdot\|^2$ that

$$\|A^\mathsf{T}\mathbf{d}_\lambda^k\|^2 \le s\|\boldsymbol{\delta}^k\|^2 + (1-s)\|A^\mathsf{T}\mathbf{d}_\lambda^{k-1}\|^2.$$

By subtracting $(1-s)\|A^\mathsf{T}\mathbf{d}_\lambda^k\|^2$ and dividing s from both sides of the above inequality, we derive

$$\|A^\mathsf{T}\mathbf{d}_\lambda^k\|^2 \le \|\boldsymbol{\delta}^k\|^2 + \frac{1-s}{s}\left(\|A^\mathsf{T}\mathbf{d}_\lambda^{k-1}\|^2 - \|A^\mathsf{T}\mathbf{d}_\lambda^k\|^2\right). \tag{4.20}$$

Case 2: $s \in (1,2)$. It follows from (4.18) that

$$\|A^\mathsf{T}\mathbf{d}_\lambda^k\|^2 = (1-s)^2\|A^\mathsf{T}\mathbf{d}_\lambda^{k-1}\|^2 + s^2\|\boldsymbol{\delta}^k\|^2 + 2s(1-s)\langle A^\mathsf{T}\mathbf{d}_\lambda^{k-1}, \boldsymbol{\delta}^k\rangle. \tag{4.21}$$

Then, by (4.21) and Cauchy-Schwartz inequality, for an $\nu > 0$ we have

$$\begin{aligned} \|A^{\mathsf{T}}\mathbf{d}_\lambda^k\|^2 &\leq (1-s)^2\|A^{\mathsf{T}}\mathbf{d}_\lambda^{k-1}\|^2 + s^2\|\boldsymbol{\delta}^k\|^2 + s(s-1)\left(\nu\|A^{\mathsf{T}}\mathbf{d}_\lambda^{k-1}\|^2 + \frac{1}{\nu}\|\boldsymbol{\delta}^k\|^2\right) \\ &= \big((1-s)^2 + s(s-1)\nu\big)\|A^{\mathsf{T}}\mathbf{d}_\lambda^{k-1}\|^2 + \left(s^2 + \frac{s(s-1)}{\nu}\right)\|\boldsymbol{\delta}^k\|^2. \end{aligned} \tag{4.22}$$

By choosing $\nu = (2-s)/s$, we have

$$(1-s)^2 + s(s-1)\nu = s-1 \qquad \text{and} \qquad s^2 + \frac{s(s-1)}{\nu} = \frac{s^2}{2-s}.$$

So, we have from (4.22) that

$$\|A^{\mathsf{T}}\mathbf{d}_\lambda^k\|^2 \leq (s-1)\|A^{\mathsf{T}}\mathbf{d}_\lambda^{k-1}\|^2 + \frac{s^2}{2-s}\|\boldsymbol{\delta}^k\|^2.$$

By subtracting $(s-1)\|A^{\mathsf{T}}\mathbf{d}_\lambda^k\|^2$ and dividing $2-s$ from both sides of the above inequality, we derive

$$\|A^{\mathsf{T}}\mathbf{d}_\lambda^k\|^2 \leq \frac{s^2}{(2-s)^2}\|\boldsymbol{\delta}^k\|^2 + \frac{s-1}{2-s}\left(\|A^{\mathsf{T}}\mathbf{d}_\lambda^{k-1}\|^2 - \|A^{\mathsf{T}}\mathbf{d}_\lambda^k\|^2\right). \tag{4.23}$$

Now, combining (4.20) and (4.23) and the definition of functions ψ_1 and ψ_2 in (4.14), we have

$$\|A^{\mathsf{T}}\mathbf{d}_\lambda^k\|^2 \leq \psi_1(s)\|\boldsymbol{\delta}^k\|^2 + \psi_2(s)\left(\|A^{\mathsf{T}}\mathbf{d}_\lambda^{k-1}\|^2 - \|A^{\mathsf{T}}\mathbf{d}_\lambda^k\|^2\right). \tag{4.24}$$

In addition, by (4.7), (4.12) and the definition of $\boldsymbol{\delta}^k$ in (4.19), we have

$$\begin{aligned} \|\boldsymbol{\delta}^k\|^2 &= \|\nabla f(\widehat{\mathbf{x}}^k) - \nabla f(\mathbf{x}^k) - \xi_x^{k+1} + \overline{\xi}_x^k\|^2 \\ &\leq \left(L\|\widehat{\mathbf{d}}_x^k\| + c_x\beta(\|\widehat{\mathbf{d}}_x^k\| + \|\widehat{\mathbf{d}}_x^{k-1}\| + \|\mathbf{d}_y^k\| + \|\mathbf{d}_y^{k-1}\|)\right)^2 \\ &= \left((L+c_x\beta)\|\widehat{\mathbf{d}}_x^k\| + c_x\beta\left(\|\widehat{\mathbf{d}}_x^{k-1}\| + \|\mathbf{d}_y^k\| + \|\mathbf{d}_y^{k-1}\|\right)\right)^2 \\ &\leq 2(L+c_x\beta)^2\|\widehat{\mathbf{d}}_x^k\|^2 + 6c_x^2\beta^2\left(\|\widehat{\mathbf{d}}_x^{k-1}\|^2 + \|\mathbf{d}_y^k\|^2 + \|\mathbf{d}_y^{k-1}\|^2\right). \end{aligned} \tag{4.25}$$

Therefore, (4.15) follows from the above inequality and (4.24). □

Now, let us denote $\mathbf{w}^k = (\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k)$, $\widehat{\mathbf{w}}^k = (\widehat{\mathbf{x}}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}^k)$ and define the potential energy functions as

$$E^{k+1} = \mathcal{L}_\beta(\mathbf{w}^{k+1}) + \Gamma^k \qquad \text{and} \qquad \widehat{E}^{k+1} = \mathcal{L}_\beta(\widehat{\mathbf{w}}^{k+1}) + \Gamma^k, \tag{4.26}$$

where

$$\Gamma^k = \frac{6(1+\tau)\psi_1(s)c_x^2\beta}{s\sigma_A}\left(\|\widehat{\mathbf{d}}_x^k\|^2 + \|\mathbf{d}_y^k\|^2\right) + \frac{(1+\tau)\psi_2(s)}{s\beta\sigma_A}\|A^\mathsf{T}\mathbf{d}_\lambda^k\|^2 \tag{4.27}$$

and τ can be any constant in $(0,1)$. Then, based on the previous lemma, we can derive the following potential energy reduction theorem.

Theorem 4.3.1. *Suppose the Assumptions 4.3.1 and 4.3.2 hold and the iterates $\{\mathbf{w}^k\}$ generated by Algorithm 4.1 satisfy the conditions (4.4), (4.6) and (4.7). Let $\tau \in (0,1)$ be the constant in the potential energies E^k and $\widehat{E}^k$ defined in (4.26). If the parameters in Algorithm 4.1 are chosen such that*

$$\overline{D}_x := \frac{1-\tau}{2(1+\tau)}\mathcal{D}_x - \frac{\psi_1(s)\left[2(L/\beta + c_x)^2 + 6c_x^2\right]}{s\sigma_A}\mathbf{I} \succeq \mathbf{0}, \tag{4.28}$$

and

$$\overline{D}_y := \frac{1-\tau}{24(1+\tau)}\mathcal{D}_y - \frac{\psi_1(s)c_x^2}{s\sigma_A}\mathbf{I} \succeq \mathbf{0}. \tag{4.29}$$

Then, denoting $\widetilde{\mathbf{d}}_x^k = \mathbf{x}^k - \widehat{\mathbf{x}}^{k-1}$, we have

$$E^{k+1} \le E^k - \frac{\tau\beta}{2}\|\widehat{\mathbf{d}}_x^k\|_{\mathcal{D}_x}^2 - \frac{\tau\beta}{2}\|\mathbf{d}_y^k\|_{\mathcal{D}_y}^2 - \frac{\tau}{s\beta}\|\mathbf{d}_\lambda^k\|^2 - \delta\beta\|\widetilde{\mathbf{d}}_x^{k+1}\|^2 \tag{4.30}$$

and

$$\widehat{E}^{k+1} \le \widehat{E}^k - \frac{\tau\beta}{2}\|\widehat{\mathbf{d}}_x^k\|_{\mathcal{D}_x}^2 - \frac{\tau\beta}{2}\|\mathbf{d}_y^k\|_{\mathcal{D}_y}^2 - \frac{\tau}{s\beta}\|\mathbf{d}_\lambda^k\|^2 - \delta\beta\|\widetilde{\mathbf{d}}_x^k\|^2, \tag{4.31}$$

where $\delta \in (0,1)$ is the parameter given in Algorithm 4.1.

Proof. First, by (4.4), (4.6) and (4.13), we have

$$\begin{aligned}
& \mathcal{L}_\beta(\widehat{\mathbf{w}}^{k+1}) - \mathcal{L}_\beta(\mathbf{w}^k) \\
= \;& \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1}) - \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) + \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) \\
& - \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) + \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^k) - \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k) \\
\leq \;& \frac{1+\tau}{s\beta}\|\mathbf{d}_\lambda^k\|^2 - \frac{\beta}{2}\|\widehat{\mathbf{d}}_x^k\|^2_{\mathcal{D}_x} - \frac{\beta}{2}\|\mathbf{d}_y^k\|^2_{\mathcal{D}_y} - \frac{\tau}{s\beta}\|\mathbf{d}_\lambda^k\|^2 \\
\leq \;& \frac{1+\tau}{s\beta\sigma_A}\|A^\mathsf{T}\mathbf{d}_\lambda^k\|^2 - \frac{\beta}{2}\|\widehat{\mathbf{d}}_x^k\|^2_{\mathcal{D}_x} - \frac{\beta}{2}\|\mathbf{d}_y^k\|^2_{\mathcal{D}_y} - \frac{\tau}{s\beta}\|\mathbf{d}_\lambda^k\|^2. \qquad (4.32)
\end{aligned}$$

In addition, by (4.15), we obtain

$$\begin{aligned}
& \frac{1+\tau}{s\beta\sigma_A}\|A^\mathsf{T}\mathbf{d}_\lambda^k\|^2 \\
\leq \;& \frac{(1+\tau)\psi_1(s)}{s\beta\sigma_A}\left[2(L+c_x\beta)^2\|\widehat{\mathbf{d}}_x^k\|^2 + 6c_x^2\beta^2\left(\|\widehat{\mathbf{d}}_x^{k-1}\|^2 + \|\mathbf{d}_y^k\|^2 + \|\mathbf{d}_y^{k-1}\|^2\right)\right] \\
& + \frac{(1+\tau)\psi_2(s)}{s\beta\sigma_A}\left(\|A^\mathsf{T}\mathbf{d}_\lambda^{k-1}\|^2 - \|A^\mathsf{T}\mathbf{d}_\lambda^k\|^2\right). \qquad (4.33)
\end{aligned}$$

Then, plugging (4.33) into (4.32), by (4.10) and $\widetilde{\mathbf{d}}_x^{k+1} = \mathbf{x}^{k+1} - \widehat{\mathbf{x}}^k$, we have

$$\begin{aligned}
& \mathcal{L}_\beta(\mathbf{w}^{k+1}) - \mathcal{L}_\beta(\mathbf{w}^k) \\
\leq \;& \mathcal{L}_\beta(\widehat{\mathbf{w}}^{k+1}) - \mathcal{L}_\beta(\mathbf{w}^k) - \delta\beta\|\mathbf{x}^{k+1} - \widehat{\mathbf{x}}^k\|^2 \\
\leq \;& \frac{6(1+\tau)\psi_1(s)c_x^2\beta}{s\sigma_A}\left(\|\widehat{\mathbf{d}}_x^{k-1}\|^2 - \|\widehat{\mathbf{d}}_x^k\|^2 + \|\mathbf{d}_y^{k-1}\|^2 - \|\mathbf{d}_y^k\|^2\right) \\
& - \frac{\tau\beta}{2}\|\widehat{\mathbf{d}}_x^k\|^2_{\mathcal{D}_x} - \frac{\tau\beta}{2}\|\mathbf{d}_y^k\|^2_{\mathcal{D}_y} - \frac{\tau}{s\beta}\|\mathbf{d}_\lambda^k\|^2 - (1+\tau)\beta\left(\|\mathbf{d}_x^k\|^2_{\overline{\mathcal{D}}_x} + 12\|\mathbf{d}_y^k\|^2_{\overline{\mathcal{D}}_y}\right) \\
& + \frac{(1+\tau)\psi_2(s)}{s\beta\sigma_A}\left(\|A^\mathsf{T}\mathbf{d}_\lambda^{k-1}\|^2 - \|A^\mathsf{T}\mathbf{d}_\lambda^k\|^2\right) - \delta\beta\|\widetilde{\mathbf{d}}_x^{k+1}\|^2, \qquad (4.34)
\end{aligned}$$

where $\delta \in (0,1)$, $\overline{D}_x \succeq \mathbf{0}$ and $\overline{D}_y \succeq \mathbf{0}$ are defined in (4.28) and (4.29), respectively. Thus, (4.30) follows from (4.34) and the definition of E^{k+1} in (4.26). Similarly, by (4.10) and

$\widetilde{\mathbf{d}}_x^k = \mathbf{x}^k - \widehat{\mathbf{x}}^{k-1}$, we have

$$\begin{aligned}\mathcal{L}_\beta(\widehat{\mathbf{w}}^{k+1}) - \mathcal{L}_\beta(\widehat{\mathbf{w}}^k) &\leq \mathcal{L}_\beta(\widehat{\mathbf{w}}^{k+1}) - \mathcal{L}_\beta(\mathbf{w}^k) - \delta\beta\|\mathbf{x}^k - \widehat{\mathbf{x}}^{k-1}\|^2 \\ &= \mathcal{L}_\beta(\widehat{\mathbf{w}}^{k+1}) - \mathcal{L}_\beta(\mathbf{w}^k) - \delta\beta\|\widetilde{\mathbf{d}}_x^k\|^2.\end{aligned}$$

So, plugging (4.33) into (4.32), we can similarly derive by the definition of $\widehat{E}^{k+1}$ in (4.26) that (4.31) holds. □

4.3.2. Global Convergence and Sublinear Convergence Rate

We say $\mathbf{w}^* = (\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*)$ is a stationary point of problem (4.1) if $\mathbf{0} \in \partial\mathcal{L}(\mathbf{w}^*)$, namely,

$$\mathbf{0} = \nabla f(\mathbf{x}^*) - A^\mathsf{T}\boldsymbol{\lambda}^*, \quad \mathbf{0} \in \partial g(\mathbf{y}^*) - B^\mathsf{T}\boldsymbol{\lambda}^* \quad \text{and} \quad A\mathbf{x}^* + B\mathbf{y}^* = \mathbf{b}. \tag{4.35}$$

Then, it is obvious that $\mathbf{w}^k = (\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k)$ is a stationary point of problem (4.1) if $R^{k+1} = 0$, where R^k is defined in (4.8). Hence, in the following global convergence theorem, we assume $R^k \neq 0$ for all k and an infinite sequence $\{\mathbf{w}^k\}$ is generated by Algorithm 4.1.

Theorem 4.3.2. *Suppose the Assumptions 4.3.1 and 4.3.2 hold and the iterates $\{\mathbf{w}^k\}$ generated by Algorithm 4.1 satisfy the conditions (4.4), (4.5), (4.6) and (4.7). If the parameters in Algorithm 4.1 are chosen such that (4.28) and (4.29) hold, and $\{E^k\}$ or $\{\widehat{E}^k\}$ defined in (4.26) are bounded from below, then there exists an F^* such that*

$$\lim_{k\to\infty} \mathcal{L}(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k) = \lim_{k\to\infty} \mathcal{L}_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\lambda}^k) = \lim_{k\to\infty} E^k = F^*. \tag{4.36}$$

In addition, we have

$$\lim_{k\to\infty} dist(\mathbf{0}, \partial\mathcal{L}(\mathbf{w}^k)) = \lim_{k\to\infty} dist(\mathbf{0}, \partial\mathcal{L}_\beta(\mathbf{w}^k)) = 0 \tag{4.37}$$

and any limit point $\mathbf{w}^*$ of $\{\mathbf{w}^k\}$ is a stationary point of problem (4.1).

Proof. Without loss of generality, let us assume $\{E^k\}$ are bounded from below, since the proof is almost identical if $\{\widehat{E}^k\}$ are bounded from below. Then, we obtain from (4.30) that

$$\begin{aligned} & c\sum_{k=0}^{K}\left\{\|\widehat{\mathbf{d}}_x^k\|_{\mathcal{D}_x}^2+\|\mathbf{d}_y^k\|_{\mathcal{D}_y}^2+\|\mathbf{d}_\lambda^k\|^2+\|\widetilde{\mathbf{d}}_x^{k+1}\|^2\right\} \\ \leq \quad & E^0-E^{K+1}\leq E^0-\overline{P}, \end{aligned} \tag{4.38}$$

where $c=\min\{\frac{\tau\beta}{2},\frac{\tau}{s\beta},\delta\beta\}$ and $\overline{P}$ is the lower bound of E^k. Then, (4.38), $\mathcal{D}_x\succ\mathbf{0}$ and $\mathcal{D}_y\succ\mathbf{0}$ imply that

$$\lim_{k\to\infty}\|\widetilde{\mathbf{d}}_x^k\|=0,\quad \lim_{k\to\infty}\|\widehat{\mathbf{d}}_x^k\|=0,\quad \lim_{k\to\infty}\|\mathbf{d}_y^k\|=0\quad\text{and}\quad \lim_{k\to\infty}\|\mathbf{d}_\lambda^k\|=0. \tag{4.39}$$

In addition, by (4.39), $\mathbf{d}_\lambda^k=-s\beta\widehat{\mathbf{r}}^{k+1}$ and the definition of R^k in (4.8), we have

$$\lim_{k\to\infty}\|\widehat{\mathbf{r}}^k\|=0\quad\text{and}\quad \lim_{k\to\infty}R^k=\lim_{k\to\infty}(\|\widehat{\mathbf{d}}_x^{k-1}\|+\|\mathbf{d}_y^{k-1}\|+\|\widehat{\mathbf{r}}^k\|)=0. \tag{4.40}$$

So, denoting $\mathbf{r}^k=A\mathbf{x}^k+B\mathbf{y}^k-\mathbf{b}$ and $\mathbf{d}_x^k=\mathbf{x}^{k+1}-\mathbf{x}^k$, we have from $\mathbf{r}^k=\widehat{\mathbf{r}}^k+A\widetilde{\mathbf{d}}_x^k$, $\|\mathbf{d}_x^k\|\leq\|\mathbf{x}^{k+1}-\widehat{\mathbf{x}}^k\|+\|\widehat{\mathbf{x}}^k-\mathbf{x}^k\|=\|\widetilde{\mathbf{d}}_x^{k+1}\|+\|\widehat{\mathbf{d}}_x^k\|$, (4.39) and (4.40) that

$$\lim_{k\to\infty}\|\mathbf{r}^k\|=0\qquad\text{and}\qquad \lim_{k\to\infty}\|\mathbf{d}_x^k\|=0. \tag{4.41}$$

By (4.30), we have $\{E^k\}$ is a monotonically nonincreasing sequence, which together with $\{E^k\}$ being bounded from below implies $\lim_{k\to\infty}E^k=F^*$ for some F^*. Then, it follows from the definition of E^k, (4.39) and (4.41) that (4.36) holds.

On the other hand, we have by direct calculation

$$\begin{aligned}
\partial_x \mathcal{L}_\beta(\mathbf{w}^k) &= \partial_x \mathcal{L}(\mathbf{w}^k) + \beta A^\mathsf{T} \mathbf{r}^k = \nabla f(\mathbf{x}^k) - A^\mathsf{T} \boldsymbol{\lambda}^k + \beta A^\mathsf{T} \mathbf{r}^k \\
&= \nabla_x \mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}^{k-1}) - A^\mathsf{T} \mathbf{d}_\lambda^{k-1} + (\nabla f(\mathbf{x}^k) - \nabla f(\widehat{\mathbf{x}}^{k-1})), \\
\partial_y \mathcal{L}_\beta(\mathbf{w}^k) &= \partial_y \mathcal{L}(\mathbf{w}^k) + \beta B^\mathsf{T} \mathbf{r}^k = \partial_y g(\mathbf{y}^k) - B^\mathsf{T} \boldsymbol{\lambda}^k + \beta B^\mathsf{T} \mathbf{r}^k \\
&= \partial_y \mathcal{L}_\beta(\mathbf{x}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}^{k-1}) - B^\mathsf{T} (\mathbf{d}_\lambda^{k-1} - \beta A \mathbf{d}_x^{k-1}), \\
\partial_\lambda \mathcal{L}_\beta(\mathbf{w}^k) &= \partial_\lambda \mathcal{L}(\mathbf{w}^k) = -\mathbf{r}^k. \qquad (4.42)
\end{aligned}$$

Then, it follows from (4.5), (4.7), (4.39) and (4.41) that (4.37) holds. In addition, for any limiting point $\mathbf{w}^*$ of $\{\mathbf{w}^k\}$, it follows from (4.37) and the definition of the limiting subdifferential $\partial \mathcal{L}(\mathbf{w}^*)$ that (4.35) holds. Hence, $\mathbf{w}^*$ of $\{\mathbf{w}^k\}$ is a stationary point of problem (4.1). □

From Theorem 4.3.2 and (4.40), we can see that for any limiting point $\mathbf{w}^*$ of $\{\mathbf{w}^k\}$, we have $\mathcal{L}(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*) = F(\mathbf{x}^*, \mathbf{y}^*) = f(\mathbf{x}^*) + g(\mathbf{y}^*) = F^*$. In addition, we can observe from (4.38) that

$$\min_{k \in \{1,\ldots,K\}} \left\{ \|\widetilde{\mathbf{d}}_x^{k+1}\|^2 + \|\widehat{\mathbf{d}}_x^k\|^2 + \|\mathbf{d}_y^k\|^2 + \|\widehat{\mathbf{r}}^{k+1}\|^2 \right\} = \mathcal{O}(1/K),$$

which together with (4.5) and (4.7) implies

$$\min_{k \in \{1,\ldots,K\}} \left\{ \operatorname{dist}(\mathbf{0}, \partial \mathcal{L}(\mathbf{w}^k)) \right\} = \mathcal{O}(1/\sqrt{K}).$$

In Theorem 4.3.2, we assume the parameters in Algorithm 4.1 are chosen such that the potential energy sequence $\{E^k\}$ or $\{\widehat{E}^k\}$ is uniformly bounded from below. The following theorem gives a sufficient condition to ensure the uniform lower bound of $\{\widehat{E}^k\}$, which in turn implies the the uniform lower bound of $\{E^k\}$ since $\lim_{k\to\infty} \|\widehat{\mathbf{d}}_x^k\| = 0$.

Theorem 4.3.3. *Suppose there exists a* $\overline{\beta} > 0$ *such that*

$$\inf \left\{ f(\widehat{\mathbf{x}}^{k-1}) + g(\mathbf{y}^k) + \frac{\overline{\beta}}{2}\|A\widehat{\mathbf{x}}^{k-1} + B\mathbf{y}^k - \mathbf{b}\|^2 \right\} =: \overline{P} > -\infty. \tag{4.43}$$

Then, under the conditions of Theorem 4.3.1 and $\beta \geq \overline{\beta}$*, we have* $\widehat{E}^k \geq \overline{P}$ *for all* k*.*

Proof. Since $\beta \geq \overline{\beta}$, it follows from $\boldsymbol{\lambda}^k = \boldsymbol{\lambda}^{k-1} - s\beta(A\widehat{\mathbf{x}}^{k-1} + B\mathbf{y}^k - \mathbf{b})$ and (4.43) that

$$\begin{aligned}
& \mathcal{L}_\beta(\widehat{\mathbf{w}}^k) = \mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}^k) \\
\geq \;& f(\widehat{\mathbf{x}}^{k-1}) + g(\mathbf{y}^k) - (\boldsymbol{\lambda}^k)^\mathsf{T}(A\widehat{\mathbf{x}}^{k-1} + B\mathbf{y}^k - \mathbf{b}) + \frac{\overline{\beta}}{2}\|A\widehat{\mathbf{x}}^{k-1} + B\mathbf{y}^k - \mathbf{b}\|^2 \\
\geq \;& \overline{P} + \frac{1}{s\beta}(\boldsymbol{\lambda}^k)^\mathsf{T}(\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k-1}) \\
= \;& \overline{P} + \frac{1}{2s\beta}\left(\|\boldsymbol{\lambda}^k\|^2 - \|\boldsymbol{\lambda}^{k-1}\|^2 + \|\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k-1}\|^2\right).
\end{aligned}$$

Hence, by the definition of $\widehat{E}^k$ in (4.26) and the above inequality, we have

$$\sum_{k=1}^{\infty}\left(\widehat{E}^k - \overline{P}\right) \geq \sum_{k=1}^{\infty}\left(\mathcal{L}_\beta(\widehat{\mathbf{w}}^k) - \overline{P}\right) \geq -\frac{1}{s\beta}\|\boldsymbol{\lambda}^0\|^2. \tag{4.44}$$

By Theorem 4.3.1, $\widehat{E}^k$ is monotonically decreasing. So, if there exists a $\overline{k}$ such that $\widehat{E}^{\overline{k}} < \overline{P}$, we will have $\widehat{E}^k < \overline{P}$ for all $k > \overline{k}$, which implies $\sum_{k=1}^{\infty}\left(\widehat{E}^k - \overline{P}\right) = -\infty$, which is a contradiction to (4.44). Hence, we have $\widehat{E}^k \geq \overline{P}$ for all k. □

Remark 4.3.1. *The condition (4.43) in Theorem 4.3.3 is obviously satisfied if*

$$\inf \left\{ f(\mathbf{x}) + g(\mathbf{y}) + \frac{\overline{\beta}}{2}\|A\mathbf{x} + B\mathbf{y} - \mathbf{b}\|^2 \right\} > -\infty \tag{4.45}$$

for all $\mathbf{x}$ *and* $\mathbf{y}$*. And in many applications, the function* $F(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{y})$ *is uniformly bounded from below, for example, in statistical learning both the graph-guided fused lasso model [76] and the smoothly clipped absolute deviation (SCAD) model [129] have nonnegative objective function value. Therefore, (4.45) holds.*

4.3.3. Linear Convergence Rate

In this subsection, we discuss the linear convergence of $\{E^k\}$ and $\{\mathbf{w}^k\}$ under proper conditions. Let Ω^* be the set of all stationary points of problem (4.1) satisfying (4.35), i.e.,

$$\Omega^* = \{(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*) : A^\mathsf{T}\boldsymbol{\lambda}^* = \nabla f(\mathbf{x}^*), B^\mathsf{T}\boldsymbol{\lambda}^* \in \partial g(\mathbf{y}^*), A\mathbf{x}^* + B\mathbf{y}^* = \mathbf{b}\}. \tag{4.46}$$

Note that Ω^* is a closed set. In the following, let us denote $\mathbf{w}^* = (\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*) \in \Omega^*$. We need the following additional assumption to show the linear convergence.

Assumption 4.3.3. *(a) For any $\xi \geq \inf_{\mathbf{w}} \mathcal{L}_\beta(\mathbf{w})$, there exist $\epsilon > 0$ and $\tau > 0$ such that*

$$dist(\mathbf{w}, \Omega^*) \leq \tau\, dist(\mathbf{0}, \partial\mathcal{L}_\beta(\mathbf{w})), \tag{4.47}$$

whenever $dist(\mathbf{0}, \partial\mathcal{L}_\beta(\mathbf{w})) \leq \epsilon$ and $\mathcal{L}_\beta(\mathbf{w}) \leq \xi$.

(b) Ω^* is nonempty and there exists $\omega^* > 0$ such that $\|\mathbf{w}_1 - \mathbf{w}_2\| \geq \omega^*$ whenever $\mathbf{w}_1$, $\mathbf{w}_2 \in \Omega^*$ and $F(\mathbf{x}_1, \mathbf{y}_1) \neq F(\mathbf{x}_2, \mathbf{y}_2)$.

(c) Function g is locally weakly convex near

$$\Omega_y^* := \{\mathbf{y} : \textit{there exist } \mathbf{x} \textit{ and } \boldsymbol{\lambda} \textit{ such that } (\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \in \Omega^*\},$$

that is, there exist $\varepsilon, \sigma, \delta > 0$ such that for any $\mathbf{y}_1, \mathbf{y}_2$ with $dist(\mathbf{y}_1, \Omega_y^*) \leq \epsilon$, $dist(\mathbf{y}_2, \Omega_y^*) \leq \epsilon$ and $\|\mathbf{y}_1 - \mathbf{y}_2\| \leq \delta$ and for any $\boldsymbol{\nu} \in \partial g(\mathbf{y}_2)$, it has

$$g(\mathbf{y}_1) \geq g(\mathbf{y}_2) + \langle \boldsymbol{\nu}, \mathbf{y}_1 - \mathbf{y}_2 \rangle - \sigma\|\mathbf{y}_1 - \mathbf{y}_2\|^2. \tag{4.48}$$

We have the following comments on Assumption 4.3.3. Assumption 4.3.3 (a) is a local error bound condition. Similar local error bound conditions have been often used in the convergence rate analysis of many algorithms [7, 93, 94, 117, 123]. Assumption 4.3.3

(b) essentially requires that the isocost surface of F restricted on Ω^* is properly separated. For more examples and discussions on functions satisfying the error bound conditions and the isocost properties, one may refer to references [116, 117, 140, 123]. Assumption 4.3.3 (c) requires function g is locally weakly convex near the projection of the set Ω^* of stationary points onto the $\mathbf{y}$-coordinates. Convex functions and Lipschitz continuously differential functions obviously satisfy this requirement. For more properties on weakly convex functions as well as its relations to lower-$\mathcal{C}^2$ functions, one may refer to references [2, 108].

We now give the following theorem on the linear convergence of the energy sequence $\{E^k\}$. The linear convergence of energy sequence $\{\widehat{E}^k\}$ can be similarly proved.

Theorem 4.3.4. *Suppose the conditions in Theorem 4.3.2 and Assumption 4.3.3 hold. Then, for the iterates $\{\mathbf{w}^k\}$ generated by Algorithm 4.1, the following holds:*

(i) $\lim_{k\to\infty} dist(\mathbf{w}^k, \Omega^*) = 0$*;*

(ii) if $\{\mathbf{w}^k\}$ has at least one cluster point, then for all k sufficiently large,

$$0 \leq E^{k+1} - F^* \leq \theta(E^k - F^*), \tag{4.49}$$

where $\theta \in (0,1)$ is some constant, E^k is defined in (4.26) and $F^* = \lim_{k\to\infty} E^k$ is defined in (4.36).

Proof. By (4.36) and (4.37), there exists a $\zeta \geq \inf_{\mathbf{w}} \mathcal{L}_\beta(\mathbf{w})$ such that $\mathcal{L}_\beta(\mathbf{w}^k) \leq \zeta$ for all k and $\lim_{k\to\infty} dist(\mathbf{0}, \partial\mathcal{L}_\beta(\mathbf{w}^k)) = 0$. Hence, conclusion (i) follows from Assumption 4.3.3 (a) with $\xi = \zeta$.

We now prove conclusion (ii). For any iterate $\mathbf{w}^k$, let us define a $\overline{\mathbf{w}}^k \in \Omega^*$ such that $\mathrm{dist}(\mathbf{w}^k, \Omega^*) = \|\mathbf{w}^k - \overline{\mathbf{w}}^k\|$. Since Ω^* is closed, the existence of $\overline{\mathbf{w}}^k$ is guaranteed. Then, by conclusion (i), we have

$$\lim_{k\to\infty} \|\mathbf{w}^k - \overline{\mathbf{w}}^k\| = 0. \tag{4.50}$$

In addition, we have from (4.39) and $\|\mathbf{w}^k - \mathbf{w}^{k-1}\| \leq \|\mathbf{d}_x^{k-1}\| + \|\mathbf{d}_y^{k-1}\| + \|\mathbf{d}_\lambda^{k-1}\|$ that

$$\lim_{k\to\infty} \|\mathbf{w}^k - \mathbf{w}^{k-1}\| = 0. \tag{4.51}$$

Therefore, we have from $\|\overline{\mathbf{w}}^k - \overline{\mathbf{w}}^{k-1}\| \leq \|\overline{\mathbf{w}}^k - \mathbf{w}^k\| + \|\mathbf{w}^k - \mathbf{w}^{k-1}\| + \|\mathbf{w}^{k-1} - \overline{\mathbf{w}}^{k-1}\|$, (4.50) and (4.51) that

$$\lim_{k\to\infty} \|\overline{\mathbf{w}}^k - \overline{\mathbf{w}}^{k-1}\| = 0. \tag{4.52}$$

So, by Assumption 4.3.3 (b) and $\overline{\mathbf{w}}^k \in \Omega^*$, there exists a constant $\overline{F}^*$ such that

$$\mathcal{L}_\beta(\overline{\mathbf{w}}^k) = \mathcal{L}_\beta(\overline{\mathbf{x}}^k, \overline{\mathbf{y}}^k, \overline{\boldsymbol{\lambda}}^k) = F(\overline{\mathbf{x}}^k, \overline{\mathbf{y}}^k) = \overline{F}^* \tag{4.53}$$

for all k sufficiently large. Now, by our assumption, $\{\mathbf{w}^k\}$ has a cluster point $\mathbf{w}^*$, i.e., there exists a subsequence $\{\mathbf{w}^{k_i}\}$ converging to $\mathbf{w}^*$. Then, we have from Theorem 4.3.2 that $\mathbf{w}^* \in \Omega^*$, and in addition, by (4.50), we have

$$\lim_{i\to\infty} \|\overline{\mathbf{w}}^{k_i} - \mathbf{w}^*\| \leq \lim_{i\to\infty} (\|\overline{\mathbf{w}}^{k_i} - \mathbf{w}^{k_i}\| + \|\mathbf{w}^{k_i} - \mathbf{w}^*\|) = 0.$$

Hence, we have from (4.53), $\mathbf{w}^* \in \Omega^*$ and Assumption 4.3.3 (b) again that $\mathcal{L}_\beta(\mathbf{w}^*) = \overline{F}^*$. So, by the lower semicontinuity of the function $\mathcal{L}_\beta(\cdot)$, we have

$$\overline{F}^* = \mathcal{L}_\beta(\mathbf{w}^*) \leq \lim_{i\to\infty} \mathcal{L}_\beta(\mathbf{w}^{k_i}) = F^*, \tag{4.54}$$

where $F^* = \lim_{k\to\infty} E^k = \lim_{k\to\infty} \mathcal{L}_\beta(\mathbf{w}^k)$ is given in Theorem 4.3.2.

By the definition of $\mathcal{L}_\beta(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda})$ in (2.28) and the update of $\boldsymbol{\lambda}^k$ in Algorithm 4.1, we have

$$\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}^k) - \mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}) = \frac{1}{s\beta}(\boldsymbol{\lambda} - \boldsymbol{\lambda}^k)^\mathsf{T}(\boldsymbol{\lambda}^{k-1} - \boldsymbol{\lambda}^k), \tag{4.55}$$

and

$$\begin{aligned}\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y}^k,\boldsymbol{\lambda})-\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y},\boldsymbol{\lambda}) &= g(\mathbf{y}^k)-g(\mathbf{y})+\boldsymbol{\lambda}^\mathsf{T}B(\mathbf{y}-\mathbf{y}^k) \\ &+\frac{\beta}{2}\left(\|A\widehat{\mathbf{x}}^{k-1}+B\mathbf{y}^k-\mathbf{b}\|^2-\|A\widehat{\mathbf{x}}^{k-1}+B\mathbf{y}-\mathbf{b}\|^2\right),\end{aligned} \tag{4.56}$$

and

$$\begin{aligned}\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y},\boldsymbol{\lambda})-\mathcal{L}_\beta(\mathbf{x},\mathbf{y},\boldsymbol{\lambda}) &= f(\widehat{\mathbf{x}}^{k-1})-f(\mathbf{x})+\boldsymbol{\lambda}^\mathsf{T}A(\mathbf{x}-\widehat{\mathbf{x}}^{k-1}) \\ &+\frac{\beta}{2}\left(\|A\widehat{\mathbf{x}}^{k-1}+B\mathbf{y}-\mathbf{b}\|^2-\|A\mathbf{x}+B\mathbf{y}-\mathbf{b}\|^2\right).\end{aligned} \tag{4.57}$$

Then, by setting $(\mathbf{x},\mathbf{y},\boldsymbol{\lambda})=\overline{\mathbf{w}}^k$ in (4.55), (4.56) and (4.57), for all k sufficiently large, we have from (4.53) and (4.54) that

$$\begin{aligned} & \mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y}^k,\boldsymbol{\lambda}^k)-F^* \\ \leq\ & \mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y}^k,\boldsymbol{\lambda}^k)-\overline{F}^*=\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1},\mathbf{y}^k,\boldsymbol{\lambda}^k)-\mathcal{L}_\beta(\overline{\mathbf{x}}^k,\overline{\mathbf{y}}^k,\overline{\boldsymbol{\lambda}}^k) \\ \leq\ & \frac{1}{s\beta}(\overline{\boldsymbol{\lambda}}^k-\boldsymbol{\lambda}^k)^\mathsf{T}(\boldsymbol{\lambda}^{k-1}-\boldsymbol{\lambda}^k)+\frac{L}{2}\|\overline{\mathbf{x}}^k-\widehat{\mathbf{x}}^{k-1}\|^2 \\ & +\frac{1}{2s^2\beta}\|\mathbf{d}_\lambda^{k-1}\|^2+g(\mathbf{y}^k)-g(\overline{\mathbf{y}}^k)+\langle B^\mathsf{T}\overline{\boldsymbol{\lambda}}^k,\overline{\mathbf{y}}^k-\mathbf{y}^k\rangle,\end{aligned} \tag{4.58}$$

where the inequality comes from Lipschitz continuity of f, $A^\mathsf{T}\overline{\boldsymbol{\lambda}}^k=\nabla f(\overline{\mathbf{x}}^k)$, $A\overline{\mathbf{x}}^k+B\overline{\mathbf{y}}^k=\mathbf{b}$ and $\mathbf{d}_\lambda^{k-1}=-s\beta\widehat{\mathbf{r}}^k$. From (4.5), there exists a $\xi_y^k\in\partial_y\mathcal{L}_\beta(\mathbf{x}^{k-1},\mathbf{y}^k,\boldsymbol{\lambda}^{k-1})$, i.e.,

$$\boldsymbol{\nu}^k:=\xi_y^k+B^\mathsf{T}\boldsymbol{\lambda}^{k-1}-\beta B^\mathsf{T}(A\mathbf{x}^{k-1}+B\mathbf{y}^k-\mathbf{b})\in\partial g(\mathbf{y}^k) \tag{4.59}$$

with $\|\xi_y^k\|\leq c_y\beta\|\mathbf{d}_y^{k-1}\|$. So, we have

$$\begin{aligned}\|\boldsymbol{\nu}^k-B^\mathsf{T}\overline{\boldsymbol{\lambda}}^k\| &\leq \|\xi_y^k\|+\|B^\mathsf{T}(\boldsymbol{\lambda}^{k-1}-\overline{\boldsymbol{\lambda}}^k)\|+\beta\|B^\mathsf{T}(A\mathbf{x}^{k-1}+B\mathbf{y}^k-\mathbf{b})\| \\ &\leq c_y\beta\|\mathbf{d}_y^{k-1}\|+\|B\|(\|\mathbf{d}_\lambda^{k-1}\|+\|\boldsymbol{\lambda}^k-\overline{\boldsymbol{\lambda}}^k\|) \\ &\quad+\beta\|B\|(\|\widehat{\mathbf{r}}^k\|+\|A\widehat{\mathbf{d}}_x^{k-1}\|).\end{aligned} \tag{4.60}$$

Now, by (4.50), we have $\lim_{k\to\infty} \|\mathbf{y}^k - \overline{\mathbf{y}}^k\| = 0$ and $\lim_{k\to\infty} \operatorname{dist}(\mathbf{y}^k, \Omega_y^*) = 0$. Hence, it follows from Assumption 4.3.3 (c) that

$$g(\overline{\mathbf{y}}^k) \geq g(\mathbf{y}^k) + \langle \boldsymbol{\nu}^k, \overline{\mathbf{y}}^k - \mathbf{y}^k \rangle - \sigma \|\overline{\mathbf{y}}^k - \mathbf{y}^k\|^2$$

for all k sufficiently large, where $\sigma > 0$ is a constant, which implies

$$\begin{aligned} & g(\mathbf{y}^k) - g(\overline{\mathbf{y}}^k) + \langle B^\mathsf{T} \overline{\boldsymbol{\lambda}}^k, \overline{\mathbf{y}}^k - \mathbf{y}^k \rangle \\ = \; & g(\mathbf{y}^k) - g(\overline{\mathbf{y}}^k) + \langle \boldsymbol{\nu}^k, \overline{\mathbf{y}}^k - \mathbf{y}^k \rangle + \langle B^\mathsf{T} \overline{\boldsymbol{\lambda}}^k - \boldsymbol{\nu}^k, \overline{\mathbf{y}}^k - \mathbf{y}^k \rangle \\ \leq \; & \sigma \|\overline{\mathbf{y}}^k - \mathbf{y}^k\|^2 + \|B^\mathsf{T} \overline{\boldsymbol{\lambda}}^k - \boldsymbol{\nu}^k\| \|\overline{\mathbf{y}}^k - \mathbf{y}^k\|. \end{aligned}$$

Hence, by (4.58), (4.60), $\|\overline{\mathbf{x}}^k - \widehat{\mathbf{x}}^{k-1}\|^2 \leq 2(\|\overline{\mathbf{x}}^k - \mathbf{x}^k\|^2 + \|\widetilde{\mathbf{d}}_x^k\|^2)$ and $\mathbf{d}_\lambda^{k-1} = -s\beta\widehat{\mathbf{r}}^k$, there exist two constants $c_1 > 0$ and $c_2 > 0$ such that

$$\begin{aligned} & \mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}^k) - F^* \\ \leq \; & \frac{1}{s\beta}(\overline{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k)^\mathsf{T}(\boldsymbol{\lambda}^{k-1} - \boldsymbol{\lambda}^k) + \frac{L}{2}\|\overline{\mathbf{x}}^k - \widehat{\mathbf{x}}^{k-1}\|^2 \\ & + \frac{1}{2s^2\beta}\|\mathbf{d}_\lambda^{k-1}\|^2 + \sigma\|\overline{\mathbf{y}}^k - \mathbf{y}^k\|^2 + \|B^\mathsf{T}\overline{\boldsymbol{\lambda}}^k - \boldsymbol{\nu}^k\| \|\overline{\mathbf{y}}^k - \mathbf{y}^k\| \\ \leq \; & c_1(\|\widehat{\mathbf{d}}_x^{k-1}\|^2 + \|\mathbf{d}_y^{k-1}\|^2 + \|\mathbf{d}_\lambda^{k-1}\|^2 + \|\widetilde{\mathbf{d}}_x^k\|^2) + c_2\|\mathbf{w}^k - \overline{\mathbf{w}}^k\|^2 \end{aligned} \tag{4.61}$$

for all k sufficiently large, where $\widetilde{\mathbf{d}}_x^k = \mathbf{x}^k - \widehat{\mathbf{x}}^{k-1}$. By (4.5), (4.7), (4.42), $\mathbf{d}_\lambda^{k-1} = -s\beta\widehat{\mathbf{r}}^k$,

$\mathbf{r}^k = \widehat{\mathbf{r}}^k + A\widetilde{\mathbf{d}}_x^k$, and $\mathbf{d}_x^{k-1} = \widetilde{\mathbf{d}}_x^k + \widehat{\mathbf{d}}_x^{k-1}$, we have

$$\begin{aligned}
& dist(\mathbf{0}, \partial\mathcal{L}_\beta(\mathbf{w}^k)) \\
\leq \;& \|\nabla_x \mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}^{k-1}) - A^\mathsf{T}\mathbf{d}_\lambda^{k-1}\| + \|\nabla f(\mathbf{x}^k) - \nabla f(\widehat{\mathbf{x}}^{k-1})\| + \|\mathbf{r}^k\| \\
& + dist\left(B^\mathsf{T}(\mathbf{d}_\lambda^{k-1} - \beta A\mathbf{d}_x^{k-1}), \partial_y \mathcal{L}_\beta(\mathbf{x}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}^{k-1})\right) \\
\leq \;& c_x\beta(\|\widehat{\mathbf{d}}_x^{k-1}\| + \|\mathbf{d}_y^{k-1}\|) + \|A^\mathsf{T}\mathbf{d}_\lambda^{k-1}\| + c_y\beta\|\mathbf{d}_y^{k-1}\| + \|B^\mathsf{T}(\mathbf{d}_\lambda^{k-1} - \beta A\mathbf{d}_x^{k-1})\| \\
& + L\|\widetilde{\mathbf{d}}_x^k\| + \frac{1}{s\beta}\|\mathbf{d}_\lambda^{k-1}\| + \|A\widetilde{\mathbf{d}}_x^k\| \\
\leq \;& c_3(\|\widehat{\mathbf{d}}_x^{k-1}\| + \|\mathbf{d}_y^{k-1}\| + \|\mathbf{d}_\lambda^{k-1}\| + \|\widetilde{\mathbf{d}}_x^k\|),
\end{aligned}$$

where $c_3 = \max\{(c_x + \|B^\mathsf{T}A\|)\beta, (c_x + c_y)\beta, 1/(s\beta) + \|A\| + \|B\|, L + \|A\| + \beta\|B^\mathsf{T}A\|\} > 0$. So, by Assumption 4.3.3 (a), we have

$$\|\mathbf{w}^k - \overline{\mathbf{w}}^k\| = dist(\mathbf{w}^k, \Omega) \leq \tau dist(\mathbf{0}, \partial\mathcal{L}_\beta(\mathbf{w}^k)) \leq \tau c_3(\|\widehat{\mathbf{d}}_x^{k-1}\| + \|\mathbf{d}_y^{k-1}\| + \|\mathbf{d}_\lambda^{k-1}\| + \|\widetilde{\mathbf{d}}_x^k\|)$$

for all k sufficiently large, which together with (4.61) gives

$$\mathcal{L}_\beta(\widehat{\mathbf{x}}^{k-1}, \mathbf{y}^k, \boldsymbol{\lambda}^k) - F^* \leq \overline{c}(\|\widehat{\mathbf{d}}_x^{k-1}\|^2 + \|\mathbf{d}_y^{k-1}\|^2 + \|\mathbf{d}_\lambda^{k-1}\|^2 + \|\widetilde{\mathbf{d}}_x^k\|^2), \tag{4.62}$$

where $\overline{c} = c_1 + 4c_2c_3^2\tau^2$. Hence, let $d^k = \|\widehat{\mathbf{d}}_x^k\|^2 + \|\mathbf{d}_y^k\|^2 + \|\mathbf{d}_\lambda^k\|^2 + \|\widetilde{\mathbf{d}}_x^{k+1}\|^2$, it follows from the definition of E^k in (4.26), (4.10) and (4.62) that

$$\begin{aligned}
E^{k+1} - F^* \;\leq\;& \mathcal{L}_\beta(\widehat{\mathbf{x}}^k, \mathbf{y}^{k+1}, \boldsymbol{\lambda}^{k+1}) - \delta\beta\|\mathbf{x}^{k+1} - \widehat{\mathbf{x}}^k\|^2 - F^* \\
& + \frac{6(1+\tau)\psi_1(s)c_x^2\beta}{s\sigma_A}\left(\|\widehat{\mathbf{d}}_x^k\|^2 + \|\mathbf{d}_y^k\|^2\right) + \frac{(1+\tau)\psi_2(s)}{s\beta\sigma_A}\|A^\mathsf{T}\mathbf{d}_\lambda^k\|^2 \\
\leq\;& \gamma\; d^k,
\end{aligned} \tag{4.63}$$

where $\gamma = \overline{c} + \max\{6(1+\tau)\psi_1(s)c_x^2\beta^2, (1+\tau)\|A\|^2\psi_2(s)\}/(s\beta\sigma_A)$. Additionally, we have by (4.30), $\mathcal{D}_x \succ \mathbf{0}$ and $\mathcal{D}_y \succ \mathbf{0}$ that $E^k \geq F^*$ and

$$E^{k+1} \leq E^k - \overline{\gamma}d^k, \tag{4.64}$$

where $\overline{\gamma} = \min\{\frac{\tau\beta}{2}\sigma_{\mathcal{D}_x}, \frac{\tau\beta}{2}\sigma_{\mathcal{D}_y}, \frac{\tau}{s\beta}, \delta\beta\} > 0$, $\sigma_{\mathcal{D}_x} > 0$ and $\sigma_{\mathcal{D}_y} > 0$ are the smallest eigenvalue of $\mathcal{D}_x$ and $\mathcal{D}_y$, respectively. Thus, by (4.63) and (4.64), for k sufficiently large, we have $0 \leq E^{k+1} - F^* \leq \theta(E^k - F^*)$, where $\theta = \gamma/(\gamma + \overline{\gamma}) \in (0, 1)$. □

Based on the linear convergence result in the previous theorem, we can establish the following linear convergence of the iterates $\{\mathbf{w}^k\}$.

Theorem 4.3.5. *Suppose the conditions in Theorem 4.3.2 and Assumption 4.3.3 hold. If the sequence $\{\mathbf{w}^k\}$ generated by Algorithm 4.1 has one cluster point, then $\{\mathbf{w}^k\}$ converges R-linearly to a stationary point of problem (4.1).*

Proof. We have from $\mathcal{D}_x$, $\mathcal{D}_y \succ \mathbf{0}$, (4.30) and $E^k \geq F^*$ for all $k \geq 0$ that

$$\begin{aligned}
\|\widehat{\mathbf{d}}_x^k\|^2 &\leq \frac{2}{\tau\beta\sigma_{\mathcal{D}_x}}(E^k - E^{k+1}) \leq M_1(E^k - F^*),\\
\|\mathbf{d}_y^k\|^2 &\leq \frac{2}{\tau\beta\sigma_{\mathcal{D}_y}}(E^k - E^{k+1}) \leq M_1(E^k - F^*),\\
\|\mathbf{d}_\lambda^k\|^2 &\leq \frac{s\beta}{\tau}(E^k - E^{k+1}) \leq M_1(E^k - F^*), \quad \text{and}\\
\|\widetilde{\mathbf{d}}_x^{k+1}\|^2 &\leq \frac{1}{\delta\beta}(E^k - E^{k+1}) \leq M_1(E^k - F^*)
\end{aligned} \tag{4.65}$$

where $M_1 = \max\{2/(\tau\beta\sigma_{\mathcal{D}_x}), 2/(\tau\beta\sigma_{\mathcal{D}_y}), s\beta/\tau, 1/(\delta\beta)\}$. In addition, by Theorem 4.3.4, there exists a constant $M_2 > 0$ such that $0 \leq E^k - F^* \leq M_2\theta^k$ for all $k \geq 0$, where $\theta \in (0, 1)$ is the constant in (4.49). Hence, it follows from (4.65) that

$$\|\widehat{\mathbf{d}}_x^k\| \leq Mq^k, \qquad \|\mathbf{d}_y^k\| \leq Mq^k, \qquad \|\mathbf{d}_\lambda^k\| \leq Mq^k \quad \text{and} \quad \|\widetilde{\mathbf{d}}_x^{k+1}\| \leq Mq^k,$$

where $M = \sqrt{M_1M_2}$ and $q = \sqrt{\theta} \in (0, 1)$. Therefore, we have

$$\|\mathbf{w}^{k+1} - \mathbf{w}^k\| \leq \|\widehat{\mathbf{d}}_x^k\| + \|\widetilde{\mathbf{d}}_x^{k+1}\| + \|\mathbf{d}_y^k\| + \|\mathbf{d}_\lambda^k\| \leq 4Mq^k.$$

Then, for any $m_2 > m_1 \geq 1$, we have

$$\|\mathbf{w}^{m_2} - \mathbf{w}^{m_1}\| \leq \sum_{k=m_1}^{m_2-1} \|\mathbf{w}^{k+1} - \mathbf{w}^k\| \leq \frac{4M}{1-q} q^{m_1},$$

which implies the sequence $\{\mathbf{w}^k\}$ is a Cauchy sequence and hence convergent. Suppose $\{\mathbf{w}^k\}$ converges to $\mathbf{w}^*$. Letting $m_2 \to \infty$ in the above inequality, we have

$$\|\mathbf{w}^* - \mathbf{w}^{m_1}\| \leq \frac{4M}{1-q} q^{m_1},$$

which shows $\{\mathbf{w}^k\}$ converges R-linearly to $\mathbf{w}^*$. Finally, Theorem 4.3.2 ensures $\mathbf{w}^*$ is a stationary point of (4.1). □

4.4. Inexact Subproblem Solution

Depending on various (e.g., smooth, convex or sparse) properties of the function g, one can design different algorithms to solve the $\mathbf{y}$-subproblem (4.2) inexactly to find $\mathbf{y}^{k+1}$ satisfying the conditions (4.4) and (4.5). Here, in this subsection, we just propose a generalized accelerated gradient method to find an inexact solution satisfying (4.6) and (4.7) of the $\mathbf{x}$-subproblem. Note that the $\mathbf{x}$-subproblem (4.3) is equivalent to

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^{n_x}} \ \Phi^k(\mathbf{x}) \ &:= \ f(\mathbf{x}) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{D}_x^k} + \mathbf{x}^\mathsf{T}\mathbf{p}^k + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{A^\mathsf{T}A} \\ &= \ h^k(\mathbf{x}) + \phi^k(\mathbf{x}), \end{aligned} \tag{4.66}$$

where $\mathbf{p}^k = -A^\mathsf{T}\left[\boldsymbol{\lambda}^k - \beta(A\mathbf{x}^k + B\mathbf{y}^{k+1} - \mathbf{b})\right]$, $\phi^k(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{p}^k + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{A^\mathsf{T}A}$ and

$$h^k(\mathbf{x}) = f(\mathbf{x}) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{D}_x^k}. \tag{4.67}$$

We also need the following assumptions in the subsection.

Assumption 4.4.1. *(a) The optimal value of* $\mathbf{x}$*-subproblem is bounded from below, i.e.,* $\Phi^* = \min_{\mathbf{x}\in\mathbb{R}^{n_x}} \ \Phi^k(\mathbf{x}) > -\infty$*, where the function* Φ^k *is defined in (4.66).*

(b) There exist constants $L_1 > 0$ and $L_2 > 0$ such that for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{n_x}$, it holds

$$-\frac{L_1}{2}\|\mathbf{z}_1 - \mathbf{z}_2\|^2 \leq f(\mathbf{z}_2) - f(\mathbf{z}_1) - \langle \nabla f(\mathbf{z}_1), \mathbf{z}_2 - \mathbf{z}_1 \rangle \leq \frac{L_2}{2}\|\mathbf{z}_1 - \mathbf{z}_2\|^2.$$

Obviously, we have $\max\{L_1, L_2\} \leq L$ by (4.12).

Assumption 4.4.2. *We assume the proximal matrix $\mathcal{D}_x^k$ chosen in the $\mathbf{x}$-subproblem is positive definite and upper bounded, i.e.,*

$$\overline{\eta}\mathbf{I} \succeq \mathcal{D}_x^k \succeq \eta\mathbf{I} \quad \textit{for some } \overline{\eta} \geq \eta > 0. \tag{4.68}$$

Under Assumptions 4.4.1 and 4.4.2, it follows from the definition h^k in (4.67) that

$$-\frac{\mu}{2}\|\mathbf{z}_1 - \mathbf{z}_2\|^2 \leq h^k(\mathbf{z}_2) - h^k(\mathbf{z}_1) - \langle \nabla h^k(\mathbf{z}_1), \mathbf{z}_2 - \mathbf{z}_1 \rangle \leq \frac{\Lambda}{2}\|\mathbf{z}_1 - \mathbf{z}_2\|^2 \tag{4.69}$$

for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{n_x}$, where $\mu = \max\{L_1 - \beta\eta, 0\}$ and $\Lambda = L_2 + \beta\overline{\eta}$.

Since we focus on solving the $\mathbf{x}$-subproblem where the outer iteration number k is fixed, for the simplicity of notation, in the following of this subsection we simply write Φ, h, ϕ and Λ for Φ^k, h^k, ϕ^k and Λ^k, respectively. Then, our algorithm for solving (4.66) is described in Algorithm 4.2, which is a generalization of the accelerated gradient method proposed in [65] for solving convex subproblems of ADMM to the case when f is not necessarily convex.

Theorem 4.4.1. *Suppose Assumptions 4.4.1 and 4.4.2 hold. Then, for the sequence $\{\mathbf{x}_t\}$ generated by Algorithm 4.2, we have*

$$\lim_{t\to\infty} \|\nabla\Phi(\mathbf{x}_t)\| = \lim_{t\to\infty} \|\nabla\Phi(\widehat{\mathbf{x}}_t)\| = 0. \tag{4.70}$$

Proof. First, apparently, by the definitions in (4.69), we have $\Lambda > \mu \geq 0$ since $\eta > 0$. When $\mu = 0$, we have h is a convex function, and it follows from Algorithm 4.2 that $\tau = 0$

```
Initialization:  Choose $\Theta > \Lambda$; Set $\breve{\mathbf{x}}_1 = \mathbf{x}_1 = \mathbf{x}^k$ and $\tau = 1 - \sqrt{\frac{\Theta-\mu}{\Theta+\mu}}$.
For $t = 1, 2, 3, \ldots$
    Set $\beta_t = \max\{\overline{\beta}_t, \tau\}$, where $\overline{\beta}_t = 2/(t+1)$.
    $\widehat{\mathbf{x}}_t = \beta_t \breve{\mathbf{x}}_t + (1-\beta_t)\mathbf{x}_t$.
    Set $\gamma_t = \beta_t \Theta (t+1)/t$.
    $\breve{\mathbf{x}}_{t+1} = \arg\min \left\{ \langle \nabla h(\widehat{\mathbf{x}}_t), \mathbf{x} \rangle + \frac{\gamma_t}{2} \|\mathbf{x} - \breve{\mathbf{x}}_t\|^2 + \phi(\mathbf{x}) \right\}$.
    $\mathbf{x}_{t+1} = \beta_t \breve{\mathbf{x}}_{t+1} + (1-\beta_t)\mathbf{x}_t$.
end
```

Algorithm 4.2. A generalized accelerated gradient method (G-AGM) for solving **x**-subproblem

and $\beta_t = \overline{\beta}_t$ for all $t \geq 1$. In this case, Algorithm 4.2 reduces to a standard accelerated gradient method (see algorithms developed in [65, 66]) for solving convex composite optimization which guarantees $\lim_{t\to\infty} \Phi(\mathbf{x}_t) = \lim_{t\to\infty} \Phi(\widehat{\mathbf{x}}_t) = \Phi^* > -\infty$. Hence, (4.70) holds.

In the following, we discuss the convergence of Algorithm 4.2 when $\mu > 0$. From the updates of $\mathbf{x}_{t+1}$ and $\widehat{\mathbf{x}}_t$, we have

$$\beta_t(\breve{\mathbf{x}}_{t+1} - \widehat{\mathbf{x}}_t) + (1-\beta_t)(\mathbf{x}_t - \widehat{\mathbf{x}}_t) = \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t = \beta_t \mathbf{s}_t, \tag{4.71}$$

where $\mathbf{s}_t = \breve{\mathbf{x}}_{t+1} - \breve{\mathbf{x}}_t$. Then, by (4.69) and (4.71), the following relations hold

$$\begin{aligned} h(\mathbf{x}_{t+1}) &\leq h(\widehat{\mathbf{x}}_t) + \langle \nabla h(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t \rangle + \frac{\Lambda}{2} \|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|^2 \\ &= h(\widehat{\mathbf{x}}_t) + \langle \nabla h(\widehat{\mathbf{x}}_t), \mathbf{x}_t - \widehat{\mathbf{x}}_t \rangle + \langle \nabla h(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{\Lambda \beta_t^2}{2} \|\mathbf{s}_t\|^2 \\ &\leq h(\mathbf{x}_t) + \frac{\mu}{2} \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \langle \nabla h(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{\Lambda \beta_t^2}{2} \|\mathbf{s}_t\|^2 . \end{aligned} \tag{4.72}$$

Furthermore, by (4.71), (4.72), $\mathbf{x}_{t+1} = \beta_t \breve{\mathbf{x}}_{t+1} + (1-\beta_t)\mathbf{x}_t$ and the convexity of function ϕ, we have

$$\begin{aligned}
&\Phi(\mathbf{x}_{t+1}) = h(\mathbf{x}_{t+1}) + \phi(\mathbf{x}_{t+1}) \\
\leq\ & \beta_t \left[h(\mathbf{x}_t) + \langle \nabla h(\widehat{\mathbf{x}}_t), \breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\rangle + \phi(\breve{\mathbf{x}}_{t+1})\right] + (1-\beta_t)\left[h(\mathbf{x}_t) + \phi(\mathbf{x}_t)\right] \\
& + \frac{\mu}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \frac{\Lambda\beta_t^2}{2}\|\mathbf{s}_t\|^2 \\
=\ & \beta_t \Big[h(\mathbf{x}_t) + \langle \nabla h(\widehat{\mathbf{x}}_t), \breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\rangle + \frac{\gamma_t}{2}\|\mathbf{s}_t\|^2 + \phi(\breve{\mathbf{x}}_{t+1})\Big] \\
& + (1-\beta_t)\Phi(\mathbf{x}_t) + \frac{\mu}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \frac{\Lambda\beta_t^2 - \gamma_t\beta_t}{2}\|\mathbf{s}_t\|^2. \qquad (4.73)
\end{aligned}$$

Now, it follows from

$$\breve{\mathbf{x}}_{t+1} = \arg\min\Big\{\langle \nabla h(\widehat{\mathbf{x}}_t), \mathbf{x}\rangle + \frac{\gamma_t}{2}\left\|\mathbf{x} - \breve{\mathbf{x}}_t\right\|^2 + \phi(\mathbf{x})\Big\}, \qquad (4.74)$$

and $\mathbf{s}_t = \breve{\mathbf{x}}_{t+1} - \breve{\mathbf{x}}_t$ that

$$\begin{aligned}
& \langle \nabla h(\widehat{\mathbf{x}}_t), \breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\rangle + \frac{\gamma_t}{2}\|\mathbf{s}_t\|^2 + \phi(\breve{\mathbf{x}}_{t+1}) \\
\leq\ & \frac{\gamma_t}{2}\left(\|\mathbf{x}_t - \breve{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|^2\right) + \phi(\mathbf{x}_t) - \frac{1}{2}\left\|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\right\|_{\mathcal{M}}^2, \qquad (4.75)
\end{aligned}$$

where $\mathcal{M} = \beta A^{\mathsf{T}} A$, and

$$\nabla h(\widehat{\mathbf{x}}_t) + \gamma_t \mathbf{s}_t + \nabla\phi(\breve{\mathbf{x}}_{t+1}) = \mathbf{0}. \qquad (4.76)$$

By (4.73) and (4.75), we have

$$\begin{aligned}
\Phi(\mathbf{x}_{t+1}) \leq\ & \beta_t \Big[h(\mathbf{x}_t) + \frac{\gamma_t}{2}\left(\|\mathbf{x}_t - \breve{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|^2\right) + \phi(\mathbf{x}_t) - \frac{1}{2}\left\|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\right\|_{\mathcal{M}}^2\Big] \\
& + (1-\beta_t)\Phi(\mathbf{x}_t) + \frac{\mu}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \frac{\Lambda\beta_t^2 - \gamma_t\beta_t}{2}\|\mathbf{s}_t\|^2 \\
\leq\ & \Phi(\mathbf{x}_t) + \frac{\mu}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \frac{\beta_t\gamma_t}{2}\left(\|\mathbf{x}_t - \breve{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|^2\right) \\
& - \frac{\beta_t}{2}\left\|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\right\|_{\mathcal{M}}^2 - \frac{(\Theta - \Lambda)\beta_t^2}{2}\|\mathbf{s}_t\|^2, \qquad (4.77)
\end{aligned}$$

where the last inequality follows from

$$\gamma_t\beta_t - \Lambda\beta_t^2 = \beta_t^2\Theta(t+1)/t - \Lambda\beta_t^2 \geq (\Theta - \Lambda)\beta_t^2.$$

Now, note that

$$\breve{\mathbf{x}}_t - \mathbf{x}_t = \frac{1}{\beta_t}(\widehat{\mathbf{x}}_t - \mathbf{x}_t) \quad \text{and} \quad \breve{\mathbf{x}}_{t+1} - \mathbf{x}_t = \frac{1}{\beta_t}(\mathbf{x}_{t+1} - \mathbf{x}_t). \tag{4.78}$$

Then, we have from (4.77) that

$$\begin{aligned} \Phi(\mathbf{x}_{t+1}) \ &\leq \ \Phi(\mathbf{x}_t) + \frac{\mu + \gamma_t/\beta_t}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 - \frac{\gamma_t/\beta_t}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &\quad -\frac{\beta_t}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2_{\mathcal{M}} - \frac{(\Theta - \Lambda)\beta_t^2}{2}\|\mathbf{s}_t\|^2. \end{aligned} \tag{4.79}$$

For $t \geq 2$, by (4.78), we obtain

$$\begin{aligned} \widehat{\mathbf{x}}_t - \mathbf{x}_t \ &= \ \beta_t(\breve{\mathbf{x}}_t - \mathbf{x}_t) = \beta_t(\breve{\mathbf{x}}_t - \mathbf{x}_{t-1} + \mathbf{x}_{t-1} - \mathbf{x}_t) \\ &= \ \beta_t\left(\frac{1}{\beta_{t-1}}(\mathbf{x}_t - \mathbf{x}_{t-1}) + \mathbf{x}_{t-1} - \mathbf{x}_t\right) \\ &= \ \theta_t(\mathbf{x}_t - \mathbf{x}_{t-1}), \end{aligned} \tag{4.80}$$

where $\theta_t = \frac{\beta_t}{\beta_{t-1}}(1 - \beta_{t-1})$. In addition, by defining $\beta_0 = 1$ and $\mathbf{x}_0 = \mathbf{x}_1$, we can see (4.80) holds for all $t \geq 1$. Hence, for $t \geq 1$ it follows from (4.79) that

$$\begin{aligned} \Phi(\mathbf{x}_{t+1}) \ &\leq \ \Phi(\mathbf{x}_t) + \frac{(\gamma_t/\beta_t + \mu)\theta_t^2}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 - \frac{\gamma_t/\beta_t}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &\quad -\frac{\beta_t}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2_{\mathcal{M}} - \frac{(\Theta - \Lambda)\beta_t^2}{2}\|\mathbf{s}_t\|^2. \end{aligned} \tag{4.81}$$

Since $\gamma_t/\beta_t = \Theta(t+1)/t$, we have

$$\gamma_t/\beta_t - \gamma_{t+1}/\beta_{t+1} = \Theta/(t^2 + t) > 0.$$

So, we have from (4.81) that

$$\begin{aligned} &\Phi(\mathbf{x}_{t+1}) + \frac{\eta_{t+1}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ \leq \ &\Phi(\mathbf{x}_t) + \frac{\eta_t}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 - \frac{\gamma_{t+1}/\beta_{t+1} - \eta_{t+1}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &-\frac{\beta_t}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2_{\mathcal{M}} - \frac{(\Theta - \Lambda)\beta_t^2}{2}\|\mathbf{s}_t\|^2, \end{aligned} \tag{4.82}$$

where $\eta_t = (\gamma_t/\beta_t + \mu)\theta_t^2$.

Now, by the choice of β_t in Algorithm 4.2 and $\mu > 0$, we have

$$\beta_t = \max\{\overline{\beta}_t, \tau\}, \tag{4.83}$$

where $\tau = 1 - \sqrt{(\Theta - \mu) / (\Theta + \mu)} > 0$. So, for all $t \geq 1$, we have $\beta_t/\beta_{t-1} \leq 1$ and

$$\theta_t = \beta_t/\beta_{t-1}(1 - \beta_{t-1}) \leq 1 - \beta_{t-1} \leq \sqrt{(\Theta - \mu) / (\Theta + \mu)} < 1. \tag{4.84}$$

Then, by (4.84) and $\gamma_t/\beta_t = \Theta(t+1)/t > \Theta$, for all $t \geq 1$, we have

$$\begin{aligned} \gamma_t/\beta_t - \eta_t &= \gamma_t/\beta_t - (\gamma_t/\beta_t + \mu)\theta_t^2 = \gamma_t/\beta_t(1 - \theta_t^2) - \mu\theta_t^2 \\ &\geq \Theta(1 - \theta_t^2) - \mu\theta_t^2 = \Theta - (\Theta + \mu)\theta_t^2 \geq \Theta - (\Theta + \mu)\frac{\Theta - \mu}{\Theta + \mu} = \mu. \end{aligned} \tag{4.85}$$

Hence, it follows from (4.82), (4.83) and (4.85) that

$$\begin{aligned} &\Phi(\mathbf{x}_{t+1}) + \frac{\eta_{t+1}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ \leq\ &\Phi(\mathbf{x}_t) + \frac{\eta_t}{2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 - \frac{\mu}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &-\frac{\beta_t}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|_{\mathcal{M}}^2 - \frac{(\Theta - \Lambda)\tau^2}{2}\|\mathbf{s}_t\|^2 \end{aligned} \tag{4.86}$$

for all $t \geq 1$. Since $\Phi(\mathbf{x})$ is bounded from below by Assumption 4.4.1, we can obtain from (4.86), $\mu > 0$, $\tau > 0$ and $\Theta > \Lambda$ that

$$\sum_{t=\bar{t}}^{\infty} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 < \infty \quad \text{and} \quad \sum_{t=\bar{t}}^{\infty} \|\breve{\mathbf{x}}_{t+1} - \breve{\mathbf{x}}_t\|^2 = \sum_{t=t_0}^{\infty} \|\mathbf{s}_t\|^2 < \infty,$$

which implies

$$\lim_{t\to\infty} \|\mathbf{x}_{t+1} - \mathbf{x}_t\| = 0 \quad \text{and} \quad \lim_{t\to\infty} \|\breve{\mathbf{x}}_{t+1} - \breve{\mathbf{x}}_t\| = 0. \tag{4.87}$$

Since $\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t = \beta_t(\breve{\mathbf{x}}_{t+1} - \breve{\mathbf{x}}_t)$, we have from (4.87) that $\lim_{t\to\infty} \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\| = 0$. Then, we have from (4.78) that

$$\lim_{t\to\infty} \|\breve{\mathbf{x}}_t - \mathbf{x}_t\| \leq 1/\tau \lim_{t\to\infty} \|\widehat{\mathbf{x}}_t - \mathbf{x}_t\| = 0. \tag{4.88}$$

Therefore, (4.70) follows from (4.76), (4.87),(4.88) and the Lipschitz continuity of ∇f and $\nabla\phi$. □

By Theorem 4.4.1, any cluster point of $\{\mathbf{x}_t\}$ will be a stationary point of the $\mathbf{x}$-subproblem (4.66). Now suppose $\liminf_{t\to\infty}\|\mathbf{x}_t-\mathbf{x}^k\|>0$. Otherwise, $\mathbf{x}^k$ is a stationary point of the $\mathbf{x}$-subproblem. We discuss the iterates $\{\mathbf{x}_t\}$ generated by Algorithm 4.2 will essentially satisfy conditions (4.6) and (4.7). First, by $\nabla\Phi(\mathbf{x}) = \nabla_x\mathcal{L}_\beta(\mathbf{x},\mathbf{y}^{k+1},\boldsymbol{\lambda}^k) + \beta\mathcal{D}_x^k(\mathbf{x}-\mathbf{x}^k)$ and $\lim_{t\to\infty}\nabla\Phi(\mathbf{x}_t) = \mathbf{0}$, the condition (4.7) will be satisfied by setting $\widehat{\mathbf{x}}^k = \mathbf{x}_t$ for any $c_x > \overline{\eta}$ and all t sufficiently large. Second, since $\mathbf{x}_0 = \mathbf{x}_1 = \mathbf{x}^k$, we have from (4.82) that

$$\Phi(\widehat{\mathbf{x}}^k) = \Phi(\mathbf{x}_t) \le \Phi(\mathbf{x}_1) = \Phi(\mathbf{x}^k) \tag{4.89}$$

for $t \ge 1$. Note that $\Phi(\widehat{\mathbf{x}}^k) \le \Phi(\mathbf{x}^k)$ is equivalent to

$$\frac{\beta}{2}\|\widehat{\mathbf{x}}^k-\mathbf{x}^k\|^2_{\mathcal{D}_x^k} + \mathcal{L}_\beta(\widehat{\mathbf{x}}^k,\mathbf{y}^{k+1},\boldsymbol{\lambda}^k) \le \mathcal{L}_\beta(\mathbf{x}^k,\mathbf{y}^{k+1},\boldsymbol{\lambda}^k).$$

So, with the choice of $\mathcal{D}_x^k$ satisfying condition (4.68), the condition (4.6) holds with $\mathcal{D}_x = \eta\mathbf{I}$ by setting $\widehat{\mathbf{x}}^k = \mathbf{x}_t$ for all $t \ge 1$.

4.5. Numerical Experiments

In this section, we evaluate the performance of Algorithm 4.1 on numerical experiments. The convergence results require that the parameters in Algorithm 4.1 are chosen such that (4.28) and (4.29) hold and $\{E^k\}$ or $\{\widehat{E}^k\}$ defined in (4.26) are bounded from below. However, the condition (4.28) depends on the Lipschitz constant L, which is usually unknown for general nonlinear function f and a poor estimate of its value may severely deteriorate the algorithm performance. Fortunately, a closer inspection on the convergence proof (see inequality (4.25)) reveals that the convergence results still hold as long as

$$\|\nabla f(\widehat{\mathbf{x}}^k) - \nabla f(\mathbf{x}^k)\| \leq L\|\widehat{\mathbf{x}}^k - \mathbf{x}^k\| \tag{4.90}$$

holds for all k sufficiently large. Here, L may be some constant smaller than the true Lipschitz constant. Hence, in numerical experiments, we gradually estimate the Lipschitz constant by starting with some $L^0 > 0$ and for $k = 0, 1, \ldots$, update L^k as

$$L^{k+1} = \begin{cases} \rho L^k, & \text{if } \|\nabla f(\widehat{\mathbf{x}}^k) - \nabla f(\mathbf{x}^k)\| > L^k\|\widehat{\mathbf{x}}^k - \mathbf{x}^k\|, \\ L^k, & \text{otherwise}, \end{cases} \tag{4.91}$$

where $\rho > 1$ is some parameter. Since ∇f is Lipschitz continuous, we see that L^k can only be increased finite number of times. Hence, L^k will remain as a constant L such that (4.90) holds for all k sufficiently large. Under the above choice of L^k , we dynamically update β by $\beta^k = L^k/c_\beta$ at the k-th iteration for some $c_\beta \in (0, 1)$. We require that L^0 and c_β are chosen such that for all $\beta \geq \beta^0 = L^0/c_\beta$, the functions $\mathcal{L}_x^k(\cdot)$ and $\mathcal{L}_y^k(\cdot)$ are bounded from below and (4.45) holds with $\overline{\beta} = \beta^0$. Hence, we can always solve the subproblems inexactly as required by Algorithm 4.1, and $\{E^k\}$ (also $\{\widehat{E}^k\}$) will be bounded from below by Theorem 4.3.3. So, to ensure global convergence, by Theorem 4.3.2 and the above setting, we only need to require c_β and the parameters in Algorithm 4.1 are chosen such that

$$\frac{\varphi(\tau)}{2}\mathcal{D}_x^k - \frac{\psi_1(s)\left[2(c_\beta + c_x)^2 + 6c_x^2\right]}{s\sigma_A}\mathbf{I} \succeq \mathbf{0} \quad \text{and} \quad \frac{\varphi(\tau)}{24}\mathcal{D}_y^k - \frac{\psi_1(s)c_x^2}{s\sigma_A}\mathbf{I} \succeq \mathbf{0} \tag{4.92}$$

for some $\tau \in (0, 1)$, where $\varphi(\tau) = (1 - \tau)/(1 + \tau)$. In our numerical experiments, the parameters are chosen as

$$c_\beta = c_x = \frac{1}{13},\ \mathcal{D}_x^k = \mathcal{D}_y^k = \frac{1}{6}\mathbf{I},\ s = 1,\ \rho = 1.01,\ \eta = 1.2,\ \text{and } \delta = 0.1.$$

The above choices of parameters satisfy condition (4.92) with τ sufficiently small in $(0,1)$, since $\sigma_A = 1$ in our experiments.

4.5.1. The SCAD Penalty Problem

Recall the following smoothly clipped absolute deviation (SCAD) penalty problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) := \frac{1}{2}\|H\mathbf{x}-\mathbf{u}\|^2 + \sum_{i=1}^{n} p_\kappa\left(|\mathbf{x}_i|\right), \tag{4.93}$$

where $H \in \mathbb{R}^{m\times n}, \mathbf{u} \in \mathbb{R}^m$ and the nonconvex SCAD penalty $p_\kappa(\cdot)$ is defined as

$$p_\kappa(\theta) := \begin{cases} \kappa\theta, & \theta \le \kappa, \\ \frac{-\theta^2+2c\kappa\theta-\kappa^2}{2(c-1)}, & \kappa < \theta \le c\kappa, \\ \frac{(c+1)\kappa^2}{2}, & \theta > c\kappa, \end{cases}$$

with $c > 2$ and $\kappa > 0$ being the knots of the quadratic spline function. Clearly, the above problem can be reformulated as a special case of (4.1):

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^n} \ \frac{1}{2}\|H\mathbf{x}-\mathbf{u}\|^2 + \sum_{i=1}^{n} p_\kappa\left(|\mathbf{y}_i|\right) \\ &\text{s.t.} \quad \mathbf{x}-\mathbf{y} = \mathbf{0}. \end{aligned} \tag{4.94}$$

Then, (4.94) is in the form of (4.1) as $f(\mathbf{x}) = \frac{1}{2}\|H\mathbf{x}-\mathbf{u}\|^2$, $g(\mathbf{y}) = \sum_{i=1}^{n} p_\kappa\left(|\mathbf{y}_i|\right)$, $A = \mathbf{I}$, $B = -\mathbf{I}$ and $\mathbf{b} = \mathbf{0}$. Applying I-ADMM in Algorithm 4.1 and G-AGM in Algorithm 4.2 with $\mathcal{D}_y^k = \eta_y\mathbf{I}$ and $D_x^k = \eta_x\mathbf{I}$, we have the following updates:

$$\begin{cases} \mathbf{y}^{k+1} = \arg\min_{\mathbf{y}\in\mathbb{R}^n} \left\{ \sum_{i=1}^{n} p_\kappa\left(|\mathbf{y}_i|\right) + \frac{(1+\eta_y)\beta}{2}\left\|\mathbf{y} - \frac{\mathbf{x}^k+\eta_y\mathbf{y}^k-\boldsymbol{\lambda}^k/\beta}{1+\eta_y}\right\|^2 \right\}, \\ \breve{\mathbf{x}}_{t+1} = \frac{1}{\gamma_t+\beta}\left[\boldsymbol{\lambda}^k + \beta(\eta_x\mathbf{x}^k + \mathbf{y}^{k+1}) - (H^\mathsf{T}H + \beta\eta_x\mathbf{I})\widehat{\mathbf{x}}_t + \gamma_t\breve{\mathbf{x}}_t + H^\mathsf{T}\mathbf{u}\right], \end{cases}$$

where the $\mathbf{y}$-subproblem has closed form solution as we mentioned in Chapter 3.

We choose $L^0 = 1$ in this experiment, since the function f here is nonnegative. We compare I-ADMM with several algorithms for solving the SCAD penalty problem, which are NL-ADMM [107], P-ADMM [86], BP-ADMM (Algorithm 2, [14]),

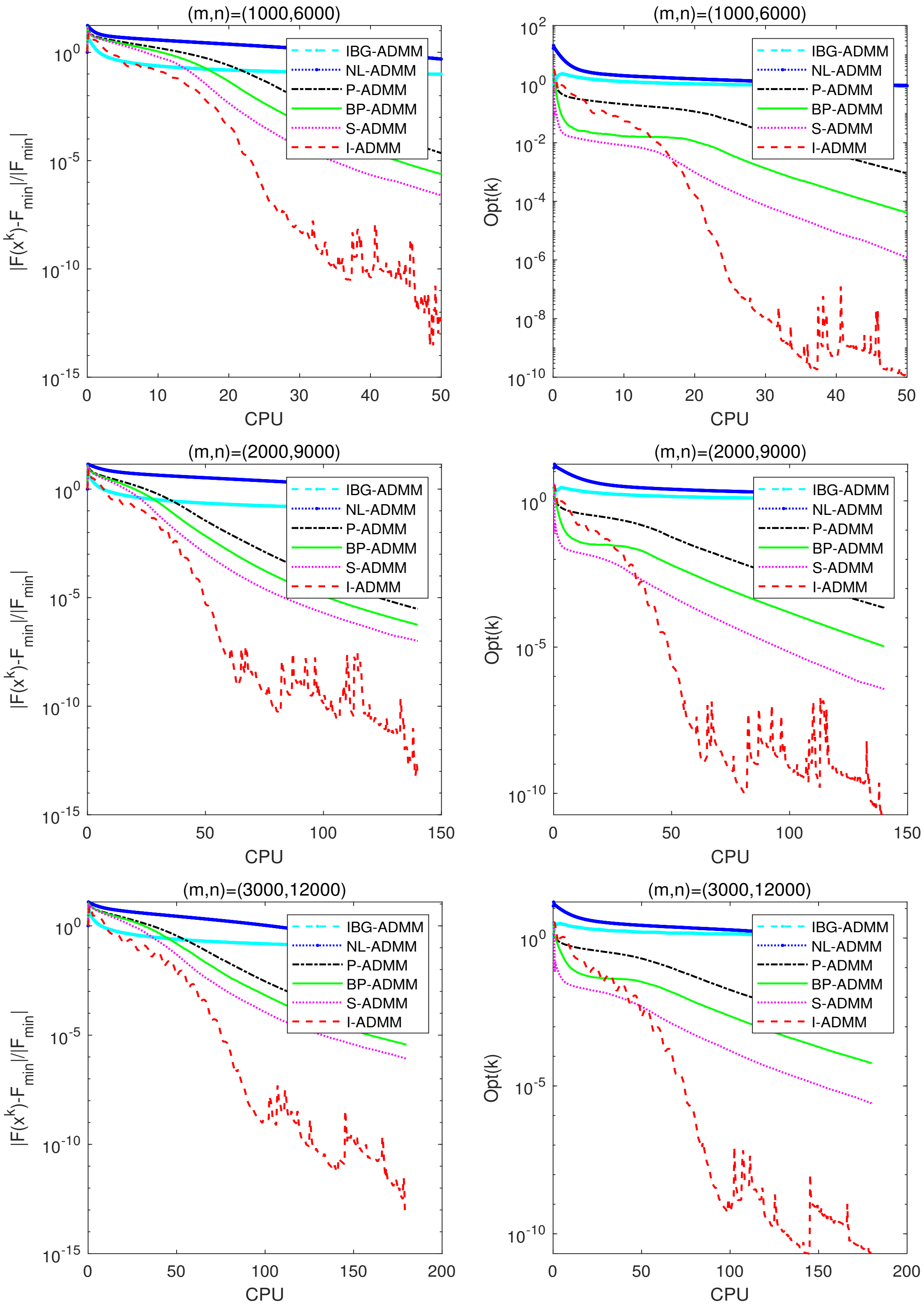


Figure 4.1. Comparison of state-of-the-art algorithms for the SCAD penalty problem

S-ADMM [79] and IBG-ADMM [129], where

- NL-ADMM uses the tuned value $\beta = 300$ and $s = 1.6$ as the dual stepsize;
- P-ADMM uses $\beta = 5.1L$ as the penalty value according to [86, Example 1];
- BP-ADMM uses $t_k = \beta$ which is 1.2 times the maximal value satisfying the involved conditions (14) and (15) in [14] (also see [14, Assumption 1]);
- S-ADMM uses the tuned stepsizes $(\alpha, \theta) = (0.05, 1.2)$ and the penalty parameter is chosen to be larger than the maximal eigenvalue of the involved quadratic function(see [79, Assumption 3.1]);
- IBG-ADMM [129] solves (4.94) by introducing variable $\mathbf{y} = H\mathbf{x} - \mathbf{u}$ (see [129, Section 4.2] for more details on the implementation and parameter settings).

Same as those used in [129], the parameters in function p_κ is set as $(c, \kappa) = (3.7, 0.1)$. We first generate a matrix $\overline{H}$ with each component $\overline{H}_{ij} \sim \mathcal{N}(0, 1)$. We then normalize each column of $\overline{H}$ and take it as H. We take $\mathbf{x}^* \in \mathbb{R}^m$ to be a random sparse vector with the density $100/n$ and then set $\mathbf{u} = H\mathbf{x}^* + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 100/n)$. The following optimality error $\mathrm{Opt(k)} = \max\left\{\|\mathbf{x}^k - \mathbf{y}^k\|, \|H^\mathsf{T}(H\mathbf{x}^k - \mathbf{u}) - \boldsymbol{\lambda}^k\|\right\}$ is used for the iterates generated by algorithms except IBG-ADMM, while for IBG-ADMM, we have $\mathrm{Opt(k)} = \max\left\{\|H\mathbf{x}^k - \mathbf{y}^k - \mathbf{u}\|, \|\mathbf{y}^k + \boldsymbol{\lambda}^k\|\right\}$ since it solves problem (4.94) in a different format.

Figure 4.1 presents the convergence curves of $|F(\mathbf{x}^k) - F_{\min}|/|F_{\min}|$ and Opt(k) versus CPU time, where $F_{\min}$ is the minimum of the objective values obtained by all the comparison algorithms. We can see from Figure 4.1 that I-ADMM performs significantly better than other comparison algorithms in terms of the running time to a certain accuracy.

I-ADMM can always obtain an objective function value with higher accuracy and smaller optimality error Opt(k). This efficiency is contributed by the adaptive inexact subproblem solution, the expansion linesearch step and the adaptive way for updating the Lipschitz constant in (4.91).

4.5.2. The Nonconvex Quadratic Programming Problem

In this subsection, we consider the following nonconvex quadratic programming (NQP) problem

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\mathbf{x}^\mathsf{T} G\mathbf{x} - \mathbf{g}^\mathsf{T}\mathbf{x} \\ &\text{s.t.} \quad \mathbf{v} \le A\mathbf{x} \le \mathbf{u}, \end{aligned} \tag{4.95}$$

where $G \in \mathbb{R}^{n\times n}$ is symmetric but may not be positive semidefinite, $A \in \mathbb{R}^{m\times n}$, $\mathbf{g} \in \mathbb{R}^n$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^m$. When $A = I$, the problem (4.95) will be reduced to a bound constrained quadratic programming problem, which already has many applications. Note that since efficient projection on the feasible set of (4.95), which is a polyhedral, is in general nontrivial.

The problem (4.95) can be also rewritten in the format of (4.1) as

$$\min_{(\mathbf{x},\mathbf{y})\in\mathbb{R}^n\times\mathbb{R}^m} \ \frac{1}{2}\mathbf{x}^\mathsf{T} G\mathbf{x} - \mathbf{g}^\mathsf{T}\mathbf{x} + \delta_{\mathcal{C}}(\mathbf{y}) \qquad \text{subject to} \quad A\mathbf{x} = \mathbf{y}, \tag{4.96}$$

where $\delta_{\mathcal{C}}$ is the indicator function of the set $\mathcal{C} = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{v} \le \mathbf{y} \le \mathbf{u}\}$. Applying I-ADMM in Algorithm 4.1 and G-AGM in Algorithm 4.2 to the problem (4.96) with $\mathcal{D}_y^k = \eta_y\mathbf{I}$ and $D_x^k = \eta_x\mathbf{I}$ involves solving the following subproblems:

$$\mathbf{y}^{k+1} = \arg\min_{\mathbf{y}\in\mathbb{R}^n} \delta_{\mathcal{C}}(\mathbf{y}) + \frac{(1+\eta_y)\beta}{2}\left\|\mathbf{y} - \mathbf{q}\right\|^2 \quad \text{and} \quad (\frac{\gamma_t}{\beta}\mathbf{I} + A^\mathsf{T} A)\breve{\mathbf{x}}_{t+1} = \mathbf{b},$$

where $\mathbf{q} := \frac{A\mathbf{x}^k + \eta_y\mathbf{y}^k - \boldsymbol{\lambda}^k/\beta}{1+\eta_y}$ and $\mathbf{b} := \frac{1}{\beta}A^\mathsf{T}\boldsymbol{\lambda}^k + \eta_x\mathbf{x}^k + A^\mathsf{T}\mathbf{y}^{k+1} - (\eta_x\mathbf{I} + \frac{1}{\beta}G)\widehat{\mathbf{x}}_t + \frac{1}{\beta}(\gamma_t\breve{\mathbf{x}}_t + \mathbf{g})$.

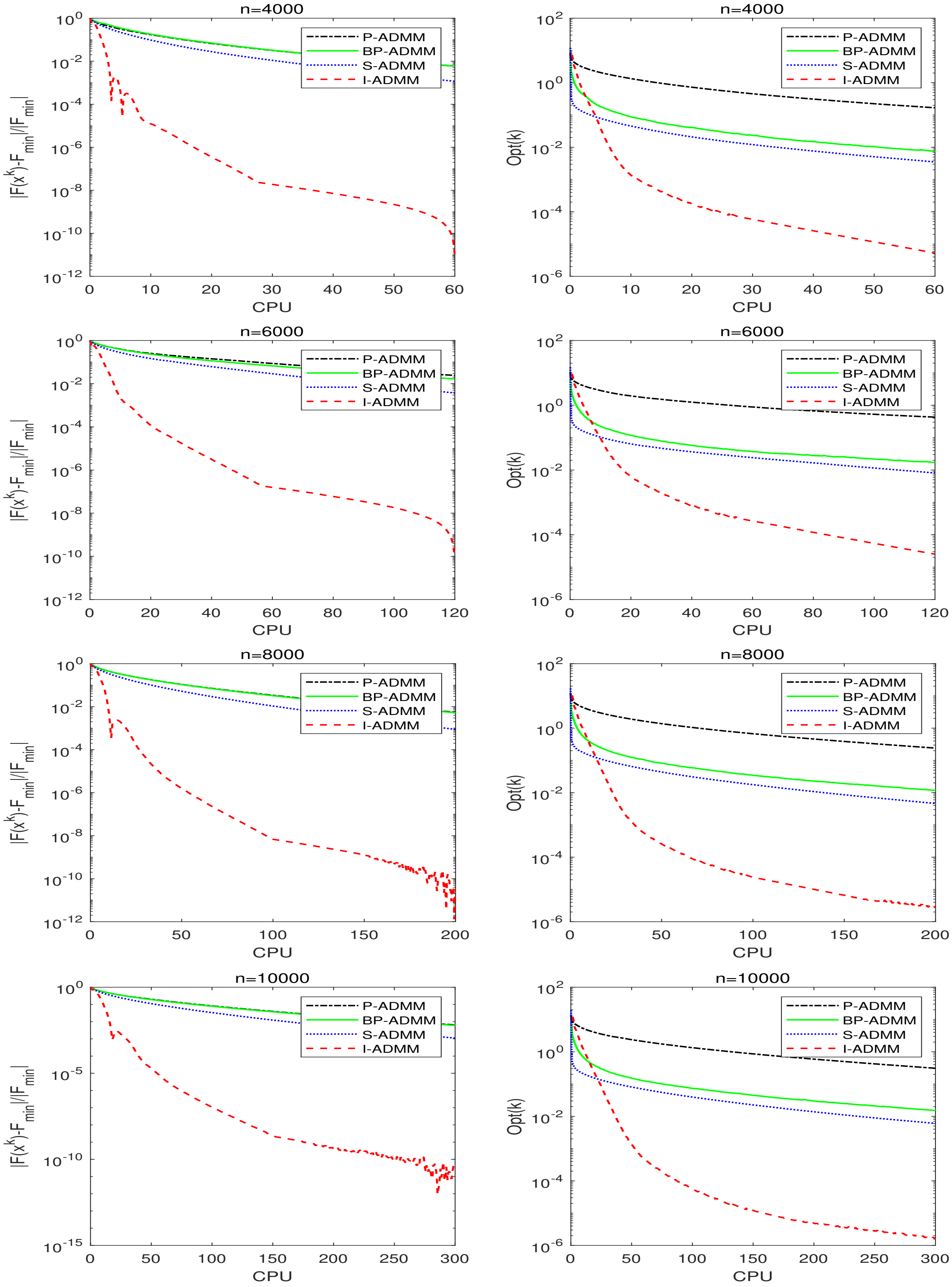


Figure 4.2. Comparison of different algorithms for solving the NQP problem.

Observe that both of the above subproblems admit closed form solutions. And when $m \ll n$, the Sherman-Morrison-Woodbury Formula should be used to solve $\breve{\mathbf{x}}_{t+1}$ as

$$\breve{\mathbf{x}}_{t+1} = \frac{\beta}{\gamma_t}\mathbf{b} - \frac{\beta^2}{\gamma_t^2}A^\mathsf{T}\Big(\mathbf{I} + \frac{\beta}{\gamma_t}AA^\mathsf{T}\Big)^{-1}A\mathbf{b}.$$

In our numerical experiments, A is always generated to be an orthogonal matrix, i.e., $A^\mathsf{T}A = \mathbf{I}$. Note that even for A being an orthogonal matrix, projection on the feasible set of problem (4.95) is in general still nontrivial. Then we randomly generate a matrix U with entries from standard normal distribution and make an orthogonal matrix A where the columns are orthogonal and spanned by the columns of U. The initialization of G and $\mathbf{g}$ is the same as that in the numerical experiments in Chapter 3. The vector $\mathbf{u}$ is set to $10 * \mathbf{e}$ where $\mathbf{e}$ is a all-one vector and the elements in $\mathbf{v}$ are all zero. To ensure the x-subproblem is bounded below, we set $L^0 = 2|\min\{\lambda_{\min}(G), 0\}| + 1$, where $\lambda_{\min}(G)$ is the smallest eigenvalue of G.

We compare I-ADMM with the aforementioned algorithms P-ADMM, BP-ADMM and S-ADMM. The rest two algorithms IBG-ADMM and NL-ADMM are not presented since their performance is much worse on solving this problem. We plot both $|F(\mathbf{x}^k) - F_{\min}|/|F_{\min}|$ and Opt(k), which is given by $\text{Opt(k)} = \max\{\|A\mathbf{x}^k - \mathbf{y}^k\|, \|G\mathbf{x}^k - \mathbf{g} - A^\mathsf{T}\boldsymbol{\lambda}^k\|\}$, against the CPU time in Figure 4.2 for n belonging to $\{2000, 4000, 6000, 8000, 10000\}$, where $F_{\min}$ denotes the minimum of the objective values obtained by the four algorithms. From Figure 4.2, we can see that I-ADMM converges much faster and obtains a solution with higher accuracy than other algorithms under the same CPU time budget.

Chapter 5. Conclusion

In this chapter, we first summarize the proposed algorithms, the unified proximal gradient method with extrapolation (UPG) and the inexact alternating direction method of multipliers (I-ADMM), and add some remarks on the relationship between the two types of the problems that UPG and I-ADMM solve. Then, we present the work that might be of interests in future.

The goal of UPG is to solve nonconvex and nonsmooth composite optimization problems, where the objective function is the sum of two components. One of the components is smooth but possibly nonconvex and the other one is convex but can be nonsmooth. UPG exploits an extrapolation step where extrapolation parameter is estimated by a linesearch technique to accelerate convergence. The algorithm obtains the optimal convergence rate for the convex case and a linear convergence rate when the problem is nonconvex under additional proper assumptions. Furthermore, a stochastic generalization of the method is proposed to solve nonconvex composite optimization problems where a component in the objective is an averaged finite sum. Numerical examples demonstrate the efficiency of both methods.

I-ADMM solves separable nonconvex and nonsmooth optimization problems with linear equality constraints. Though the objective function is also given by the summation of two parts, the two functions in the objective can be both nonconvex with one of which being nonsmooth. I-ADMM solves each subproblem inexactly to an adaptive accuracy and allows a larger range of dual stepsize. Global convergence and a linear convergence rate of I-ADMM are established under proper assumptions. In addition, a generalized accelerated gradient method is proposed to solve the smooth but possibly nonconvex subproblem.

By allowing adaptive inexact subproblem solutions, the expansion linesearch step and the adaptive way of updating the Lipschitz constant, I-ADMM performs significantly better than other state-of-the-art ADMM algorithms as shown in numerical experiments.

If we take a close look at the problems UPG and I-ADMM handle, it is interesting to observe that these two types of problems are in fact interchangeable under some conditions. If we introduce a variable $\mathbf{y}$ and replace $\mathbf{x}$ by $\mathbf{y}$ in the function p of problem (3.1), i.e., the one that UPG solves, add an indicator function $\delta_{\mathcal{X}}(\mathbf{y})$ in the objective function for the constraint $\mathbf{x} \in \mathcal{X}$ and an equality constraint $\mathbf{x} = \mathbf{y}$, then problem (3.1) is converted to the form of problem (4.1) with $g(\mathbf{y}) = p(\mathbf{y}) + \delta_{\mathcal{X}}(\mathbf{y})$ being convex, $A = \mathbf{I}$, $B = -\mathbf{I}$, and $\mathbf{b} = \mathbf{0}$. For the other direction, namely, transforming problem (4.1) to the form of problem (3.1), we have to restrict g in (4.1) to be convex as it is required in problem (3.1) and require the linear equality constraints have some special form, for example, $A = B = \mathbf{0}$, $\mathbf{b} = \mathbf{0}$, and $\mathcal{X} = \{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n : \mathbf{x} = \mathbf{y}\}$ or $A = \mathbf{I}$, $B = -\mathbf{I}$, $\mathbf{b} = \mathbf{0}$, and $\mathcal{X} = \mathbb{R}^n$. Therefore, problem (3.1) can be viewed as a special case of (4.1) and I-ADMM has a more general frame. Hence, given an optimization problem that meets the settings of both UPG and I-ADMM, UPG is preferred if we can find explicit solutions of the proximal subproblems in UPG, in other words, p has the structure such that the proximal subproblem is easy to solve. Otherwise, I-ADMM should be chosen. Obviously if there is no convexity specified for any part of the objective function, I-ADMM is supposed to be applied since UPG requires p is convex.

On the other hand, there are possible extensions of our current work and interesting topics for future research. As we mentioned previously, although numerical examples are conducted to show the possible practical applications of SUPG, we are unable to provide the convergence results and convergence rate at the moment. So, this will be a continuous work in future. In the experiments, we compare SUPG with prox-SVRG, another potential research is to implement the experiments with more state-of-the-art algorithms, for example, the accelerated mini-batch proximal gradient method with variance reduction in [104]. Note that we require part of the objective function, $p(\mathbf{x})$, is convex for UPG, an extension of UPG could be made to solve composite problems when $p(\mathbf{x})$ is nonconvex, in other words, there is no convexity required for any component in the objective function. Results for similar idea can be found in [87]. Another possibility is to extend I-ADMM to multi-block case as successful generalizations of ADMM have been presented for both convex and nonconvex cases, e.g., in [89, 35, 90, 119, 134, 63]. Additionally, generalizing I-ADMM to solve stochastic optimization problems could be an interesting topic as well, since in this way, the method could further improve the convergence speed with stochastic gradients. However, it might be complicated because we have to take into account the divergence brought by inexact solutions and not using full gradients. The variance reduced step and carefully choosing mini-batch size could be helpful to avoid the situation that the stochastic gradients are far away from the gradients in the deterministic case. Some recent results can be found in [91, 3, 92]. Additionally, when solutions with very high accuracy are required, methods only involving first-order information might get very slow at the last stage. Then a possible work to handle this issue is to combine these methods with second-order techniques to accelerate the convergence, e.g., semi-smooth Newton's method.

Extrapolation is a useful strategy for accelerating convergence of vector sequences, which has been widely used in optimization methods. As one may notice, an extrapolation step, which is a linear combination of the iterates from the past iteration, is applied in both proposed algorithms to accelerate convergence. The extrapolation parameter is adaptively selected by some linesearch technique. In fact, there is one extrapolation method called nonlinear extrapolation introduced in [33] for unconstrained quadratic minimization, where the core idea is to find a linear combination of all past iterates to minimize the square of the norm of the gradient at the extrapolated point by using the iterates and the corresponding gradients generated by first-order methods. The mechanism behind it is that, if the function is differentiable, the gradient norm at a local minimizer equals zero. The nonlinearity comes from the fact that the coefficients in the linear combination actually depend nonlinearly on the gradient at the extrapolated point. This method could be a potential way to implement extrapolation in optimization methods as it finds an optimal combination of iterates such that the gradient norm at the extrapolation point is minimal at each iteration. Obviously one drawback of this method is that we have to store all previous iterates and the gradients, which needs much memory. A potential way to overcome it is to limit the number of past iterates we want to use at the extrapolation step. More details about this extrapolation method and other acceleration and extrapolation techniques can be found in [33, 137, 17].

Bibliography

[1] Zeyuan Allen-Zhu and Yang Yuan. Improved svrg for non-strongly-convex or sum-of-non-convex objectives, 2015.

[2] D Aussel, Aris Daniilidis, and L Thibault. Subsmooth sets: functional characterizations and related concepts. *Transactions of the American Mathematical Society*, 357(4):1275–1301, 2005.

[3] Jianchao Bai, William W Hager, and Hongchao Zhang. An inexact accelerated stochastic admm for separable convex optimization. *Computational Optimization and Applications*, pages 1–40, 2022.

[4] Jianchao Bai, Jicheng Li, Fengmin Xu, and Hongchao Zhang. Generalized symmetric admm for separable convex optimization. *Computational optimization and applications*, 70(1):129–170, 2018.

[5] Rina Foygel Barber and Emil Y Sidky. Convergence for nonconvex admm, with applications to ct imaging. *arXiv preprint arXiv:2006.07278*, 2020.

[6] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[7] Amir Beck and Marc Teboulle. A linearly convergent dual-based gradient projection algorithm for quadratically constrained convex minimization. *Mathematics of Operations Research*, 31(2):398–417, 2006.

[8] Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE transactions on image processing*, 18(11):2419–2434, 2009.

[9] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[10] Stephen Becker, Jé rôme Bobin, and Emmanuel J. Candès. NESTA: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, jan 2011.

[11] Stephen R. Becker, Emmanuel J. Candès, and Michael C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, jul 2011.

[12] Dimitri Bertsekas. *Convex optimization theory*, volume 1. Athena Scientific, 2009.

[13] Robert G Bland, Donald Goldfarb, and Michael J Todd. The ellipsoid method: A survey. *Operations research*, 29(6):1039–1091, 1981.

[14] Radu Ioan Boţ and Dang-Khoa Nguyen. The proximal alternating direction method of multipliers in the nonconvex setting: convergence analysis and rates. *Mathematics of Operations Research*, 45(2):682–712, 2020.

[15] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

[16] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[17] Claude Brezinski. A general extrapolation algorithm. *Numerische Mathematik*, 35:175–187, 1980.

[18] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the complexity of steepest descent, newton's and regularized newton's methods for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization*, 20(6):2833–2852, 2010.

[19] Augustin Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

[20] Akshay Chandrashekaran and Ian R Lane. Speeding up hyper-parameter optimization by extrapolation of learning curves using previous builds. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*, pages 477–492. Springer, 2017.

[21] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[22] Nikolaos Chatzipanagiotis, Darinka Dentcheva, and Michael M Zavlanos. An augmented lagrangian method for distributed optimization. *Mathematical Programming*, 152(1):405–434, 2015.

[23] Caihua Chen, Raymond H Chan, Shiqian Ma, and Junfeng Yang. Inertial proximal admm for linearly constrained separable convex optimization. *SIAM Journal on Imaging Sciences*, 8(4):2239–2267, 2015.

[24] Caihua Chen, Bingsheng He, Yinyu Ye, and Xiaoming Yuan. The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, 155(1):57–79, 2016.

[25] Gong Chen and Marc Teboulle. A proximal-based decomposition method for convex minimization problems. *Mathematical Programming*, 64(1):81–101, 1994.

[26] Hongmei Chen, Guoyong Gu, and Junfeng Yang. A golden ratio proximal alternating direction method of multipliers for separable convex optimization. *Journal of Global Optimization*, pages 1–22, 2022.

[27] Run Chen. *On the Extensions of the Predictor-Corrector Proximal Multiplier (PCPM) Algorithm and Their Applications.* PhD thesis, Purdue University Graduate School, 2020.

[28] Xi Chen, Qihang Lin, Seyoung Kim, Jaime G Carbonell, and Eric P Xing. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2):719–752, 2012.

[29] Vašek Chvátal. Linear programming wh freeman and company. *New York*, pages 13–26, 1983.

[30] Ying Cui, Xudong Li, Defeng Sun, and Kim-Chuan Toh. On the convergence properties of a majorized admm for linearly constrained convex optimization problems with coupled objective functions. *arXiv preprint arXiv:1502.00098*, 2015.

[31] George Dantzig. Linear programming and extensions. In *Linear programming and extensions*. Princeton university press, 2016.

[32] George B Dantzig. Linear programming. *Operations research*, 50(1):42–47, 2002.

[33] Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245, 2021.

[34] Dirk Den Hertog. *Interior point approach to linear, quadratic and convex programming: algorithms and complexity*, volume 277. Springer Science & Business Media, 2012.

[35] Wei Deng, Ming-Jun Lai, Zhimin Peng, and Wotao Yin. Parallel multi-block admm with o (1/k) convergence. *Journal of Scientific Computing*, 71:712–736, 2017.

[36] Neil K Dhingra, Sei Zhen Khong, and Mihailo R Jovanović. The proximal augmented lagrangian method for nonsmooth composite optimization. *IEEE Transactions on Automatic Control*, 64(7):2861–2868, 2018.

[37] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[38] Jonathan Eckstein and Dimitri P Bertsekas. On the douglas—rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical programming*, 55:293–318, 1992.

[39] Jonathan Eckstein and Paulo JS Silva. A practical relative error criterion for aug-

mented lagrangians. *Mathematical Programming*, 141(1-2):319–348, 2013.

[40] Jonathan Eckstein and Wang Yao. Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives. *Pac. J. Optim.*, 11(4):619–644, 2015.

[41] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[42] Wilhelm Forst and Dieter Hoffmann. *Optimization—theory and practice*. Springer Science & Business Media, 2010.

[43] Michel Fortin and Roland Glowinski. *Augmented Lagrangian methods: applications to the numerical solution of boundary-value problems*. Elsevier, 2000.

[44] D. Gabay. Chapter ix applications of the method of multipliers to variational inequalities. In Michel Fortin and Roland Glowinski, editors, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, volume 15 of *Studies in Mathematics and Its Applications*, pages 299–331. Elsevier, 1983.

[45] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

[46] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & mathematics with applications*, 2(1):17–40, 1976.

[47] Alexander Vladimirovich Gasnikov and Yu E Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58(1):48–64, 2018.

[48] Saeed Ghadimi. Conditional gradient type methods for composite nonlinear and stochastic optimization. *Mathematical Programming*, 173(1):431–464, 2019.

[49] Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016.

[50] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Generalized uniformly optimal methods for nonlinear programming, 2015.

[51] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approx-

imation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016.

[52] Luana E Gibbons, Donald W Hearn, Panos M Pardalos, and Motakuri V Ramana. Continuous characterizations of the maximum clique problem. *Mathematics of Operations Research*, 22(3):754–768, 1997.

[53] Roland Glowinski. *Numerical Methods for Nonlinear Variational Problems.* Springer Berlin, Heidelberg, 1984.

[54] Roland Glowinski. On alternating direction methods of multipliers: a historical perspective. *Modeling, simulation and optimization for science and technology*, pages 59–82, 2014.

[55] Roland Glowinski. Admm and non-convex variational problems. In *Splitting Methods in Communication, Imaging, Science, and Engineering*, pages 251–299. Springer, 2016.

[56] Roland Glowinski and Americo Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(R2):41–76, 1975.

[57] Max LN Goncalves, Jefferson G Melo, and Renato DC Monteiro. Convergence rate bounds for a proximal admm with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems. *arXiv preprint arXiv:1702.01850*, 2017.

[58] Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Huang, and Jieping Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems, 2013.

[59] Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Huang, and Jieping Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *international conference on machine learning*, pages 37–45. PMLR, 2013.

[60] Martin Grötschel, László Lovász, and Alexander Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

[61] Lovász László Schrijver Alexander Grötschel, Martin. *Geometric algorithms and combinatorial optimization.* Springer, 1988.

[62] Guoyong Gu, Bingsheng He, and Junfeng Yang. Inexact alternating-direction-based contraction methods for separable linearly constrained convex optimization. *Journal of Optimization Theory and Applications*, 163(1):105–129, 2014.

[63] Ke Guo, Deren Han, David ZW Wang, and Tingting Wu. Convergence of admm for multi-block nonconvex separable optimization models. *Frontiers of Mathematics in China*, 12:1139–1162, 2017.

[64] Ke Guo, Xiaoming Yuan, and Shangzhi Zeng. Convergence analysis of ista and fista for "strongly+ semi" convex programming, 2016.

[65] William W Hager and Hongchao Zhang. Inexact alternating direction methods of multipliers for separable convex optimization. *Computational Optimization and Applications*, 73:201–235, 2019.

[66] William W Hager and Hongchao Zhang. Convergence rates for an inexact admm applied to separable convex optimization. *Computational Optimization and Applications*, 77(3):729–754, 2020.

[67] Deren Han and Xiaoming Yuan. A note on the alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, 155(1):227–238, 2012.

[68] Deren Han, Xiaoming Yuan, and Wenxing Zhang. An augmented lagrangian based parallel splitting method for separable convex minimization with applications to image processing. *Mathematics of Computation*, 83(289):2263–2291, 2014.

[69] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[70] Bingsheng He, Li-Zhi Liao, Deren Han, and Hai Yang. A new inexact alternating directions method for monotone variational inequalities. *Mathematical Programming*, 92(1):103–118, 2002.

[71] Bingsheng He, Zheng Peng, and Xiangfeng Wang. Proximal alternating direction-based contraction methods for separable linearly constrained convex optimization. *Frontiers of Mathematics in China*, 6(1):79–114, 2011.

[72] Bingsheng He, Xiaoming Yuan, and Wenxing Zhang. A customized proximal point algorithm for convex minimization with linear constraints. *Computational Optimization and Applications*, 56(3):559–572, 2013.

[73] Magnus R Hestenes. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5):303–320, 1969.

[74] Mingyi Hong, Tsung-Hui Chang, Xiangfeng Wang, Meisam Razaviyayn, Shiqian Ma, and Zhi-Quan Luo. A block successive upper-bound minimization method of multipliers for linearly constrained convex optimization. *Mathematics of Operations Research*, 45(3):833–861, 2020.

[75] Mingyi Hong, Zhi-Quan Luo, and Meisam Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1):337–364, 2016.

[76] Feihu Huang, Songcan Chen, and Zhaosong Lu. Stochastic alternating direction method of multipliers with variance reduction for nonconvex optimization. *arXiv preprint arXiv:1610.02758*, 2016.

[77] Toshihide Ibaraki and Naoki Katoh. *Resource allocation problems: algorithmic approaches.* MIT press, 1988.

[78] Naoki Ito, Akiko Takeda, and Kim-Chuan Toh. A unified formulation and fast accelerated proximal gradient method for classification. *The Journal of Machine Learning Research*, 18(1):510–558, 2017.

[79] Zehui Jia, Xue Gao, Xingju Cai, and Deren Han. The convergence rate analysis of the symmetric admm for the nonconvex separable optimization problems. *Journal of Industrial and Management Optimization*, 17(4):1943–1971, 2021.

[80] Jakub Konečnỳ, Jie Liu, Peter Richtárik, and Martin Takáč. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):242–255, 2015.

[81] Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[82] Andrei Kulunchakov and Julien Mairal. A generic acceleration framework for stochastic composite optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[83] Guanghui Lan, Zhize Li, and Yi Zhou. A unified variance-reduced accelerated gradient method for convex optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[84] Kenneth Lange. *Optimization*, volume 95. Springer Science & Business Media, 2013.

[85] Guoyin Li and Ting Kei Pong. Splitting methods for nonconvex composite optimization. *arXiv preprint arXiv:1407.0753*, 2014.

[86] Guoyin Li and Ting Kei Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4):2434–2460, 2015.

[87] Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. *Advances in neural information processing systems*, 28, 2015.

[88] Tao Lin, Lingjing Kong, Sebastian Stich, and Martin Jaggi. Extrapolation for large-batch training in deep learning. In *International Conference on Machine Learning*, pages 6094–6104. PMLR, 2020.

[89] Tian-Yi Lin, Shi-Qian Ma, and Shu-Zhong Zhang. On the sublinear convergence rate of multi-block admm. *Journal of the Operations Research Society of China*, 3:251–274, 2015.

[90] Tianyi Lin, Shiqian Ma, and Shuzhong Zhang. On the global linear convergence of the admm with multiblock variables. *SIAM Journal on Optimization*, 25(3):1478–1497, 2015.

[91] Yuanyuan Liu, Fanhua Shang, and James Cheng. Accelerated variance reduced stochastic admm. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[92] Yuanyuan Liu, Fanhua Shang, Hongying Liu, Lin Kong, Licheng Jiao, and Zhouchen Lin. Accelerated variance reduction stochastic admm for large-scale machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4242–4255, 2020.

[93] Zhi-Quan Luo and Paul Tseng. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30(2):408–425, 1992.

[94] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 1993.

[95] Harry M Markowitz. Foundations of portfolio theory. *The journal of finance*, 46(2):469–477, 1991.

[96] Angelo Miele, PE Moseley, AV Levy, and GM Coggins. On the method of multipliers for mathematical programming problems. *Journal of optimization Theory and Applications*, 10(1):1–33, 1972.

[97] Kazuo Murota. Linear programming. In *Computer Vision: A Reference Guide*, pages 1–7. Springer, 2020.

[98] Katta G Murty and Santosh N Kabadi. Some np-complete problems in quadratic and nonlinear programming. Technical report, 1985.

[99] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[100] Yu. E. Nesterov. A method of solving a convex programming problem with conver-

gence rate $0(1/k^2)$. *Sov. Math., Dokl.*, 27:372–376, 1983.

[101] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[102] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[103] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

[104] Atsushi Nitanda. Stochastic proximal gradient descent with acceleration techniques. *Advances in Neural Information Processing Systems*, 27, 2014.

[105] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[106] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[107] Linbo Qiao, Bofeng Zhang, Jinshu Su, and Xicheng Lu. Linearized alternating direction method of multipliers for constrained nonconvex regularized optimization. In *Asian Conference on Machine Learning*, pages 97–109. PMLR, 2016.

[108] R Tyrrell Rockafellar. Favorable classes of lipschitz continuous functions in subgradient optimization. 1981.

[109] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer Verlag, Heidelberg, Berlin, New York, 1998.

[110] Cornelis Roos. A full-newton step o (n) infeasible interior-point algorithm for linear optimization. *SIAM Journal on Optimization*, 16(4):1110–1136, 2006.

[111] Ron Shamir. The efficiency of the simplex method: a survey. *Management science*, 33(3):301–334, 1987.

[112] Ron Shefi and Marc Teboulle. Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. *SIAM Journal on Optimization*, 24(1):269–297, 2014.

[113] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. A proximal gradient algorithm for decentralized composite optimization. *IEEE Transactions on Signal Processing*, 63(22):6013–6023, 2015.

[114] Mikhail V Solodov and Benar Fux Svaiter. An inexact hybrid generalized proximal point algorithm and some new results on the theory of bregman functions. *Mathematics of Operations Research*, 25(2):214–230, 2000.

[115] Paul Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125(2):263–295, 2010.

[116] Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117:387–423, 2009.

[117] Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. *Computational Optimization and Applications*, 47(2):179–206, 2010.

[118] Jean-Philippe Vial. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8(2):231–259, 1983.

[119] Fenghui Wang, Wenfei Cao, and Zongben Xu. Convergence of multi-block bregman admm for nonconvex composite problems. *Science China Information Sciences*, 61:1–12, 2018.

[120] Fenghui Wang, Zongben Xu, and Hong-Kun Xu. Convergence of bregman alternating direction method with multipliers for nonconvex composite problems. *arXiv preprint arXiv:1410.8625*, 2014.

[121] Xiao Wang, Shuxiong Wang, and Hongchao Zhang. Inexact proximal stochastic gradient method for convex composite optimization. *Computational Optimization and Applications*, 68:579–618, 2017.

[122] Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of admm in nonconvex nonsmooth optimization. 2015.

[123] Bo Wen, Xiaojun Chen, and Ting Kei Pong. Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. *SIAM Journal on Optimization*, 27(1):124–145, 2017.

[124] Bo Wen, Xiaojun Chen, and Ting Kei Pong. A proximal difference-of-convex algorithm with extrapolation. *Computational optimization and applications*, 69:297–324, 2018.

[125] Zhongming Wu and Min Li. General inertial proximal gradient method for a class of nonconvex nonsmooth optimization problems. *Computational Optimization and Applications*, 73:129–158, 2019.

[126] Guangzeng Xie, Yitan Wang, Shuchang Zhou, and Zhihua Zhang. Interpolatron: Interpolation or extrapolation schemes to accelerate optimization for deep neural networks, 2018.

[127] Huiliang Xie and Jian Huang. Scad-penalized regression in high-dimensional par-

tially linear models. 2009.

[128] Jiaxin Xie, Anping Liao, and Xiaobo Yang. An inexact alternating direction method of multipliers with relative error criteria. *Optimization Letters*, 11(3):583–596, 2017.

[129] Jiawei Xu and Miantao Chao. An inertial bregman generalized alternating direction method of multipliers for nonconvex optimization. *Journal of Applied Mathematics and Computing*, pages 1–27, 2021.

[130] Jinming Xu, Ye Tian, Ying Sun, and Gesualdo Scutari. Distributed algorithms for composite optimization: unified framework and convergence analysis. *IEEE Transactions on Signal Processing*, 69:3555–3570, 2021.

[131] Yangyang Xu. Hybrid jacobian and gauss–seidel proximal block coordinate update methods for linearly constrained convex programming. *SIAM Journal on Optimization*, 28(1):646–670, 2018.

[132] Junfeng Yang and Yin Zhang. Alternating direction algorithms for \ell_1-problems in compressive sensing. *SIAM journal on scientific computing*, 33(1):250–278, 2011.

[133] Lei Yang, Ting Kei Pong, and Xiaojun Chen. Alternating direction method of multipliers for a class of nonconvex and nonsmooth problems with applications to background/foreground extraction. *SIAM Journal on Imaging Sciences*, 10(1):74–110, 2017.

[134] Maryam Yashtini. Multi-block nonconvex nonsmooth proximal admm: Convergence and rates under kurdyka–łojasiewicz property. *Journal of Optimization Theory and Applications*, 190(3):966–998, 2021.

[135] Maryam Yashtini. Convergence and rate analysis of a proximal linearized admm for nonconvex nonsmooth optimization. *Journal of Global Optimization*, 84(4):913–939, 2022.

[136] Xiyu Yu and Dacheng Tao. Variance-reduced proximal stochastic gradient descent for non-convex composite optimization. *arXiv preprint arXiv:1606.00602*, 2016.

[137] M Redivo Zaglia. *Extrapolation methods: theory and practice.* Elsevier, 2013.

[138] Lingmin Zeng and Jun Xie. Group variable selection via scad-l2. *Statistics*, 48(1):49–66, 2014.

[139] Wenliang Zhong and James Kwok. Gradient descent with proximal average for nonconvex and composite regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

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

[141] Zeyuan Allen Zhu and Yang Yuan. Univr: A universal variance reduction framework for proximal stochastic gradient method. *ArXiv*, abs/1506.01972, 2015.

[142] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010.

Vita

Miao Zhang was born in Chengdu, Sichuan, China in 1994. She completed her undergraduate studies at Huazhong Agricultural University in June 2016. She earned a Master of Science degree in Applied Statistics at Huazhong Agricultural University in June 2018. She continued her studies at Louisiana State University to pursue studies in Mathematics in August 2018. She earned a Master of Science degree in Mathematics in 2020. She is currently a candidate for a Doctor of Philosophy degree in Mathematics, which will be awarded in May 2023.