First-order Methods with Extrapolation for Some Structured Nonconvex Problems and Their Applications

ZU CHENCHEN

PhD

The Hong Kong Polytechnic University

2021

The Hong Kong Polytechnic University

Department of Applied Mathematics

First-order Methods with Extrapolation for Some Structured Nonconvex Problems and Their Applications

ZU Chenchen

A thesis submitted in partial fulfilment of the requirements

for the degree of doctor of philosophy

August 2021

CERTIFICATE OF ORIGINALITY

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

____________________(Signed)

___ZU Chenchen____(Name of student)

Dedicate to my parents.

Abstract

Nowadays, many optimization problems are presented on a large scale. Therefore, numerous techniques have been introduced to accelarate optimization algorithms. A practical one is the extrapolation strategy. In this thesis, we propose and explore two extrapolated first-order algorithms and then evaluate their efficiency by numerical experiments. We also construct sparse portfolio selection models and observe their theoretical and numerical characteristics.

The first part investigates an extrapolated inexact quasisubgradient method with diminishing, constant, and dynamic stepsizes for a quasiconvex minimization problem. The convergence in objective values and the iteration complexity of the method are established under a Hölder condition. With an additional assumption of weak sharp minima, a sublinear convergence rate for iterates is obtained for some special diminishing stepsize and extrapolation rule. For the constant and dynamic stepsizes, a linear rate of convergence to (a ball of) the optimal solution set is provided with a specially selected extrapolation step. In a similar way, we study a primal-dual extrapolated quasisubgradient method with the diminishing and constant stepsizes for finding a saddle point of a quasiconvex-quasiconcave function. The numerical testing shows that extrapolation improves the performance in terms of the number of iterations needed for reaching an approximate optimal solution.

In the second part, a penalty extrapolated alternating direction method of multipliers (ADMM) is proposed to solve a generalized bilinear programming problem. The algorithm is divided into inner and outer iterations. The inner iterations are constructed by using a proximal ADMM strategy with extrapolation for a quadratic penalty relaxation. The outer iterations are composed of updating the penalty parameter. The subsequential convergence and the iteration complexity $\mathcal{O}(1/k)$ are established for the inner extrapolated ADMM algorithm, and its global convergence is obtained by virtue of the Kurdyka-Łojasiewicz theory. Afterward, the convergence to the stationary point is also provided for the outer algorithm. In numerical testing, the efficiency of extrapolation is again demonstrated in the penalty ADMM algorithm. Besides, we compare the proposed algorithm with a semidefinite relaxation method. The resulting optimal values of both methods are close, but the proposed ADMM algorithm terminates within a shorter time.

In the final part, we study an l_p-sparse minimax model and an l_1-sparse minimax Sharpe ratio model and observe a descent property of the l_p norm of the optimal portfolio. A parametric algorithm is also designed for finding a global solution of the l_1-sparse minimax Sharpe ratio model. Numerically, we compare the three sparse minimax models with two sparse mean-variance models and test the effect of the regularization parameter on the sparsity, return, risk, and short selling by using the weekly historical data of 1200 stocks. We also apply the proposed penalty ADMM method to the l_1-sparse minimax Sharpe ratio model. As indicated in numerical experiments, more sparse portfolios of all the sparse minimax models tend to have lower rates of return and lower levels of risk. However, for the sparse mean-variance models, the corresponding changes are not so significant.

Acknowledgements

Carrying out research is always a non-isolated activity. Here, I express my sincere gratitude and regards to individuals who supported me in various aspects during my PhD study.

First of all, I would like to thank my supervisor, Professor Yang Xiaoqi, for his patience and encouragement. He helped me all the time with the research and writing of this thesis. This research would not be possible without his selfless support.

I would also like to express my gratitude to Dr. Hu Yaohua for his enlightening guidance and kind assistance in conducting the work of the extrapolated inexact quasisubgradient method.

Furthermore, I am very grateful to my friends and colleagues: Yao Wenfang, Carisa Kwok Wai Yu, Meng Kaiwen, Li Minghua, Dong Zhilong, Wu Yuqia, Sun Ying, Lin Qianying, Wang Qingzheng, and Hao Meiling.

Last but not least, my deepest gratitude and love, of course, belong to my parents, Chen Tong and Zu Xiaojun, for their unconditional love, company, encouragement, and support throughout my life.

Contents

List of Tables

List of Figures

List of Notations

$\mathbb{R}$	Set of real numbers.
$\mathbb{R}^n$	Set of n-dimensional real vectors.
$\mathbb{R}^{m\times n}$	Set of $m \times n$ real matrices.
$\mathbb{N}$	Set of nonnegative integers.
$\mathbf{S}$	Unit sphere centered at the origin.
$\mathbf{1}$	Vector of ones.
$\\|\bar{x}\\|$	Euclidean norm of the vector $\bar{x}$.
$\\|\bar{x}\\|_p$	l_p norm of the vector $\bar{x}$ $(0 < p \leq 1)$.
$\langle \bar{x}, \bar{y} \rangle$	Inner product of the vectors $\bar{x}$ and $\bar{y}$.
$\lceil a \rceil$	Ceiling function of the real number a.
$\liminf / \limsup$	Limit inferior/superior.
$\mathrm{dist}(\bar{x}, X)$	Euclidean distance between the vector $\bar{x}$ and the set X.
$P_X(\bar{x})$	Euclidean projection of the vector $\bar{x}$ onto the set X.
$N_X(\bar{x})$	Normal cone to the convex set X at the vector $\bar{x}$.
$\delta_X(\bar{x})$	Indicator function of the set X at the vector $\bar{x}$.
$\mathrm{dom} f$	Effective domain of the function f.
$\mathrm{lev}_{<c} f / \mathrm{lev}_{>c} f$	Strict level set of the function f.
$\mathrm{lev}_{\leq c} f / \mathrm{lev}_{\geq c} f$	Level set of the function f.
∇f	Gradient of the function f.

$\nabla_x f$	Partial gradient of the function f with respect to the variable x.
∂f	(Convex) subdifferential of the function f.
$\partial^F f$	Fréchet subdifferential of the function f.
$\partial^L f$	Limiting subdifferential of the function f.
$\bar{\partial}^* f$	Quasisubdifferential of the function f.
$\bar{\partial}^*_\epsilon f$	ϵ-Quasisubdifferential of the function f.

Chapter 1

Literature Review and Introduction

In this day and age, optimization problems arise in numerous fields such as engineering, economics, biology, and aerostatics. Due to the low requirement of storage, first-order optimization algorithms have attracted widespread attention. On the other hand, as many application problems are large-scale, various techniques have been proposed to design fast algorithms. A popular one is the extrapolation strategy. Therefore, we are interested in using the extrapolation step to speed up first-order methods. This thesis focuses on two extrapolated first-order algorithms for solving some structured nonconvex problems, particularly an extrapolated inexact quasisubgradient method for a quasiconvex programming problem and a penalty extrapolated proximal alternating direction method of multipliers (ADMM) for a generalized bilinear programming problem. In addition, several sparse portfolio selection models are also studied, where a critical descent property is satisfied for the l_p regularizer of the optimal portfolio.

This chapter will first present a literature review on the concerned issues and our motivations, then follow notations and definitions of the thesis and background knowledge of first-order optimization methods. The content of Chapters 2-4 will be briefly introduced afterward.

1.1 Extrapolation Strategy

The *extrapolation step*, also known as the *inertial force* in the literature, has been widely used in numerical optimization to accelerate the convergence rate and enhance numerical performances. In general, algorithms with extrapolation adopt the updating rules on an extrapolated point instead of the last iterate. For an optimization problem with x^k and x^{k-1} being the previous two iterates, an extrapolated point is given by $\hat{x}^k := x^k + \alpha^k(x^k - x^{k-1})$, where $\alpha^k \geq 0$ is the extrapolation parameter.

For some classes of nondifferentiable convex optimization problems, the iteration complexity is improved from $\mathcal{O}(1/k)$ to $\mathcal{O}(1/k^2)$ by using specially structured extrapolation rules in proximal gradient algorithms (see Beck & Teboulle (2009) and Nesterov (2013)) and a primal-dual algorithm (see Chambolle & Pock (2011)). Without the theoretically accelerated rate of convergence, more general schemes of the extrapolation step have been investigated for various types of convex optimization methods (see Alvarez (2004), Maingé & Merabet (2010), Boţ et al. (2015), Chen et al. (2015), Chambolle & Pock (2016), Johnstone & Moulin (2017), Wen et al. (2017), Alves et al. (2020), and Attouch & Cabot (2020)). In Boţ et al. (2015), Chen et al. (2015), Chambolle & Pock (2016), and Alves et al. (2020), better numerical results are obtained with the extrapolation strategy.

For nonconvex optimization problems, the convergence of different extrapolated algorithms has been established with an assumption of the Kurdyka-Łojasiewicz property (see Ochs et al. (2014), Pock & Sabach (2016), Alecsa et al. (2019), Wu & Li (2019), Jia et al. (2019), Zhang et al. (2019), Chao et al. (2020)). Moreover, Goudou & Munier (2009) and Maingé (2009) studied a proximal point method with extrapolation for quasiconvex optimizations in Hilbert space, where the weak convergence is satisfied. As indicated in numerical experiments conducted by Ochs et al. (2014), Pock & Sabach (2016), Wu & Li (2019), and Zhang et al. (2019), the extrapolation step enhances the performance of some nonconvex algorithms when the extrapolation parameter is appropriately selected.

To the best of our knowledge, no subgradient method with extrapolation has been studied for nondifferentiable convex or nonconvex problems in the literature. For ADMM-type methods, only Chen et al. (2015) constructed extrapolated proximal ADMM algorithms for a separable convex problem.

1.2 Quasiconvex Programming and Quasisubgradient Method

Quasiconvex programming, minimizing a quasiconvex function over a closed and convex set, appears in various areas, for example, engineering, economics, and decision science (see Avriel et al. (1988), Crouzeix et al. (1998), dos Santos Gromicho (1998), Hadjisavvas et al. (2005), and Ramík & Vlach (2012)). Many gradient-type methods have been investigated for continuously differentiable quasiconvex minimization problems. Kiwiel & Murty (1996) discussed the convergence property of the steepest descent method with Armijo's stepsize. Motivated by Kiwiel & Murty (1996), Quiroz et al. (2008) generalized the classical Armijo line search and constructed the steepest descent method for quasiconvex functions on Riemannian manifolds. Moreover, projected gradient methods with the convergence to a stationary point have been used to solve quasiconvex problems (see Cruz & Pérez (2010)) and quasiconvex multiobjective problems (see Cruz et al. (2011)). Proximal-method variants have also been adopted on the minimization problem of quasiconvex functions. Replacing the classical proximal distance with some second-order homogeneous distances, Attouch & Teboulle (2004) and Pan & Chen (2007) introduced proximal-like methods for quasiconvex programming. After that, Quiroz et al. (2008), Souza et al. (2010), and Langenberg & Tichatschke (2012) extended the proximal algorithm by Bregman distances, and Quiroz et al. (2015) proposed an inexact proximal method with an induced proximal distance for some classes of quasiconvex minimization problems. Furthermore, Goudou & Munier (2009) and Maingé (2009) studied a proximal point method with extrapolation for quasiconvex optimizations in Hilbert space, where the weak convergence is satisfied.

The subgradient method also plays an essential role in quasiconvex programming. Subgradient methods for solving nondifferentiable convex optimization problems began with the works of Ermol'ev (1966) and Polyak (1967) and were further developed by Shor (1985), Bertsekas et al. (2003), and Auslender & Teboulle (2004). Based on the Greenberg-Pierskalla subdifferential (Greenberg & Pierskalla, 1973), the quasi-subdifferential (see the definition in Subsection 1.5.2) was introduced to construct the quasisubgradient method for quasiconvex optimizations. For the minimization of a quasiconvex function f over a closed and convex set X (i.e., $\min_{x \in X} \; f(x)$), the standard quasisubgradient method is given by

$$x^{k+1} = P_X \left(x^k - v^k g(x^k) \right),$$

where $g(x^k)$ is a quasisubgradient of f at x^k, v^k is a stepsize, and $P_X(\cdot)$ denotes the Euclidean projection onto X.

Kiwiel (2001) explored convergence properties of quasisubgradient methods in Hilbert spaces with a diminishing stepsize, including an application to surrogate relaxation. Konnov (2003) studied an inexact quasisubgradient method with exact and inexact dynamic stepsizes for quasiconvex optimization problems. Inspired by Polyak (1978) and Nedić & Bertsekas (2010), Hu et al. (2015) added noise in an inexact quasisubgradient to establish an approximate quasisubgradient method with diminishing and constant stepsizes and presented convergence results in both objective values and iterates and the finite convergence to approximate optimality. Afterward, Hu et al. (2020) provided a unified convergence analysis for a sequence satisfying a basic inequality and applied it to various types of quasisubgradient methods, which includes an abstract convergence theorem in Yu et al. (2019) as a special case. Furthermore, Hishinuma & Iiduka (2020) proposed a quasisubgradient method for quasiconvex problems with respect to a fixed point set and showed its numerical superiority to some existing algorithms. In Kiwiel (2001), Konnov (2003), and Hu et al. (2020), a sublinear convergence in objective values is obtained under a Hölder condition, and a sublinear or linear convergence for iterates is obtained under both Hölder and weak sharp minima conditions.

1.3 Generalized Bilinear Programming and ADMM

A function with two variables is said to be *bilinear* if it is linear in one variable when the other one is fixed. More precisely, a function $f(x, y)$ is bilinear if it is linear in x for a fixed y and linear in y for a fixed x. *Bilinear programming* (BLP), minimizing a bilinear function subject to linear constraints, emerges in numerous application problems such as the assignment problem, game theory, and location allocation (see Konno (1971), Reklaitis et al. (1983), and Papalambros & Wilde (2000)). Recently, industrial applications of the BLP problem have also drawn widespread attention, e.g., the nonnegative matrix factorization and nonnegative matrix factorization completion (see Pauca et al. (2006), Xu et al. (2012), and Hajinezhad et al. (2016)).

Based on bilinear programming, Al-Khayyal (1992) discussed a more complicated bilinear optimization problem with bilinear constraints and called it the *generalized bilinear programming* (GBLP) problem. The pooling problem in Foulds et al. (1992), Audet et al. (2004), and Erbeyoğlu & Bilge (2016), farm management problem in Bloemhof-Ruwaard & Hendrix (1996), and global supply chain problem in Vidal & Goetschalckx (2001) provided specializations of this class of problems. Moreover, all quadratically constrained quadratic programming (QCQP) problems (see Luo et al. (2010) and Anstreicher (2012)) and all linear sum-of-ratios problems (see Benson (2007) and Jiao & Liu (2015)) can be reformulated as a GBLP problem (see Subsections 3.6.2 and 3.6.3).

Research on global algorithms for GBLP problems is abundant, and the most has centered on branch-and-bound methods with different relaxation strategies. Sherali & Alameddine (1992), Liberti & Pantelides (2006), and Sherali & Adams (2013) used the reformulation linearization technique to extend branch-and-bound algorithms. Al-Khayyal (1992), Tawarmalani et al. (2010), and Fukuda & Kojima (2001) provided some tight convex approximations to reformulate subproblems in the branch-and-bound scheme. Another common relaxation technique is the Lagrangian relaxation (see Ben-Tal et al. (1994) and Almutairi & Elhedhli (2009)). Besides branch-and-bound methods, Konno & Kuno (1992) and Kuno et al. (1992) constructed a parametric algorithm by solving a master problem for a class of GBLP problems. Floudas & Aggarwal (1990) and Osman & Demirli (2010) applied generalized Benders decompositions to some specific GBLP problems. Moreover, convex transformations have been composed to approximate a bilinear integer nonlinear programming problem (see Harjunkoski et al. (1997) and Harjunkoski et al. (1998)), where any GBLP problem is included as a specialization.

In terms of the local method, alternating algorithms have been used to solve some linearly constrained nonconvex problems, accommodating BLP problems as instances. The major types are the block coordinate descent (BCD) method and the alternating direction method of multipliers (ADMM) (see Tseng (1993) and Tseng (2001) for BCD; see Xu et al. (2012) for ADMM). For problems with nonconvex constraints, BCD or ADMM methods with the convergence to a Nash point or a stationary point were provided by Xu & Yin (2013) and Hajinezhad & Shi (2018). In Xu & Yin (2013), three BCD variants were proposed to solve a multi-convex optimization problem, which covers the GBLP structure. However, those algorithms are limited to the feasibility of subproblems, which is not guaranteed for many generalized bilinear problems. In Hajinezhad & Shi (2018), an ADMM algorithm was adopted on a structured bilinear problem, where the framework is related to GBLP.

Inspired by the mentioned literature, we are interested in constructing an ADMM-type method for GBLP problems. The *alternating direction method of multipliers* (ADMM) was introduced by Gabay & Mercier (1976) to solve the following two-block separable convex optimization problem

$$\begin{aligned} \min_{x,z} \quad & f(x) + h(z) \\ \text{s.t.} \quad & Ax + Bz = c, \end{aligned} \tag{1.1}$$

where f and h are proper, closed, and convex functions. The *augmented Lagrangian function* of (1.1) is written as

$$\mathcal{L}_\rho(x, z, \mu) := f(x) + h(z) + \mu^{\mathrm{T}}(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|^2,$$

where μ is the multiplier and $\rho > 0$ is the (augmented) penalty parameter of $\mathcal{L}_\rho$. The standard ADMM for (1.1) is given by

$$\begin{aligned} x^{k+1} &= \arg\min_x \mathcal{L}_\rho(x, z^k, \mu^k), \\ z^{k+1} &= \arg\min_z \mathcal{L}_\rho(x^{k+1}, z, \mu^k), \\ \mu^{k+1} &= \mu^k + \rho(Ax^{k+1} + Bz^{k+1} - c). \end{aligned} \tag{1.2}$$

Eckstein & Bertsekas (1992) and Fukushima (1992) provided the convergence of (1.2), and He & Yuan (2012, 2015) obtained its iteration complexity $\mathcal{O}(1/k)$. Apart from the standard scheme, modified ADMM methods have also been studied in He et al. (2002), He et al. (2006), and Chen et al. (2015). From Chen et al. (2016), the straight extension of (1.2) may diverge when (1.1) is extended to a multi-block structure. However, under mild conditions or with revised subproblems, He et al.

(2012), Chang et al. (2014), and Shen & Pan (2015) established the convergence of ADMM algorithms for multi-block separable convex problems. With relatively restrictive assumptions, a linear convergence or a sublinear convergence with the rate $o(1/k)$ can be obtained for some multi-block ADMM method (see Hong & Luo (2017) and Deng et al. (2017)).

For nonconvex problems, ADMM variants with the convergence to a stationary point have been adopted on specific application problems, such as the matrix completion problem in Xu et al. (2012) and the sharing and consensus problem in Hong et al. (2016). Li & Pong (2015), Yang et al. (2017), and Hajinezhad & Shi (2018) proposed ADMM-type algorithms for some classes of nonconvex and nonsmooth problems and conducted numerical experiments to evaluate the effectiveness of their methods. Furthermore, the ADMM scheme has also been applied to more general nonconvex problems with novel techniques and weaker assumptions (see Wang et al. (2019) and Jiang et al. (2019)). The subsequential convergence of all those algorithms is established with a sufficiently large associated penalty parameter, and their global convergence is obtained by virtue of the Kurdyka-Łojasiewicz theory (see Attouch et al. (2010)).

1.4 Sparse Portfolio Selection Models

In 1952, Markowitz (1952) formulated the portfolio selection problem as the mean-variance model, which has since then become a milestone in portfolio selection and remains a dominant technique in use today (see Das et al. (2011) and Markowitz & Van Dijk (2003)). In the mean-variance framework, two critical elements, return and risk, are defined by the expected return and variance of a portfolio, respectively. The expression of return is straightforward, but the measure of risk is not definite. In fact, various risk measures have been proposed to replace the portfolio variance and establish alternative portfolio selection rules, typically the ones with a linear structure. Sharpe (1967) measured the risk by market responsiveness and composed a linear approximation of the mean-variance model. Moreover, some new linear risk measures such as a mean absolute deviation (Konno & Yamazaki, 1991), a minimax risk measure (Young, 1998), and an l_∞ risk function (Cai et al., 2000) have been demonstrated to be competitive in empirical studies.

In modern society, portfolios including many securities are not desirable, especially for large-scale investments or retail investors. Therefore, finding sparse optimal portfolios becomes an essential issue in portfolio selection. Many selection criteria have been adopted to seek sparse portfolios based on the mean-variance model. By considering nonnegativity constrained portfolios, Jagannathan & Ma (2003) obtained an optimal portfolio consisting of only around 24 stocks for a 500-stock universe. Qi et al. (2019) applied their portfolio model to 1800-stock problems, and the minimum number of the selected stocks can be 62 on average. Woodside-Oriakhi et al. (2011) provided a series of heuristic algorithms for a cardinality constrained mean-variance model, which generate an efficient frontier with the number of stocks included fixed.

Furthermore, the regularization method is also a promising method for pursuing sparse portfolios in the mean-variance framework. For this method, an l_p norm or l_p regularizer ($0 < p \leq 1$) is added in the objective function or constraint set to modify the original model. The regularization technique has been widely applied in the industry. In particular, the l_1 regularizer has been used to seek sparsity for problems of image reconstruction, data analysis, machine learning, and so on (see Tibshirani (1996), Daubechies et al. (2004), and Beck & Teboulle (2009)). Furthermore, it has been shown by Chartrand (2007) and Saab et al. (2008) that the method with l_p ($0 < p < 1$) norm rather than l_1 norm produces more sparse solutions for some industrial applications, although the computation of l_p-sparse formulations is more complicated. In respect of the portfolio selection, some l_1-norm mean-variance models have been illustrated to be effective for promoting the sparsity of portfolios (see Brodie et al. (2009), DeMiguel et al. (2009), and Dai & Wen (2018)). When $0 < p < 1$, Chen et al. (2013) and Fastrich et al. (2015) claimed that sparse optimal portfolios of l_p-regularized mean-variance models have satisfactory numerical performances. In fact, research on the sparse portfolio selection has been centered on the mean-variance model, but few studies have focused on linear portfolio models. Therefore, we take a linear minimax model (Young, 1998) as an example to investigate l_p-sparse $(0 < p \leq 1)$ linear portfolio models (see Subsection 4.3.1).

In addition to classical portfolio selection frameworks, we are also curious about a sparse minimax model based on the Sharpe ratio. In performance assessment, the Treynor index (Treynor, 1965), Sharpe index (Sharpe, 1966), and Jensen index (Jensen, 1968) are the three popular performance measures to rank the performance of portfolios. The Sharpe index, also known as the Sharpe ratio, was first put forward by Sharpe (1966) as an extension of the Treynor index. The classical Sharpe ratio is a quotient of an expected excess return (numerator) and risk (denominator). The risk in the Sharpe ratio is defined by the volatility or standard deviation of the portfolio under consideration. By virtue of the quotient structure, different Sharpe-type ratios were composed, where the volatility is replaced by alternative risk measures, such as the Martin ratio (Martin & McCann, 1989), Sortino ratio (Sortino & Van Der Meer, 1991), and Sterling ratio (McCafferty, 2002). Apart from ranking performances, the Sharpe ratio is also employed as the objective function to construct portfolio selection models (see Benninga & Czaczkes (2014) and Elton et al. (2014)). With generalized Sharpe ratios, the Sharpe ratio maximization model can be extended. In this way, we establish a generalized Sharpe ratio model by the minimax risk measure in Young (1998) and study its l_1-sparse formulation in Subsection 4.3.2.

1.5 Preliminaries

In this section, we present notations and definitions of the thesis and then introduce several essential subdifferentials and first-order conditions of different optimization problems, which will be frequently mentioned in our convergence analysis.

1.5.1 Notations and Definitions

Throughout the thesis, notations and definitions are standard. We restrict ourselves to a Euclidean space. Let $\mathbb{R}$ be the set of real numbers, $\mathbb{R}^n$ be the set of n-dimensional real vectors, $\mathbb{R}^{m\times n}$ be the set of $m \times n$ real matrices, and $\mathbb{N}$ be the set of nonnegative integers. Notations $\|\cdot\|$, $\langle \cdot, \cdot \rangle$, and $\lceil \cdot \rceil$ are used to represent the Euclidean norm of a vector, the inner product of two vectors, and the ceiling function of a real number, respectively. The limit inferior (resp. limit superior) of a sequence is denoted by $\liminf$ (resp. $\limsup$).

For a vector $\bar{x} \in \mathbb{R}^n$ and a closed and convex set $X \subseteq \mathbb{R}^n$, $\operatorname{dist}(\bar{x}, X)$ denotes the *Euclidean distance* between $\bar{x}$ and X, i.e.,

$$\operatorname{dist}(\bar{x}, X) := \min_{x\in X} \|x - \bar{x}\|\,;$$

$P_X(\bar{x})$ denotes the *Euclidean projection* of $\bar{x}$ onto X, i.e.,

$$P_X(\bar{x}) := \operatorname*{arg\,min}_{x\in X} \|x - \bar{x}\|\,;$$

$N_X(\bar{x})$ denotes the *normal cone* to X at $\bar{x}$, i.e.,

$$N_X(\bar{x}) := \{v \in \mathbb{R}^n : \langle v, x - \bar{x}\rangle \leq 0,\ \forall x \in X\};$$

$\delta_X(\bar{x})$ denotes the *indicator function* of X at $\bar{x}$, i.e.,

$$\delta_X(\bar{x}) := \begin{cases} 0, & \text{if } \bar{x} \in X, \\ +\infty, & \text{if } \bar{x} \notin X. \end{cases}$$

For a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the *effective domain* of f is defined by

$$\mathrm{dom} f := \{x \in \mathbb{R}^n : f(x) < +\infty\}.$$

f is said to be *proper* if there exists $x \in \mathbb{R}^n$ such that $f(x) < +\infty$, or equivalently, $\mathrm{dom} f \neq \emptyset$. f is said to be *lower semi-continuous* on $\mathbb{R}^n$ if

$$\liminf_{y \to x} f(y) = f(x) \quad \text{for all} \quad x \in \mathbb{R}^n.$$

f is said to be *Lipschitz continuous* on $D \subseteq \mathbb{R}^n$ if there exists $L > 0$ such that

$$|f(x) - f(y)| \leq L \, \|x - y\| \quad \text{for all} \quad x, y \in D.$$

f is said to be *convex* if $\mathrm{dom} f$ is a convex set and

$$f(tx + (1-t)y) \leq t f(x) + (1-t) f(y) \quad \text{for all} \quad x, y \in \mathrm{dom} f \text{ and } t \in [0,1].$$

f is said to be *concave* if $-f$ is convex. Moreover, if $f - \frac{\gamma}{2} \|x\|^2$ is convex for some $\gamma > 0$, we say that f is *strongly convex* with modulus γ. When f is continuously differentiable, an equivalent definition of f to be strongly convex with γ is

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\gamma}{2} \|x - y\| \quad \text{for all} \quad x, y \in \mathrm{dom} f.$$

f is said to be *quasiconvex* if $\mathrm{dom} f$ is a convex set and

$$f(tx + (1-t)y) \leq \max\{f(x), f(y)\} \quad \text{for all} \quad x, y \in \mathrm{dom} f \text{ and } t \in [0,1].$$

f is said to be *quasiconcave* if $-f$ is quasiconvex. For each $c \in \mathbb{R}$, the *(strict) level sets* of f are written as

$$\begin{aligned} &\mathrm{lev}_{<c} f := \{x \in \mathbb{R}^n : f(x) < c\}, \ \mathrm{lev}_{>c} f := \{x \in \mathbb{R}^n : f(x) > c\}, \\ &\mathrm{lev}_{\leq c} f := \{x \in \mathbb{R}^n : f(x) \leq c\}, \ \mathrm{lev}_{\geq c} f := \{x \in \mathbb{R}^n : f(x) \geq c\}. \end{aligned}$$

1.5.2 Subdifferentials and First-order Conditions

We consider an unconstrained optimization problem

$$\min_{x} \ f(x), \tag{1.3}$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper function.

Convex Optimization Problem

Suppose that f in (1.3) is convex. The *(convex) subdifferential* of f at $\bar{x}$ is given by

$$\partial f(\bar{x}) := \{g \in \mathbb{R}^n : \langle g, x - \bar{x}\rangle \leq f(x) - f(\bar{x}), \ \forall x \in \mathrm{dom} f\}.$$

If f is smooth at $\bar{x}$, the gradient is the only element included, i.e., $\partial f(\bar{x}) = \{\nabla f(\bar{x})\}$.

The stationary point associated with the subdifferential is defined as follows.

Definition 1.5.1. *Suppose that f in (1.3) is convex. We say that $\bar{x}$ is a stationary point of (1.3) if $0 \in \partial f(\bar{x})$.*

Then, the first-order optimality condition of (1.3) can be expressed by the stationary point (see Rockafellar (1970, Section 27)).

Proposition 1.5.2. *Suppose that f in (1.3) is convex. Then $\bar{x}$ is an optimal solution of (1.3) if and only if $\bar{x}$ is a stationary point of (1.3).*

Nonconvex Optimization Problem

Suppose that f in (1.3) is possibly nonconvex. When f is smooth, the stationary point is standardly defined by the gradient. Specifically, we say that $\bar{x}$ is a stationary point of (1.3) if $\nabla f(\bar{x}) = 0$. This point is either a local extremum (maximum or maximum) or a saddle point of the problem. When f is nonsmooth, unfortunately, directly using the (convex) subdifferential may lose critical information. For example, the subdifferential of $f(x) = -|x|$ is empty at the origin, which is in fact the global maximum of the function. Therefore, numerous generalized subdifferentials have been proposed to investigate nonconvex and nonsmooth optimization problems (see Aussel et al. (1995), Rockafellar & Wets (2009), and Mordukhovich (2006)). The most popular are the Fréchet or regular subdifferential (Bazaraa et al., 1974) and limiting (Fréchet) subdifferential (Mordukhovich, 1976).

- The *Fréchet or regular subdifferential* of f at $\bar{x}$ is given by

$$\partial^F f(\bar{x}) := \left\{ g \in \mathbb{R}^n : \liminf_{x \neq \bar{x}, x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \langle g, x - \bar{x} \rangle}{\|x - \bar{x}\|} \geq 0 \right\};$$

- The *limiting (Fréchet) subdifferential* of f at $\bar{x}$ is given by

$$\partial^L f(\bar{x}) := \left\{ g \in \mathbb{R}^n : \exists\ x^k \xrightarrow{f} \bar{x} \text{ and } g^k \to g \text{ with } g^k \in \partial^F f(x^k) \text{ as } k \to +\infty \right\},$$

  which is the limit version of the Fréchet subdifferential.

Similar to ∂f, it holds that $\partial^F f(\bar{x}) = \partial^L f(\bar{x}) = \{\nabla f(\bar{x})\}$ if f is smooth at $\bar{x}$. When f is convex, both subdifferentials reduce to the (convex) subdifferential, i.e., $\partial^F f(\bar{x}) = \partial^L f(\bar{x}) = \partial f(\bar{x})$.

Respective stationary points for the above subdifferentials are defined as follows.

Definition 1.5.3. *We say that $\bar{x}$ is a Fréchet or directional (resp. limiting) stationary point of* (1.3) *if $0 \in \partial^F f(\bar{x})$ (resp. $0 \in \partial^L f(\bar{x})$).*

Referring to Rockafellar (1985), Van Ngai et al. (2002), Kruger (2003), Mordukhovich (2006, Chapters 1 and 3), and Rockafellar & Wets (2009, Chapters 8 and 10), we summarize some useful properties about these generalized subdifferentials and stationary points.

Proposition 1.5.4. *For* (1.3)*, the following statements hold.*

(1). (Relationship). *For any* $\bar{x} \in \mathrm{dom} f$*, one has that* $\partial^F f(\bar{x}) \subseteq \partial^L f(\bar{x})$*.*

(2). (First-order condition). *If* $\bar{x}$ *is a local minimum of* (1.3)*, then* $\bar{x}$ *is a Fréchet or directional stationary point and also a limiting stationary point of* (1.3)*.*

(3). (Addition with a smooth function). *Let* $f(x) := f_1(x) + f_2(x)$*. Then, for any* $\bar{x} \in \mathrm{dom} f_1$ *with* f_2 *being smooth on a neighborhood of* $\bar{x}$*, one has that* $\partial^F f(\bar{x}) = \partial^F f_1(\bar{x}) + \nabla f_2(\bar{x})$ *and* $\partial^L f(\bar{x}) = \partial^L f_1(\bar{x}) + \nabla f_2(\bar{x})$*.*

(4). (Addition of separable functions). *Let* $f(x) := f_1(x_1) + f_2(x_2)$*, where* f_1 *and* f_2 *are lower semi-continuous functions, and* $x := (x_1, x_2)$*. Then, for any* $\bar{x} := (\bar{x}_1, \bar{x}_2) \in \mathrm{dom} f$ *with* $\liminf\limits_{t \searrow 0, d \to 0} \frac{f_i(\bar{x}_i + td) - f_i(\bar{x})}{t} = 0$*,* $i = 1, 2$*, one has that* $\partial^F f(\bar{x}) = \partial^F f_1(\bar{x}_1) \times \partial^F f_2(\bar{x}_2)$ *and* $\partial^L f(\bar{x}) = \partial^L f_1(\bar{x}_1) \times \partial^L f_2(\bar{x}_2)$*.*

Quasiconvex Optimization Problem

Suppose that f in (1.3) is quasiconvex. Initially, all the subdifferentials for nonconvex and nonsmooth functions can be used to discuss quasiconvex optimization problems (see Pan & Chen (2007), Langenberg & Tichatschke (2012), and Quiroz et al. (2015)). On the other hand, some subdifferentials related to quasiconvexity have also been investigated (see Greenberg & Pierskalla (1973), Plastria (1985), Martínez-Legaz & Sach (1999), and Daniilidis et al. (2001)). In this thesis, we focus on the quasisubdifferential (see Kiwiel (2001)) and its inexact version (see Konnov (2003) and Hu et al. (2015)).

Let $\epsilon \geq 0$. Then the *quasisubdifferential* and ϵ*-quasisubdifferential* of f at $\bar{x}$ are respectively given by

$$\bar{\partial}^* f(\bar{x}) := \left\{ g \in \mathbb{R}^n : \langle g, x - \bar{x} \rangle \leq 0, \ \forall x \in \mathrm{lev}_{<f(\bar{x})} f \right\}$$

and

$$\bar{\partial}^*_\epsilon f(\bar{x}) := \left\{ g \in \mathbb{R}^n : \langle g, x - \bar{x} \rangle \leq 0, \ \forall x \in \mathrm{lev}_{<f(\bar{x}) - \epsilon} f \right\}.$$

Another equivalent definition is the normal cone to a level set; in particular,

$$\bar{\partial}^* f(\bar{x}) := N_{\mathrm{lev}_{<f(\bar{x})} f}(\bar{x}) \quad \text{and} \quad \bar{\partial}^*_\epsilon f(\bar{x}) := N_{\mathrm{lev}_{<f(\bar{x})-\epsilon} f}(\bar{x}).$$

It is clear that $\bar{\partial}^* f(\bar{x}) = \bar{\partial}^*_{\epsilon=0} f(\bar{x})$ and $\bar{\partial}^*_\epsilon f(\bar{x})$ is a closed and convex cone. When f is convex, one has that $\bar{\partial}^* f(\bar{x}) = \{\lambda g : g \in \partial f(\bar{x}),\ \lambda \geq 0\}$.

Moreover, for a proper quasiconcave function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the *quasisubdifferential* and *ϵ-quasisubdifferential* at $\bar{x}$ are respectively given by

$$\bar{\partial}^* h(\bar{x}) := \left\{g \in \mathbb{R}^n : \langle g, x - \bar{x}\rangle \geq 0,\ \forall x \in \mathrm{lev}_{>h(\bar{x})} h\right\}$$

and

$$\bar{\partial}^*_\epsilon h(\bar{x}) := \left\{g \in \mathbb{R}^n : \langle g, x - \bar{x}\rangle \geq 0,\ \forall x \in \mathrm{lev}_{>h(\bar{x})+\epsilon} h\right\}.$$

Constrained Optimization Problem

Now, we consider a constrained optimization problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & x \in X, \end{aligned} \tag{1.4}$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper function, and $X \subseteq \mathbb{R}^n$ is a nonempty, closed, and convex set. Equivalently, (1.4) can be translated into an unconstrained form

$$\min_x \ f(x) + \delta_X(x).$$

From definitions of the normal cone and (convex) subdifferential, we have that $N_X = \partial \delta_X$. Thus, the first-order condition of (1.4) can be expressed by the normal cone to X. We give an example for f being smooth.

Proposition 1.5.5. *Suppose that f in (1.4) is smooth. If $\bar{x}$ is a local minimum of (1.4), then $-\nabla f(\bar{x}) \in N_X(\bar{x})$, or equivalently,*

$$\langle \nabla f(\bar{x}), x - \bar{x}\rangle \geq 0 \quad \textit{for all} \quad x \in X.$$

Furthermore, if f is a smooth convex function, the above condition is sufficient and necessary for $\bar{x}$ to be an optimal solution of (1.4).

1.6 Organization of the Thesis

The rest of the thesis is organized as follows.

In Chapter 2, we generalize the inexact quasisubgradient method in Hu et al. (2015) by the extrapolation strategy for a quasiconvex optimization problem. The convergence properties of the method are explored with diminishing, constant and dynamic stepsizes. The convergence in objective values and the iteration complexity are established under a Hölder condition, where the diminishing stepsize and constant stepsize are treated in a unified way as both have similar structures. Furthermore, the rate of convergence for all the stepsizes is analyzed under both Hölder and weak sharp minima conditions. We also consider the extrapolated inexact quasisubgradient method in a primal-dual framework for solving a saddle point problem of a quasiconvex-quasiconcave function afterward. The primal-dual method is investigated with the diminishing and constant stepsizes in respect of the convergence property in objective values and the iteration complexity. Eventually, numerical experiments are carried out to evaluate the effectiveness of the extrapolation step.

In Chapter 3, we introduce a penalty extrapolated ADMM method to solve a GBLP problem. The algorithm contains inner and outer iterations. The inner algorithm is a proximal ADMM method with extrapolation for a quadratic penalty relaxation of the GBLP problem. Its subsequential convergence and iteration complexity are explored using a basic descent property, and the global convergence is discussed on the basis of the Kurdyka-Łojasiewicz theory. The outer algorithm is an update of the penalty parameter associated with the inner method, and the outer convergence to a stationary point is also investigated. In numerical testing, the effect of extrapolation parameters is examined for the inner penalty problem and a linear sum-of-ratios problem. Apart from it, we also compare the proposed algorithm with a semidefinite relaxation method for a specially structured QCQP problem.

In Chapter 4, we take the minimax risk measure in Young (1998) as the representative of linear risk measures to construct sparse linear portfolio models, specifically, an l_p-sparse ($0 < p \leq 1$) minimax model and an l_1-sparse minimax Sharpe ratio model. To find a global solution of the second model, we design a parametric method based on Konno & Kuno (1990). Additionally, since the second model can be translated into the GBLP structure, we also specialize the algorithm in Chapter 3 to it in numerical studies. In the experiments, we test the influence of model parameters and observe characteristics of the sparse minimax models, where the equal-weighted rule and l_p-sparse mean-variance models are selected as benchmarks.

In Chapter 5, we summarize the main results in this thesis.

Chapter 2

Inexact Quasisubgradient Methods with Extrapolation

2.1 Introduction

In this chapter, we consider the following quasiconvex programming problem

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.t.} \quad & x \in X, \end{aligned} \tag{2.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a quasiconvex function, and $X \subseteq \mathbb{R}^n$ is a nonempty, bounded, closed, and convex set.

For solving (2.1), we extend an inexact quasisubgradient method (see Hu et al. (2015)) by using the extrapolation strategy. For the proposed algorithm, each iteration is an inexact quasisubgradient step based on a combination of the previous two iterates. To the best of our knowledge, no subgradient method with extrapolation has been studied for nondifferential convex or nonconvex optimization problems in the literature. To obtain a critical basic inequality, we need to assume that the constraint set X is bounded. The same boundedness assumption is also found in Kiwiel (2001), Nedić & Bertsekas (2010), and Hu et al. (2015). The convergence property is explored with diminishing, constant, and dynamic stepsizes, where the third stepsize is absent in Hu et al. (2015). Under a Hölder condition of order p, we derive an extended basic inequality, a dominant tool in our analysis, as it is the case for all subgradient-type methods.

We establish the convergence in objective values, iteration complexity, and convergence rate for iterates for the proposed method. In the analysis of the convergence in objective values and the iteration complexity, we treat the diminishing stepsize and constant stepsize in a unified way as both have similar structures. When studying the iteration complexity and rate of convergence for iterates, we make use of some specially structured diminishing stepsize and extrapolation rule. When both the diminishing stepsize and extrapolation rule are decaying as a power function, we obtain explicit iteration complexities. Moreover, an additional assumption of weak sharp minima of order q is made to investigate the convergence rate for iterates. When the diminishing stepsize is decaying as a power function and the extrapolation rule is decreasing not less than a power function, the method provides a sublinear convergence rate $\mathcal{O}\left(\tau^{k^s}\right)$ (for some $0 < s < 1$ and $0 < \tau < 1$) to the optimal solution set of (2.1) or a tolerance region of the optimal solution set, which is faster than $\mathcal{O}\left(1/k^h\right)$ (for each $h > 0$). With a geometrically decreasing extrapolation step, we obtain a linear rate of convergence (in particular, $\mathcal{O}\left(\tau^k\right)$ for some $0 < \tau < 1$) to (a tolerance region of) the optimal solution set for the constant and dynamic stepsizes. Our convergence results include the relevant ones in Kiwiel (2001), Konnov (2003), Hu et al. (2015), Hu et al. (2020), and references therein as special cases. Motivated by Hu et al. (2016), we also study a primal-dual extrapolated quasisubgradient method with diminishing and constant stepsizes for finding a saddle point of a quasiconvex-quasiconcave function. The convergence in objective values and the iteration complexity of the primal-dual method are obtained by virtue of the similar analysis to the (primal) extrapolated inexact quasisubgradient method. Eventually, we test the effect of the extrapolation step and find that the quasisubgradient algorithm with extrapolation is more efficient than that without extrapolation in terms of the number of iterations needed for reaching an approximate optimal solution.

The rest of the chapter is organized as follows. We propose an extrapolated inexact quasisubgradient method and provide some preliminary properties in Section 2.2. In Sections 2.3 and 2.4, we establish the general convergence property in objective values and study the iteration complexity of the method, respectively, then follow the convergence rate for iterates in Section 2.5. Afterward, in Section 2.6, a primal-dual extrapolated inexact quasisubgradient method is introduced and explored in terms of the convergence in objective values and the iteration complexity. Eventually, we present and discuss numerical results in Section 2.7.

2.2 Inexact Quasisubgradient Method and Basic Properties

Let $\epsilon \geq 0$ be the inexactness of the quasisubdifferential, $\left\{r^k\right\}$ be the sequence of noise, $\left\{v^k\right\}$ be the sequence of stepsizes, and $\left\{\alpha^k\right\}$ be the sequence of extrapolation parameters. We use $\mathbf{S}$ to denote the unit sphere centered at the origin. Besides, f^* and X^* represent the optimal value and optimal solution set of (2.1), respectively, and X^*_ϵ represents the ϵ-approximate solution set of (2.1), that is, $X^*_\epsilon := \{x \in X : f(x) \leq f^* + \epsilon\}$. As X is bounded, then there exists $d > 0$ such that $\|x\| \leq d$ for all $x \in X$. Recall that the ϵ-quasisubdifferential of f (see Subsection 1.5.2) is defined by

$$\bar{\partial}^*_\epsilon f(\bar{x}) := \left\{g \in \mathbb{R}^n : \langle g, x - \bar{x}\rangle \leq 0,\ \forall x \in \mathrm{lev}_{<f(\bar{x})-\epsilon} f\right\}.$$

Now, we propose the following extrapolated inexact quasisubgradient method (EiQSG) for (2.1), which extends the method in Hu et al. (2015) by extrapolation,

$$\begin{aligned} \hat{x}^k &= x^k + \alpha^k(x^k - x^{k-1}), \\ x^{k+1} &= P_X\left(\hat{x}^k - v^k \tilde{g}(x^k)\right), \end{aligned} \tag{2.2}$$

where $x^0 \in \mathbb{R}^n$ and $x^{-1} = x^0$ are initial points. As the projection is adopted, the function value $f(x^k)$ does not decrease monotonically. Moreover,

$$\tilde{g}(x^k) = g(x^k) + r^k \qquad \text{and} \qquad g(x^k) \in \bar{\partial}^*_\epsilon f(x^k) \cap \mathbf{S}.$$

In (2.2), the extrapolation parameter sequence $\{\alpha^k\}$ is assumed to satisfy

$$\alpha^k \geq 0 \qquad \text{and} \qquad \sum_{k=0}^{+\infty} \alpha^k < +\infty,$$

which implies that $\lim_{k\to+\infty} \alpha^k = 0$ and $\{\alpha^k\}$ is bounded. When $\alpha^k \equiv 0$, (2.2) reduces to the method in Hu et al. (2015). Without loss of generality, we assume that $\alpha^k \leq 1$ for all $k \in \mathbb{N}$. We also assume that the noise sequence $\{r^k\}$ is bounded, then there exists $R \geq 0$ such that $\|r^k\| \leq R$ for all $k \in \mathbb{N}$. For $\{v^k\}$, we consider the following stepsizes:

- Diminishing stepsize: $v^k > 0$, $\lim_{k\to+\infty} v^k = 0$, and $\sum_{k=0}^{+\infty} v^k = +\infty$;
- Constant stepsize: $v^k \equiv v > 0$;
- Dynamic stepsize:

$$v^k = \gamma^k \left(\frac{f(x^k) - f^* - \epsilon}{L} \right)^{\frac{1}{p}} \quad \text{with} \quad 0 < \underline{\gamma} \leq \gamma^k \leq \bar{\gamma} < \frac{2}{(1+R)^2} \ ,$$

  where p and L are the order and modulus in Assumption 2.2.1, respectively. For this stepsize, the method terminates at the k^{th} iteration if $x^k \in X^*_\epsilon$.

The diminishing stepsize and constant stepsize have similar structures and can be unified as

$$v^k > 0, \quad \lim_{k\to+\infty} v^k = v \geq 0, \quad \text{and} \quad \sum_{k=0}^{+\infty} v^k = +\infty. \tag{2.3}$$

When $v = 0$, it is the diminishing stepsize. When $v^k \equiv v > 0$, it is the constant stepsize. This unification simplifies our presentation in Section 2.3. In Section 2.4, we focus on a specially structured diminishing stepsize $v^k = ck^{-s}$, where $c > 0$ and $0 < s < 1$. For this case, the diminishing and constant stepsizes can be unified as

$$v^k = v + ck^{-s} > 0 \qquad \text{with} \quad v \geq 0,\ c \geq 0,\ \text{and } 0 < s < 1. \tag{2.4}$$

When a Hölder condition is satisfied, the dynamic stepsize sequence $\{v^k\}$ is bounded (see (2.5)). Then there exists $\bar{v} > 0$ such that $v^k \leq \bar{v}$ for all $k \in \mathbb{N}$. This upper bound is helpful in convergence analysis for the dynamic stepsize.

The Hölder condition of order p has been used in the literature to study the convergence of subgradient or quasisubgradient methods, as it is a key condition to establish some properties of (quasi)subgradients and obtain a basic inequality. This condition is assumed throughout Sections 2.2-2.5.

Assumption 2.2.1. *Assume that f in (2.1) satisfies the Hölder condition restricted to X^* of order $p > 0$ with modulus $L > 0$ on $\mathbb{R}^n$, i.e.,*

$$f(x) - f^* \leq L\mathrm{dist}^p(x, X^*) \quad \textit{for all} \quad x \in \mathbb{R}^n.$$

Suppose that Assumption 2.2.1 holds. Then the sequence of dynamic stepsizes $\{v^k\}$ is bounded. Indeed, for any k satisfying $x^k \notin X^*_\epsilon$, we have that

$$v^k = \gamma^k \left(\frac{f(x^k) - f^* - \epsilon}{L}\right)^{\frac{1}{p}} \leq \bar{\gamma}\left(\frac{L\mathrm{dist}^p(x, X^*) - \epsilon}{L}\right)^{\frac{1}{p}} \leq \bar{\gamma}\left(\frac{L(2d)^p - \epsilon}{L}\right)^{\frac{1}{p}}. \tag{2.5}$$

Lemma 2.2.2 (Hu et al. (2015, Lemma 3.3)) and Lemma 2.2.3 (Konnov (2003, Proposition 2.1)) provide the relations between an ϵ-quasisubgradient and a function value under the Hölder condition of order p. In fact, Lemma 2.2.2 can be implied by Lemma 2.2.3. However, as the inequality in Lemma 2.2.2 will be repeatedly used in the proofs about diminishing and constant stepsizes, we still keep it for readers' convenience.

Lemma 2.2.2. *Suppose that Assumption 2.2.1 holds. Let $x \in X$ and $\zeta \geq 0$ satisfy $f(x) > f^* + L\zeta^p + \epsilon$, and let $g(x) \in \bar{\partial}^*_\epsilon f(x) \cap \mathbf{S}$. Then $\langle g(x), x - x^* \rangle \geq \zeta$ for all $x^* \in X^*$.*

Lemma 2.2.3. *Suppose that Assumption 2.2.1 holds. Let $x \in X$ satisfy $f(x) > f^* + \epsilon$, and let $g(x) \in \bar{\partial}^*_\epsilon f(x) \cap \mathbf{S}$. Then*

$$\langle g(x), x - x^* \rangle \geq \left(\frac{f(x) - f^* - \epsilon}{L} \right)^{\frac{1}{p}} \quad \textit{for all} \quad x^* \in X^*.$$

With the introduction of extrapolation, we establish the following extended basic inequality for EiQSG (2.2).

Lemma 2.2.4. *Let $\{x^k\}$ be the sequence generated by EiQSG (2.2). Then, for any $k \in \mathbb{N}$ and $x \in X$, one has that*

$$\begin{aligned} \|x^{k+1} - x\|^2 \leq & \|x^k - x\|^2 + 12d^2\alpha^k + 4(1+R)dv^k\alpha^k \\ & - 2v^k\langle g(x^k), x^k - x\rangle + 4Rdv^k + \left[(1+R)v^k\right]^2. \end{aligned}$$

Proof. By use of the nonexpansive property of the projection operator and (2.2), we obtain that, for any $k \in \mathbb{N}$ and $x \in X$,

$$\begin{aligned} \|x^{k+1} - x\|^2 \leq & \|\hat{x}^k - v^k\tilde{g}(x^k) - x\|^2 \\ = & \|x^k - x\|^2 + \|\hat{x}^k - x^k\|^2 + 2\langle x^k - x, \hat{x}^k - x^k\rangle \\ & - 2v^k\langle \tilde{g}(x^k), \hat{x}^k - x^k\rangle - 2v^k\langle \tilde{g}(x^k), x^k - x\rangle + (\|\tilde{g}(x^k)\|v^k)^2. \end{aligned}$$

Then, it follows from the Cauchy-Schwarz inequality, boundedness of X and $\{\alpha_k\}$, and (2.2) that, for any $k \in \mathbb{N}$ and $x \in X$,

$$\begin{aligned}\|x^{k+1} - x\|^2 \leq & \|x^k - x\|^2 + (\|x^k - x^{k-1}\|\alpha^k)^2 + 2\|x^k - x\|\|x^k - x^{k-1}\|\alpha^k \\ & + 2(1+R)\|x^k - x^{k-1}\|v^k\alpha^k - 2v^k\langle g(x^k), x^k - x\rangle \\ & + 2R\|x^k - x\|v^k + (\|\tilde{g}(x^k)\|v^k)^2 \\ \leq & \|x^k - x\|^2 + 12d^2\alpha^k + 4(1+R)dv^k\alpha^k \\ & - 2v^k\langle g(x^k), x^k - x\rangle + 4Rdv^k + \left[(1+R)v^k\right]^2,\end{aligned}$$

which completes the proof. □

2.3 Convergence in Objective Values

In this section, we study the convergence property in objective values for the proposed method with different stepsizes. Our results depend on the inexactness ϵ, level of noise R, and bound parameter d for the constraint set X.

2.3.1 Diminishing and Constant Stepsizes

As the diminishing stepsize and constant stepsize have similar structures but differ in magnitude (see (2.3)), we treat them together in the following theorem.

Theorem 2.3.1. *Let $\{x^k\}$ be the sequence generated by EiQSG (2.2) with the stepsize (2.3) and Assumption 2.2.1 hold. Then*

$$\liminf_{k\to+\infty} f(x^k) \leq f^* + L\left(\frac{v}{2}(1+R)^2 + 2Rd\right)^p + \epsilon.$$

Proof. We assume by contradiction that

$$\liminf_{k\to+\infty} f(x^k) > f^* + L\left(\frac{v}{2}(1+R)^2 + 2Rd\right)^p + \epsilon.$$

Then, there exist $\delta > 0$ and $k_0 \in \mathbb{N}$ such that

$$f(x^k) > f^* + L\left(\frac{v}{2}(1+R)^2 + 2Rd + \delta\right)^p + \epsilon \quad \text{for all} \quad k \geq k_0.$$

It follows from Lemma 2.2.2 that, for any $k \geq k_0$ and $x^* \in X^*$,

$$\langle g(x^k), x^k - x^* \rangle \geq \frac{v}{2}(1+R)^2 + 2Rd + \delta. \tag{2.6}$$

As $\lim_{k \to +\infty} v^k = v$ and $\lim_{k \to +\infty} \alpha^k = 0$, there exists $k_1 \in \mathbb{N}$ such that

$$v^k \leq v + \frac{\delta}{2(1+R)^2} \quad \text{and} \quad \alpha^k \leq \frac{\delta}{4d(1+R)} \qquad \text{for all} \quad k \geq k_1. \tag{2.7}$$

Together with (2.6), (2.7), and Lemma 2.2.4, we have that, for any $k \geq k_2 := \max\{k_0, k_1\}$ and $x^* \in X^*$,

$$\begin{aligned}\|x^{k+1} - x^*\|^2 \leq & \|x^k - x^*\|^2 + 12d^2\alpha^k + 4(1+R)dv^k\alpha^k \\ & - 2v^k\left(\frac{v}{2}(1+R)^2 + 2Rd + \delta\right) + 4Rdv^k + \left[(1+R)v^k\right]^2 \\ \leq & \|x^k - x^*\|^2 + 12d^2\alpha^k - \frac{\delta}{2}v^k,\end{aligned}$$

where, in the right-hand side of the first inequality, the second α^k and one v^k in $\left[(1+R)v^k\right]^2$ were respectively replaced by estimates in (2.7). For $n > k_2$, summing the left-hand side and right-hand side of the above inequality over $k = k_2, \cdots, n$, we obtain that

$$\|x^{n+1} - x^*\|^2 \leq \|x^{k_2} - x^*\|^2 + 12d^2\sum_{k=k_2}^{n}\alpha^k - \frac{\delta}{2}\sum_{k=k_2}^{n}v^k,$$

in contradiction when $n \to +\infty$ to the facts that $\sum_{k=0}^{+\infty}\alpha^k < +\infty$ and $\sum_{k=0}^{+\infty}v^k = +\infty$. Thus, the proof is completed. □

When $v = 0$, the following estimate is the convergence result in objective values for the diminishing stepsize

$$\liminf_{k \to +\infty} f(x^k) \leq f^* + L(2Rd)^p + \epsilon.$$

When $v^k \equiv v > 0$, the estimate in Theorem 2.3.1 is the convergence result in objective values for the constant stepsize.

2.3.2 Dynamic Stepsize

Then, we explore the convergence property for the dynamic stepsize.

Theorem 2.3.2. *Let $\{x^k\}$ be the sequence generated by EiQSG (2.2) with the dynamic stepsize and Assumption 2.2.1 hold. Then either $x^k \in X_\epsilon^*$ for some $k \in \mathbb{N}$, or*

$$\liminf_{k\to+\infty} f(x^k) \le f^* + L\left(\frac{4Rd\bar{\gamma}}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]}\right)^p + \epsilon.$$

Furthermore, if $R = 0$, then either $x^k \in X_\epsilon^*$ for some $k \in \mathbb{N}$, or $\lim_{k\to+\infty} f(x^k) = f^* + \epsilon$.

Proof. If there is $k \in \mathbb{N}$ such that $x^k \in X_\epsilon^*$, then the statement holds automatically. Now, we consider the case that $x^k \notin X_\epsilon^*$ (i.e., $f(x^k) > f^* + \epsilon$) for all $k \in \mathbb{N}$. We assume by contradiction that

$$\liminf_{k\to+\infty} f(x^k) > f^* + L\left(\frac{4Rd\bar{\gamma}}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]}\right)^p + \epsilon.$$

Then, there exist $\delta > 0$ and $k_0 \in \mathbb{N}$ such that

$$f(x^k) > f^* + L\left(\frac{4Rd\bar{\gamma}}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]} + \delta\right)^p + \epsilon \quad \text{for all} \quad k \ge k_0. \tag{2.8}$$

On the other hand, since $f(x^k) > f^* + \epsilon$ for all $k \in \mathbb{N}$, it follows from Lemmas 2.2.3 and 2.2.4 that, for any $k \ge k_0$ and $x^* \in X^*$,

$$\begin{aligned}\|x^{k+1} - x^*\|^2 \le & \|x^k - x^*\|^2 + 12d^2\alpha^k + 4(1+R)dv^k\alpha^k \\ & - 2v^k\left(\frac{f(x^k) - f^* - \epsilon}{L}\right)^{\frac{1}{p}} + 4Rdv^k + \left[(1+R)v^k\right]^2.\end{aligned}$$

Thus, replacing the v^k in the term with α^k by $\bar{v}$ and other v^k's by the formula of the dynamic stepsize, respectively, we obtain that, for any $k \geq k_0$ and $x^* \in X^*$,

$$\begin{aligned}
&\|x^{k+1}-x^*\|^2\\
\leq&\|x^k-x^*\|^2+12d^2\alpha^k+4(1+R)d\bar{v}\alpha^k+4Rd\gamma^k\left(\frac{f(x^k)-f^*-\epsilon}{L}\right)^{\frac{1}{p}}\\
&-\gamma^k[2-\gamma^k(1+R)^2]\left(\frac{f(x^k)-f^*-\epsilon}{L}\right)^{\frac{2}{p}}\\
\leq&\|x^k-x^*\|^2+4[3d+(1+R)\bar{v}]d\alpha^k+4Rd\bar{\gamma}\left(\frac{f(x^k)-f^*-\epsilon}{L}\right)^{\frac{1}{p}} \qquad (2.9)\\
&-\underline{\gamma}[2-\bar{\gamma}(1+R)^2]\left(\frac{f(x^k)-f^*-\epsilon}{L}\right)^{\frac{2}{p}}\\
=&\|x^k-x^*\|^2+4[3d+(1+R)\bar{v}]d\alpha^k+\frac{4R^2d^2\bar{\gamma}^2}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]}\\
&-\underline{\gamma}[2-\bar{\gamma}(1+R)^2]\left[\left(\frac{f(x^k)-f^*-\epsilon}{L}\right)^{\frac{1}{p}}-\frac{2Rd\bar{\gamma}}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]}\right]^2.
\end{aligned}$$

Together with (2.8) and (2.9), we have that, for any $k \geq k_0$ and $x^* \in X^*$,

$$\begin{aligned}
&\|x^{k+1}-x^*\|^2\\
\leq&\|x^k-x^*\|^2+4[3d+(1+R)\bar{v}]d\alpha^k+\frac{4R^2d^2\bar{\gamma}^2}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]}\\
&-\underline{\gamma}[2-\bar{\gamma}(1+R)^2]\left[\left(\frac{4Rd\bar{\gamma}}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]}+\delta\right)-\frac{2Rd\bar{\gamma}}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]}\right]^2\\
\leq&\|x^k-x^*\|^2+4[3d+(1+R)\bar{v}]d\alpha^k\\
&-\underline{\gamma}[2-\bar{\gamma}(1+R)^2]\left[\left(\frac{2Rd\bar{\gamma}}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]}+\delta\right)^2-\left(\frac{2Rd\bar{\gamma}}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]}\right)^2\right]\\
\leq&\|x^k-x^*\|^2+4[3d+(1+R)\bar{v}]d\alpha^k-\underline{\gamma}[2-\bar{\gamma}(1+R)^2]\delta^2.
\end{aligned}$$

For $n > k_0$, summing the left-hand side and right-hand side of the above inequality over $k = k_0, \cdots, n$, we obtain that

$$\begin{aligned}\|x^{n+1} - x^*\|^2 \leq& \|x^{k_0} - x^*\|^2 + 4[3d + (1+R)\bar{v}]d \sum_{k=k_0}^{n} \alpha^k \\ &- (n+1-k_0)\underline{\gamma}[2 - \bar{\gamma}(1+R)^2]\delta^2,\end{aligned}$$

in contradiction when $n \to +\infty$ to the facts that $\sum\limits_{k=0}^{+\infty} \alpha^k < +\infty$ and the last term tends to $+\infty$ as $n \to +\infty$.

Then, we consider the case that $R = 0$. Since $f(x^k) > f^* + \epsilon$ for all $k \in \mathbb{N}$, it follows from (2.9) that, for any $k \geq k_0$ and $x^* \in X^*$,

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 + 4(3d + \bar{v})d\alpha^k - \underline{\gamma}(2 - \bar{\gamma}) \left(\frac{f(x^k) - f^* - \epsilon}{L} \right)^{\frac{2}{p}}.$$

For $n > k_0$, summing the left-hand side and right-hand side of the above inequality over $k = k_0, \cdots, n$, we obtain that

$$\|x^{n+1} - x^*\|^2 \leq \|x^{k_0} - x^*\|^2 + 4(3d + \bar{v})d \sum_{k=k_0}^{n} \alpha^k - \underline{\gamma}(2 - \bar{\gamma}) \sum_{k=k_0}^{n} \left(\frac{f(x^k) - f^* - \epsilon}{L} \right)^{\frac{2}{p}}.$$

When $n \to +\infty$, it follows from $\sum\limits_{k=0}^{+\infty} \alpha^k < +\infty$ that

$$\lim_{k \to +\infty} \left(\frac{f(x^k) - f^* - \epsilon}{L} \right)^{\frac{2}{p}} = 0.$$

Therefore, we have that

$$\lim_{k \to +\infty} f(x^k) = f^* + \epsilon.$$

Thus, the proof is completed. □

Both theorems indicate the convergence to the optimal value within some tolerance. As the inexactness and noise are considered, the optimal value f^* may not be attained. In Theorem 2.3.1, the tolerance level is $L\left(\frac{v}{2}(1+R)^2+2Rd\right)^p+\epsilon$, the same as the formula provided in Hu et al. (2015); in Theorem 2.3.2, the tolerance level is $L\left(\frac{4Rd\bar{\gamma}}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]}\right)^p+\epsilon$. It is clear that the convergence result in objective values for the inexact quasisubgradient method with extrapolation coincides with that without extrapolation for all the stepsizes.

2.4 Iteration Complexity

This section analyzes the iteration complexity by estimating the difference between the optimal value and best objective value among the first K iterations. The best objective value at the K^{th} iteration is recorded as

$$f_{best}(x^K) := \min_{1\le k\le K} f(x^k).$$

2.4.1 Diminishing and Constant Stepsizes

We again treat the diminishing stepsize and constant stepsize together and consider a special case of (2.3) (see (2.4)). Moreover, a structured extrapolation rule is also assumed in the analysis.

Theorem 2.4.1. *Let $\delta > 0$ and $0 < \eta < \frac{2}{3}$. Let $\{x^k\}$ be the sequence generated by EiQSG (2.2) with the stepsize (2.4) and extrapolation rule $\alpha^k = o\left(v^k\right)$ and Assumption 2.2.1 hold. Then there exists $\bar{K} \in \mathbb{N}$ such that*

$$f_{best}(x^K) - f^* \le L\left(\frac{v}{2}(1+R)^2+2Rd+\delta\right)^p+\epsilon,$$

where K is the minimum integer such that

$$(2-3\eta)\delta\left((K-\bar{K}+1)v+\frac{c[(K+1)^{1-s}-\bar{K}^{1-s}]}{1-s}\right) > \operatorname{dist}^2(x^{\bar{K}}, X^*).$$

Proof. We assume by contradiction that, for any $1 \leq k \leq K$,

$$f(x^k) - f^* > L\left(\frac{v}{2}(1+R)^2 + 2Rd + \delta\right)^p + \epsilon.$$

It follows from Lemma 2.2.2 that, for any $1 \leq k \leq K$ and $x^* \in X^*$,

$$\langle g(x^k), x^k - x^* \rangle \geq \frac{v}{2}(1+R)^2 + 2Rd + \delta. \tag{2.10}$$

Given $\delta > 0$ and $0 < \eta < \frac{2}{3}$, it is easy to see that there exists $\bar{K} \in \mathbb{N}$ such that

$$v^k \leq v + \frac{\eta\delta}{(1+R)^2} \quad \text{for all} \quad k \geq \bar{K}, \tag{2.11}$$

and

$$\alpha^k \leq \min\left\{\frac{\eta\delta v^k}{12d^2}, \frac{\eta\delta}{4(1+R)d}\right\} \quad \text{for all} \quad k \geq \bar{K}. \tag{2.12}$$

Together with (2.10), (2.11), and Lemma 2.2.4, we have that, for any $\bar{K} \leq k \leq K$ and $x^* \in X^*$,

$$\begin{aligned}\|x^{k+1} - x^*\|^2 \leq& \|x^k - x^*\|^2 + 12d^2\alpha^k + 4(1+R)dv^k\alpha^k \\ &- 2v^k\langle g(x^k), x^k - x^* \rangle + 4Rdv^k + \left[(1+R)v^k\right]^2 \\ \leq& \|x^k - x^*\|^2 + 12d^2\alpha^k + 4(1+R)dv^k\alpha^k - 2v^k\left(\frac{v}{2}(1+R)^2 + 2Rd + \delta\right) \\ &+ 4Rdv^k + (1+R)^2v^k\left(v + \frac{\eta\delta}{(1+R)^2}\right) \\ =& \|x^k - x^*\|^2 + 12d^2\alpha^k + 4(1+R)dv^k\alpha^k - 2\delta v^k + \eta\delta v^k.\end{aligned}$$

Now by (2.12), for any $\bar{K} \leq k \leq K$ and $x^* \in X^*$, it holds that

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - (2-3\eta)\delta v^k = \|x^k - x^*\|^2 - (2-3\eta)\delta\left(v + ck^{-s}\right).$$

Summing the above inequality over $k = \bar{K}, \cdots, K$, we obtain that

$$(2-3\eta)\delta\sum_{k=\bar{K}}^{K}\left(v + ck^{-s}\right) \leq \|x^{\bar{K}} - x^*\|^2 - \|x^{K+1} - x^*\|^2 \leq \|x^{\bar{K}} - x^*\|^2. \tag{2.13}$$

Since k^{-s} decreases as k increases, we see that

$$\sum_{k=\bar{K}}^{K} k^{-s} \geq \int_{\bar{K}}^{K+1} t^{-s} \mathrm{d}t = \frac{(K+1)^{1-s} - \bar{K}^{1-s}}{1-s}\,.$$

Let $x^* = P_{X^*}(x^{\bar{K}})$. Then, it follows from (2.13) that

$$(2-3\eta)\delta\left((K-\bar{K}+1)v + \frac{c[(K+1)^{1-s} - \bar{K}^{1-s}]}{1-s}\right) \leq \mathrm{dist}^2(x^{\bar{K}}, X^*),$$

in contradiction to the definition of K. Thus, the proof is completed. □

When $v = 0$ and $c > 0$, the following estimate provides the tolerance level for the diminishing stepsize $v^k = ck^{-s}$

$$f_{best}(x^K) - f^* \leq L\,(2Rd + \delta)^p + \epsilon.$$

Moreover, if $\alpha^k = k^{-t}$ (for each $t > 1$), from (2.11) and (2.12), the specific forms of $\bar{K}$ and K are respectively given as

$$\bar{K} := \max\left\{\left\lceil\left(\frac{12d^2}{\eta c\delta}\right)^{\frac{1}{t-s}}\right\rceil, \left\lceil\left(\frac{4(1+R)d}{\eta\delta}\right)^{\frac{1}{t}}\right\rceil, \left\lceil\left(\frac{c(1+R)^2}{\eta\delta}\right)^{\frac{1}{s}}\right\rceil\right\}$$

and

$$K := \left\lceil\left(\bar{K}^{1-s} + \frac{(1-s)\mathrm{dist}^2(x^{\bar{K}}, X^*)}{(2-3\eta)c\delta}\right)^{\frac{1}{1-s}}\right\rceil.$$

When $v > 0$ and $c = 0$, the estimate in Theorem 2.4.1 provides the tolerance level for the constant stepsize $v^k \equiv v$. Moreover, if $\alpha^k = k^{-t}$ (for each $t > 1$), from (2.11) and (2.12), the specific forms of $\bar{K}$ and K are respectively given as

$$\bar{K} := \max\left\{\left\lceil\left(\frac{12d^2}{\eta v\delta}\right)^{\frac{1}{t}}\right\rceil, \left\lceil\left(\frac{4(1+R)d}{\eta\delta}\right)^{\frac{1}{t}}\right\rceil\right\}$$

and

$$K := \left\lceil \bar{K} + \frac{\mathrm{dist}^2(x^{\bar{K}}, X^*)}{(2-3\eta)v\delta}\right\rceil.$$

2.4.2 Dynamic Stepsize

Next, we see the complexity for the dynamic stepsize.

Theorem 2.4.2. *Let $\delta > 0$ and $0 < \eta < 1$. Let $\{x^k\}$ be the sequence generated by EiQSG (2.2) with the dynamic stepsize and Assumption 2.2.1 hold. Then there exists $\bar{K} \in \mathbb{N}$ such that*

$$f_{best}(x^K) - f^* \leq L\left(\frac{4Rd\bar{\gamma}}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]} + \delta\right)^p + \epsilon,$$

where $K := \left\lceil \bar{K} + \frac{\mathrm{dist}^2(x^{\bar{K}}, X^*)}{(1-\eta)[2-\bar{\gamma}(1+R)^2]\underline{\gamma}\delta^2} \right\rceil$.

Proof. We assume by contradiction that, for any $1 \leq k \leq K$,

$$f(x^k) - f^* > L\left(\frac{4Rd\bar{\gamma}}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]} + \delta\right)^p + \epsilon.$$

It follows from Lemmas 2.2.3 and 2.2.4 that, for any $1 \leq k \leq K$ and $x^* \in X^*$,

$$\begin{aligned}\|x^{k+1} - x^*\|^2 \leq &\|x^k - x^*\|^2 + 12d^2\alpha^k + 4d(1+R)v^k\alpha^k \\ &- 2v^k\left(\frac{f(x^k) - f^* - \epsilon}{L}\right)^{\frac{1}{p}} + 4Rdv^k + \left[(1+R)v^k\right]^2.\end{aligned}$$

Thus, replacing the v^k in the term with α^k by $\bar{v}$ and other v^k's by the formula of the dynamic stepsize, respectively, we obtain that, for any $1 \leq k \leq K$ and $x^* \in X^*$,

$$\begin{aligned}\|x^{k+1} - x^*\|^2 \leq &\|x^k - x^*\|^2 + 4[3d + (1+R)\bar{v}]d\alpha^k + 4Rd\gamma^k\left(\frac{f(x^k) - f^* - \epsilon}{L}\right)^{\frac{1}{p}} \\ &- \gamma^k[2-\gamma^k(1+R)^2]\left(\frac{f(x^k) - f^* - \epsilon}{L}\right)^{\frac{2}{p}} \\ \leq &\|x^k - x^*\|^2 + 4[3d + (1+R)\bar{v}]d\alpha^k + 4Rd\bar{\gamma}\left(\frac{f(x^k) - f^* - \epsilon}{L}\right)^{\frac{1}{p}} \\ &- \underline{\gamma}[2-\bar{\gamma}(1+R)^2]\left(\frac{f(x^k) - f^* - \epsilon}{L}\right)^{\frac{2}{p}}.\end{aligned}$$

Rearranging the terms, we have that

$$\begin{aligned}
&\|x^{k+1}-x^*\|^2\\
\leq&\|x^k-x^*\|^2+4[3d+(1+R)\bar{v}]d\alpha^k+\frac{4R^2d^2\bar{\gamma}^2}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]}\\
&-\underline{\gamma}[2-\bar{\gamma}(1+R)^2]\left[\left(\frac{f(x^k)-f^*-\epsilon}{L}\right)^{\frac{1}{p}}-\frac{2Rd\bar{\gamma}}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]}\right]^2\\
\leq&\|x^k-x^*\|^2+4[3d+(1+R)\bar{v}]d\alpha^k\\
&-\underline{\gamma}[2-\bar{\gamma}(1+R)^2]\left[\left(\frac{2Rd\bar{\gamma}}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]}+\delta\right)^2-\left(\frac{2Rd\bar{\gamma}}{\underline{\gamma}[2-\bar{\gamma}(1+R)^2]}\right)^2\right]\\
\leq&\|x^k-x^*\|^2+4[3d+(1+R)\bar{v}]d\alpha^k-\underline{\gamma}[2-\bar{\gamma}(1+R)^2]\delta^2.
\end{aligned}$$

Given $\delta>0$ and $0<\eta<1$, it is easy to see that there exists $\bar{K}\in\mathbb{N}$ such that

$$\alpha^k\leq\frac{\eta\underline{\gamma}[2-\bar{\gamma}(1+R)^2]\delta^2}{4[3d+(1+R)\bar{v}]d}\quad\text{for all}\quad k\geq\bar{K}\tag{2.14}$$

Now by (2.14), for any $\bar{K}\leq k\leq K$ and $x^*\in X^*$, it holds that

$$\|x^{k+1}-x^*\|^2\leq\|x^k-x^*\|^2-(1-\eta)[2-\bar{\gamma}(1+R)^2]\underline{\gamma}\delta^2.$$

Summing the above inequality over $k=\bar{K},\cdots,K$ and letting $x^*=P_{X^*}(x^{\bar{K}})$, we obtain that

$$(K-\bar{K}+1)(1-\eta)[2-\bar{\gamma}(1+R)^2]\underline{\gamma}\delta^2\leq\|x^{\bar{K}}-x^*\|^2-\|x^{K+1}-x^*\|^2\leq\operatorname{dist}^2(x^{\bar{K}},X^*),$$

in contradiction to the definition of K. Thus, the proof is completed. □

If $\alpha^k=k^{-t}$ (for each $t>1$), from (2.14), the specific form of $\bar{K}$ is given as

$$\bar{K}:=\left\lceil\left(\frac{4[3d+(1+R)\bar{v}]d}{\eta\underline{\gamma}[2-\bar{\gamma}(1+R)^2]\delta^2}\right)^{\frac{1}{t}}\right\rceil.$$

When $\alpha^k \equiv 0$ and $R = 0$, Theorems 2.4.1 and 2.4.2 are consistent with Theorem 3.2 in Hu et al. (2020). The following remark specifies the best iteration complexity for all the stepsizes.

Remark 2.4.3. *From Theorems 2.4.1 and 2.4.2, the computational complexity of EiQSG is influenced by the extrapolation rule, and the theoretically best complexities are $\mathcal{O}(1/k^{p\min\{s,1-s\}})$, $\mathcal{O}(1/k^p)$, and $\mathcal{O}(1/k^{\frac{p}{2}})$. In particular, all the best complexities can be achieved for the respective stepsizes when $\alpha^k \equiv 0$ or $\alpha^k = k^{-t}$ (for each $t > 1$).*

2.5 Convergence Rate for Iterates

Section 2.4 investigates the rate of convergence in objective values. In this section, we study the convergence rate for iterates under a weak sharp minima condition.

The classical weak sharp minima condition was introduced in Burke & Ferris (1993). After that, a generalized definition, the weak sharp minima condition of order q, was identified by Studniarski & Ward (1999). An assumption of weak sharp minima of order q is made in analyzing the rate of convergence in iterates.

Assumption 2.5.1. *Assume that X^* for (2.1) is a set of weak sharp minima of order $q > 0$ with modulus $\rho > 0$ over X, i.e.,*

$$f(x) - f^* \geq \rho \operatorname{dist}^q(x, X^*) \quad \textit{for all} \quad x \in X.$$

It is well-known that the optimal solution set of a linear programming problem or a convex quadratic programming problem is a set of weak sharp minima of order 1 (see Burke & Ferris (1993)). Another example (see Studniarski & Ward (1999, Example 2.1)) is

$$\begin{aligned} \min_{x,y} \quad & x^q \\ \text{s.t.} \quad & x \geq 0,\ 0 \leq y \leq 1, \end{aligned}$$

where $q > 0$. Clearly, its optimal solution set $\{(x, y) : x = 0 \text{ and } 0 \leq y \leq 1\}$ is a set of weak sharp minima of order q. When $0 < q < 1$, the problem is quasiconvex.

Lemma 2.5.2 (Huang & Yang (2003, Lemma 4.1)) and Lemma 2.5.3 (Polyak (1987, Lemmas 4 and 5)) provide several significant inequalities for the following convergence analysis.

Lemma 2.5.2. *Let $a \geq b \geq 0$, $c := \min\{1, 2^{1-t}\}$, and $t > 0$. Then $(a-b)^t \geq ca^t - b^t$.*

Lemma 2.5.3. *Let $u_k \geq 0$ and $u_{k+1} \leq (1 - ak^{-s})u_k + bk^{-t}$ for all $k \in \mathbb{N}$, where $a > 0$ and $b > 0$.*

(1). If $s = 1$, $t > 1$, and $a > t - 1$, then

$$u_k \leq \frac{b}{a - t + 1} k^{1-t} + o(k^{1-t}) \quad \text{for all} \quad k \in \mathbb{N}.$$

(2). If $0 < s < 1$ and $t > s$, then

$$u_k \leq \frac{b}{a} k^{s-t} + o(k^{s-t}) \quad \text{for all} \quad k \in \mathbb{N}.$$

To find an estimate of u_k in Lemma 2.5.3 for the case that $0 < s < 1$ and $t = s$, we establish Lemma 2.5.4, a modified version of Lemma 2.3 (ii) in Hu et al. (2020).

Lemma 2.5.4. *Let $u_k \geq 0$ and $u_{k+1} \leq (1 - ak^{-s})u_k + bk^{-s}$ for all $k \in \mathbb{N}$, where $a > 0$, $b > 0$, and $0 < s < 1$. Then*

$$u_k \leq c\tau^{k^{1-s}} + \frac{b}{a} \quad \text{for all} \quad k \geq \lceil a^{\frac{1}{s}} \rceil + 2,$$

where $c := u_{k_0} e^{\frac{a}{1-s} k_0^{1-s}}$ and $\tau := e^{-\frac{a}{1-s}}$.

Proof. For any $k \geq k_0 := \lceil a^{\frac{1}{s}} \rceil + 1$, we have that

$$u_{k+1} - \frac{b}{a} \leq \left(1 - ak^{-s}\right)\left(u_k - \frac{b}{a}\right) \leq \left(1 - ak^{-s}\right)\left(1 - a(k-1)^{-s}\right)\left(u_{k-1} - \frac{b}{a}\right)$$

$$\leq \cdots \ \cdots \ \cdots \leq \left(u_{k_0} - \frac{b}{a}\right)\prod_{i=k_0}^{k}\left(1 - ai^{-s}\right) = \left(u_{k_0} - \frac{b}{a}\right) e^{\sum_{i=k_0}^{k} \ln\left(1 - ai^{-s}\right)}.$$

As $\{\ln(1 - ai^{-s})\}$ is increasing and $\ln(1 - ai^{-s}) < -ai^{-s}$ holds for all $i \geq k_0$, one has that

$$\sum_{i=k_0}^{k} \ln\left(1 - ai^{-s}\right) < \int_{k_0}^{k+1} \ln\left(1 - at^{-s}\right) \mathrm{d}t < \int_{k_0}^{k+1} -at^{-s}\mathrm{d}t = \frac{a}{1-s}\left[k_0^{1-s} - (k+1)^{1-s}\right].$$

Then, we see that

$$u_{k+1} - \frac{b}{a} \leq \left(u_{k_0} - \frac{b}{a}\right) e^{\frac{a}{1-s}k_0^{1-s}} e^{-\frac{a}{1-s}(k+1)^{1-s}} \leq u_{k_0} e^{\frac{a}{1-s}k_0^{1-s}} \left(e^{-\frac{a}{1-s}}\right)^{(k+1)^{1-s}},$$

which completes the proof. □

Remark 2.5.5. *In Lemma 2.5.4, $\{u_k\}$ is sublinearly convergent with the rate $\mathcal{O}\left(\tau^{k^{1-s}}\right)$ for some $0 < \tau < 1$, which is faster than $\mathcal{O}\left(1/k^h\right)$ for each $h > 0$.*

Proof. The sublinear convergence can be verified by definition, specifically,

$$\lim_{k\to+\infty} \frac{\tau^{(k+1)^{1-s}}}{\tau^{k^{1-s}}} = \tau^{\lim\limits_{k\to+\infty} k^{1-s}\left[\left(1+\frac{1}{k}\right)^{1-s} - 1\right]} = \tau^{\lim\limits_{k\to+\infty} k^{1-s}\left(1+\frac{1-s}{k}+o\left(\frac{1}{k^2}\right)-1\right)} = 1,$$

where the second equality comes from the Taylor's theorem.

Now, we prove that $\tau^{k^{1-s}} = o\left(1/k^h\right)$. Letting $t := \frac{a}{1-s}k^{1-s}$ and $i_0 := \left\lceil \frac{h}{1-s} - 1 \right\rceil$, we see that

$$\lim_{k\to+\infty} \frac{\tau^{k^{1-s}}}{\frac{1}{k^h}} = \lim_{t\to+\infty} \frac{\left(\frac{1-s}{a}t\right)^{\frac{h}{1-s}}}{e^t} = \left(\frac{1-s}{a}\right)^{\frac{h}{1-s}} \lim_{t\to+\infty} \frac{t^{\frac{h}{1-s}}}{e^t}$$

$$= \left(\frac{1-s}{a}\right)^{\frac{h}{1-s}} \prod_{i=0}^{i_0}\left(\frac{h}{1-s} - i\right) \lim_{t\to+\infty} \frac{t^{\frac{h}{1-s} - i_0 - 1}}{e^t} = 0,$$

where the second line holds due to the L'Hôspital's rule. □

2.5.1 Diminishing Stepsize

When exploring the rate of convergence for the diminishing stepsize, we focus on some structured stepsize and extrapolation rule. The estimates in Lemmas 2.5.3 and 2.5.4 are used to prove a sublinear convergence rate.

Theorem 2.5.6. *Suppose that Assumptions 2.2.1 and 2.5.1 hold with $p \le q \le 2p$. Let $\{x^k\}$ be the sequence generated by EiQSG (2.2) with the diminishing stepsize $v^k = c_1 k^{-s}$ and extrapolation rule $\alpha^k \le c_2 k^{-2s}$, where $c_1 > 0$, $c_2 \ge 0$, and $0 < s \le 1$.*

(1). If $0 < s \le 1$, $R = 0$, $\epsilon = 0$, and $c_1 > (2d)^{1-\frac{q}{p}} d \left(\frac{L}{\rho}\right)^{\frac{1}{p}}$ when $s = 1$, then either $x^k \in X^*$ for some $k \in \mathbb{N}$, or there exists $C > 0$ such that

$$\operatorname{dist}^2(x^k, X^*) \le Ck^{-s} \quad \textit{for sufficiently large } k.$$

(2). If $0 < s < 1$ and either $R > 0$ or $\epsilon > 0$, then either $x^k \in X_\epsilon^*$ for some $k \in \mathbb{N}$, or there exist $C > 0$, $D > 0$, and $0 < \tau < 1$ such that

$$\operatorname{dist}^2(x^k, X^*) \le C\tau^{k^{1-s}} + D \quad \textit{for sufficiently large } k.$$

Proof. If there is $k \in \mathbb{N}$ such that $x^k \in X_\epsilon^*$, then the statements (1) and (2) hold automatically. Now, we consider the case that $x^k \notin X_\epsilon^*$ (i.e., $f(x^k) > f^* + \epsilon$) for all $k \in \mathbb{N}$. Thus, by use of Lemmas 2.2.3 and 2.2.4, we have that, for any $k \in \mathbb{N}$ and $x^* \in X^*$,

$$\begin{aligned}\|x^{k+1} - x^*\|^2 \le &\|x^k - x^*\|^2 + 12d^2 c_2 k^{-2s} + 4(1+R)dc_1c_2k^{-3s} \\ &- 2c_1k^{-s}\left(\frac{f(x^k) - f^* - \epsilon}{L}\right)^{\frac{1}{p}} + 4Rdc_1k^{-s} + (1+R)^2c_1^2k^{-2s}.\end{aligned}$$

Let $C_1 := 12d^2c_2 + 4(1+R)dc_1c_2 + (1+R)^2c_1^2$ and $x^* = P_{X^*}(x^k)$. Since $k^{-3s} \le k^{-2s}$, it holds that, for any $k \in \mathbb{N}$,

$$\operatorname{dist}^2(x^{k+1}, X^*) \le \operatorname{dist}^2(x^k, X^*) - 2c_1 k^{-s} \left(\frac{f(x^k) - f^* - \epsilon}{L} \right)^{\frac{1}{p}} + C_1 k^{-2s} + 4Rdc_1 k^{-s}. \tag{2.15}$$

We consider two cases: $\epsilon = 0$, and $\epsilon > 0$.

Case (i). Let $\epsilon = 0$. It follows from (2.15), Assumption 2.5.1, and the fact of $\operatorname{dist}^{\frac{q}{p}-2}(x^k, X^*) \ge (2d)^{\frac{q}{p}-2}$ that, for any $k \in \mathbb{N}$,

$$\operatorname{dist}^2(x^{k+1}, X^*) \le \left[1 - 2(2d)^{\frac{q}{p}-2} \left(\frac{\rho}{L} \right)^{\frac{1}{p}} c_1 k^{-s} \right] \operatorname{dist}^2(x^k, X^*) + C_1 k^{-2s} + 4Rdc_1 k^{-s}. \tag{2.16}$$

Case (ii). Let $\epsilon > 0$. It follows from Lemma 2.5.2 and (2.15) that, for any $k \in \mathbb{N}$,

$$\begin{aligned} \operatorname{dist}^2(x^{k+1}, X^*) \le \operatorname{dist}^2(x^k, X^*) - 2c_1 k^{-s} \left[\tilde{c} \left(\frac{f(x^k) - f^*}{L} \right)^{\frac{1}{p}} - \left(\frac{\epsilon}{L} \right)^{\frac{1}{p}} \right] \\ + C_1 k^{-2s} + 4Rdc_1 k^{-s}, \end{aligned}$$

where $\tilde{c} := \min\{1, 2^{1-\frac{1}{p}}\}$. Then, combining with Assumption 2.5.1 and the fact of $\operatorname{dist}^{\frac{q}{p}-2}(x^k, X^*) \ge (2d)^{\frac{q}{p}-2}$, we have that, for any $k \in \mathbb{N}$,

$$\begin{aligned} \operatorname{dist}^2(x^{k+1}, X^*) \le \left[1 - 2(2d)^{\frac{q}{p}-2} \left(\frac{\rho}{L} \right)^{\frac{1}{p}} \tilde{c} c_1 k^{-s} \right] \operatorname{dist}^2(x^k, X^*) \\ + \left[2c_1 \left(\frac{\epsilon}{L} \right)^{\frac{1}{p}} + 4Rdc_1(1 + \kappa_1) \right] k^{-s}, \end{aligned} \tag{2.17}$$

where κ_1 is an arbitrary positive number.

Next, we prove the statements (1) and (2).

(1). Let $a := 2(2d)^{\frac{q}{p}-2} \left(\frac{\rho}{L} \right)^{\frac{1}{p}} c_1$ and $b := C_1$. Consider first $s = 1$, $R = 0$, $\epsilon = 0$, and $a > 1$. Applying Lemma 2.5.3(1) to (2.16), we obtain that

$$\operatorname{dist}^2(x^k, X^*) \le \frac{b}{a-1} k^{-1} + o\left(k^{-1}\right) \quad \text{for all} \quad k \in \mathbb{N}.$$

Consider then $0 < s < 1$, $R = 0$, and $\epsilon = 0$. Applying Lemma 2.5.3(2) to (2.16), we obtain that

$$\operatorname{dist}^2(x^k, X^*) \leq \frac{a}{b} k^{-s} + o\left(k^{-s}\right) \quad \text{for all} \quad k \in \mathbb{N}.$$

From both cases, the statement (1) is true.

(2). Consider first $0 < s < 1$, $R > 0$, and $\epsilon = 0$. Let κ_2 be an arbitrary positive number. For any $k \geq \left(\frac{C_1}{4Rdc_1\kappa_2}\right)^{\frac{1}{s}}$, (2.16) reduces to

$$\operatorname{dist}^2(x^{k+1}, X^*) \leq \left[1 - 2(2d)^{\frac{q}{p}-2}\left(\frac{\rho}{L}\right)^{\frac{1}{p}} c_1 k^{-s}\right] \operatorname{dist}^2(x^k, X^*) + 4Rdc_1(1+\kappa_2)k^{-s}.$$

Now, we apply Lemma 2.5.4 to the above inequality; then, there exist $0 < \tau < 1$, $C > 0$, and $D > 0$ such that

$$\operatorname{dist}^2(x^k, X^*) \leq C\tau^{k^{1-s}} + D \quad \text{for sufficiently large } k.$$

Consider then $0 < s < 1$ and $\epsilon > 0$. We apply Lemma 2.5.4 to (2.17); then, there exist $0 < \tau < 1$, $C > 0$, and $D > 0$ such that

$$\operatorname{dist}^2(x^k, X^*) \leq C\tau^{k^{1-s}} + D \quad \text{for sufficiently large } k.$$

From both cases, the statement (2) is true. □

2.5.2 Constant Stepsize

In the discussion of the constant stepsize, we consider another specific extrapolation rule. A linear convergence rate is provided in the following theorem.

Theorem 2.5.7. *Suppose that Assumptions 2.2.1 and 2.5.1 hold with $p \leq q \leq 2p$. Let $\{x^k\}$ be the sequence generated by EiQSG (2.2) with the constant stepsize and extrapolation rule $\alpha^{k+1} \leq \zeta\alpha^k$, where $0 \leq \zeta < 1$. Then either $x^k \in X^*_\epsilon$ for some $k \in \mathbb{N}$, or there exist $C > 0$, $D > 0$, and $0 < \tau < 1$ such that*

$$\operatorname{dist}^2(x^k, X^*) \leq C\tau^k + D \quad \textit{for all} \quad k \geq 1.$$

Proof. If there is $k \in \mathbb{N}$ such that $x^k \in X^*_\epsilon$, then the statement holds automatically. Now, we consider the case that $x^k \notin X^*_\epsilon$ (i.e., $f(x^k) > f^* + \epsilon$) for all $k \in \mathbb{N}$. Thus, by use of Lemmas 2.2.3 and 2.2.4, we have that, for any $k \in \mathbb{N}$ and $x^* \in X^*$,

$$\begin{aligned}\|x^{k+1} - x^*\|^2 \leq & \|x^k - x^*\|^2 + 12d^2\alpha^k + 4(1+R)dv\alpha^k \\ & - 2v\left(\frac{f(x^k) - f^* - \epsilon}{L}\right)^{\frac{1}{p}} + 4Rdv + (1+R)^2v^2.\end{aligned}$$

Let $x^* = P_{X^*}(x^k)$. It holds that, for any $k \in \mathbb{N}$,

$$\begin{aligned}\operatorname{dist}^2(x^{k+1}, X^*) \leq & \operatorname{dist}^2(x^k, X^*) + 12d^2\alpha^k + 4(1+R)dv\alpha^k \\ & - 2v\left(\frac{f(x^k) - f^* - \epsilon}{L}\right)^{\frac{1}{p}} + 4Rdv + (1+R)^2v^2.\end{aligned} \tag{2.18}$$

We consider two cases: $\epsilon = 0$, and $\epsilon > 0$.

Case (i). Let $\epsilon = 0$. It follows from (2.18), Assumption 2.5.1, and the fact of $\operatorname{dist}^{\frac{q}{p}-2}(x^k, X^*) \geq (2d)^{\frac{q}{p}-2}$ that, for any $k \in \mathbb{N}$,

$$\operatorname{dist}^2(x^{k+1}, X^*) \leq C_1 \operatorname{dist}^2(x^k, X^*) + C_2\alpha^k + C_3,$$

where $C_1 := 1 - 2(2d)^{\frac{q}{p}-2}\left(\frac{\rho}{L}\right)^{\frac{1}{p}} v$, $C_2 := 4[3d + (1+R)v]d$, and $C_3 := 4Rdv + (1+R)^2v^2$. If $C_1 \leq 0$, then, noting $\alpha^{k+1} \leq \zeta\alpha^k$, we have that, for any $k \in \mathbb{N}$,

$$\operatorname{dist}^2(x^{k+1}, X^*) \leq C_2\alpha^k + C_3 \leq \cdots \; \cdots \leq C_2\zeta^k\alpha^0 + C_3 = \frac{C_2\alpha^0}{\zeta}\zeta^{k+1} + C_3. \tag{2.19}$$

If $0 < C_1 < 1$, then we have that, for any $k \in \mathbb{N}$,

$$\begin{aligned}\operatorname{dist}^2(x^{k+1}, X^*) & \leq C_1 \operatorname{dist}^2(x^k, X^*) + C_2\alpha^k + C_3 \\ & \leq C_1[C_1 \operatorname{dist}^2(x^{k-1}, X^*) + C_2\alpha^{k-1} + C_3] + C_2\alpha^{k-1} + C_3 \\ & \cdots \; \cdots \; \cdots \\ & \leq C_1^{k+1} \operatorname{dist}^2(x^0, X^*) + C_2 \sum_{i=0}^{k} C_1^i \alpha^{k-i} + C_3 \sum_{i=0}^{k} C_1^i.\end{aligned}$$

Let $\max\{C_1, \zeta\} < \tau < 1$. Then it follows from $\alpha^{k+1} \leq \zeta \alpha^k$ that, for any $k \in \mathbb{N}$,

$$\begin{aligned} \mathrm{dist}^2(x^{k+1}, X^*) &\leq C_1^{k+1}\mathrm{dist}^2(x^0, X^*) + C_2\alpha^0 \sum_{i=0}^{k} C_1^i \tau^{k-i} + C_3 \sum_{i=0}^{k} C_1^i \\ &\leq \mathrm{dist}^2(x^1, X^*)\ \tau^{k+1} + \frac{C_2\alpha^0}{(\tau - C_1)}\ \tau^{k+1} + \frac{C_3}{1 - C_1}. \end{aligned} \tag{2.20}$$

From (2.19) and (2.20), the statement is true for this case.

Case (ii). Let $\epsilon > 0$. It follows from Lemma 2.5.2 and (2.18) that, for any $k \in \mathbb{N}$,

$$\mathrm{dist}^2(x^{k+1}, X^*) \leq \mathrm{dist}^2(x^k, X^*) + C_2\alpha^k - 2v\left[\tilde{c}\left(\frac{f(x^k) - f^*}{L}\right)^{\frac{1}{p}} - \left(\frac{\epsilon}{L}\right)^{\frac{1}{p}}\right] + C_3,$$

where $\tilde{c} := \min\{1, 2^{1-\frac{1}{p}}\}$. Then, combining with Assumption 2.5.1 and the fact of $\mathrm{dist}^{\frac{q}{p}-2}(x^k, X^*) \geq (2d)^{\frac{q}{p}-2}$, we have that, for any $k \in \mathbb{N}$,

$$\mathrm{dist}^2(x^{k+1}, X^*) \leq \overline{C}_1\mathrm{dist}^2(x^k, X^*) + C_2\alpha^k + \overline{C}_3,$$

where $\overline{C}_1 := 1 - 2(2d)^{\frac{q}{p}-2}\left(\frac{\rho}{L}\right)^{\frac{1}{p}}\tilde{c}v$ and $\overline{C}_3 := C_3 + 2v\left(\frac{\epsilon}{L}\right)^{\frac{1}{p}}$. The rest of the proof for this case can be completed similarly as for Case (i). □

2.5.3 Dynamic Stepsize

For the proposed method with the dynamic stepsize, a linear convergence for iterates is obtained in Theorem 2.5.8. The extrapolation rule under consideration is the same as that in Theorem 2.5.7.

Theorem 2.5.8. *Suppose that Assumptions 2.2.1 and 2.5.1 hold with $q = p$. Let $\{x^k\}$ be the sequence generated by EiQSG (2.2) with the dynamic stepsize and extrapolation rule $\alpha^{k+1} \leq \zeta\alpha^k$, where $0 \leq \zeta < 1$. Then either $x^k \in X_\epsilon^*$ for some $k \in \mathbb{N}$, or there exist $C > 0$, $D > 0$, and $0 < \tau < 1$ such that*

$$\mathrm{dist}^2(x^k, X^*) \leq C\tau^k + D \quad \text{for all} \quad k \geq 1.$$

Proof. If there is $k \in \mathbb{N}$ such that $x^k \in X^*_\epsilon$, then the statement holds automatically. Now, we consider the case that $x^k \notin X^*_\epsilon$ (i.e., $f(x^k) > f^* + \epsilon$) for all $k \in \mathbb{N}$. Thus, by use of Lemmas 2.2.3 and 2.2.4, we have that, for any $k \in \mathbb{N}$ and $x^* \in X^*$,

$$\begin{aligned}\|x^{k+1} - x\|^2 \leq & \|x^k - x^*\|^2 + 12d^2\alpha^k + 4(1+R)dv^k\alpha^k \\ & - 2v^k \left(\frac{f(x^k) - f^* - \epsilon}{L}\right)^{\frac{1}{p}} + 4Rdv^k + \left[(1+R)v^k\right]^2 \\ \leq & \|x^k - x^*\|^2 + 4[3d + (1+R)\bar{v}]d\alpha^k + 4Rd\bar{v} \\ & - \underline{\gamma}[2 - \bar{\gamma}(1+R)^2] \left(\frac{f(x^k) - f^* - \epsilon}{L}\right)^{\frac{2}{p}},\end{aligned}$$

where v^k was replaced by $\bar{v}$ in the terms $4(1+R)dv^k\alpha^k$ and $4Rdv^k$ and other v^k's by the formula of the dynamic stepsize. Let $x^* = P_{X^*}(x^k)$. It holds that, for any $k \in \mathbb{N}$,

$$\begin{aligned}\mathrm{dist}^2(x^{k+1}, X^*) \leq & \mathrm{dist}^2(x^k, X^*) + 4[3d + (1+R)\bar{v}]d\alpha^k + 4Rd\bar{v} \\ & - \underline{\gamma}[2 - \bar{\gamma}(1+R)^2] \left(\frac{f(x^k) - f^* - \epsilon}{L}\right)^{\frac{2}{p}}.\end{aligned} \tag{2.21}$$

We consider two cases: $\epsilon = 0$, and $\epsilon > 0$.

Case (i). Let $\epsilon = 0$. It follows from (2.21) and Assumption 2.5.1 that, for any $k \in \mathbb{N}$,

$$\mathrm{dist}^2(x^{k+1}, X^*) \leq C_1 \mathrm{dist}^2(x^k, X^*) + C_2\alpha^k + C_3,$$

where $C_1 := 1 - \underline{\gamma}[2 - \bar{\gamma}(1+R)^2] \left(\frac{\rho}{L}\right)^{\frac{2}{p}}$, $C_2 := 4[3d + (1+R)\bar{v}]d$, and $C_3 := 4Rd\bar{v}$. If $C_1 \leq 0$, then, noting $\alpha^{k+1} \leq \zeta\alpha^k$, we have that, for any $k \in \mathbb{N}$,

$$\mathrm{dist}^2(x^{k+1}, X^*) \leq C_2\alpha^k + C_3 \leq \cdots \cdots \leq C_2\zeta^k\alpha^0 + C_3 = \frac{C_2\alpha^0}{\zeta}\zeta^{k+1} + C_3. \tag{2.22}$$

If $0 < C_1 < 1$, then we have that, for any $k \in \mathbb{N}$,

$$\begin{aligned} \mathrm{dist}^2(x^{k+1}, X^*) &\le C_1 \mathrm{dist}^2(x^k, X^*) + C_2\alpha^k + C_3 \\ &\le C_1[C_1 \mathrm{dist}^2(x^{k-1}, X^*) + C_2\alpha^{k-1} + C_3] + C_2\alpha^k + C_3 \\ &\cdots\ \cdots\ \cdots \\ &\le C_1^{k+1}\mathrm{dist}^2(x^0, X^*) + C_2\sum_{i=0}^{k} C_1^i \alpha^{k-i} + C_3\sum_{i=0}^{k} C_1^i. \end{aligned}$$

Let $\max\{C_1, \zeta\} < \tau < 1$. Then it follows from $\alpha^{k+1} \le \zeta\alpha^k$ that, for any $k \in \mathbb{N}$,

$$\begin{aligned} \mathrm{dist}^2(x^{k+1}, X^*) &\le C_1^{k+1}\mathrm{dist}^2(x^0, X^*) + C_2\alpha^0\sum_{i=0}^{k} C_1^i \tau^{k-i} + C_3\sum_{i=0}^{k} C_1^i \\ &\le \mathrm{dist}^2(x^{N+1}, X^*)\ \tau^{k+1} + \frac{C_2\alpha^0}{(\tau - C_1)}\ \tau^{k+1} + \frac{C_3}{1 - C_1}. \end{aligned} \tag{2.23}$$

From (2.22) and (2.23), the statement is true for this case.

Case (ii). Let $\epsilon > 0$. It follows from Lemma 2.5.2 and (2.21) that, for any $k \in \mathbb{N}$,

$$\begin{aligned} \mathrm{dist}^2(x^{k+1}, X^*) &\le \mathrm{dist}^2(x^k, X^*) + C_2\alpha^k + 4Rd\bar{v} \\ &\quad - \underline{\gamma}[2 - \bar{\gamma}(1+R)^2]\left[\tilde{c}\left(\frac{f(x^k) - f^*}{L}\right)^{\frac{2}{p}} - \left(\frac{\epsilon}{L}\right)^{\frac{2}{p}}\right], \end{aligned}$$

where $\tilde{c} := \min\{1, 2^{1-\frac{2}{p}}\}$. Then, combining with Assumption 2.5.1, we have that, for any $k \in \mathbb{N}$,

$$\mathrm{dist}^2(x^{k+1}, X^*) \le \overline{C}_1 \mathrm{dist}^2(x^k, X^*) + C_2\alpha^{k-1} + \overline{C}_3,$$

where $\overline{C}_1 := 1 - \tilde{c}\underline{\gamma}[2 - \bar{\gamma}(1+R)^2]\left(\frac{\rho}{L}\right)^{\frac{2}{p}}$ and $\overline{C}_3 := \underline{\gamma}[2 - \bar{\gamma}(1+R)^2]\left(\frac{\epsilon}{L}\right)^{\frac{2}{p}} + 4Rd\bar{v}$. The rest of the proof for this case can be completed similarly as for Case (i). □

Theorems 2.5.6(2), 2.5.7, and 2.5.8 present the convergence to a ball (tolerance region) of the optimal solution set, and D in the theorems represents the radius of the tolerance region. Table 2.1 lists the specific value of D in different cases.

Table 2.1: Radius of tolerance region (D)

Stepsize	$\epsilon = 0$	$\epsilon > 0$
Dimi.	$(2d)^{3-q/p}R\,(L/\rho)^{1/p}$	$\frac{(2d)^{3-q/p}R}{\tilde{c}}(L/\rho)^{1/p} + \frac{(2d)^{2-q/p}}{\tilde{c}}(\epsilon/\rho)^{1/p}$
Cons.	$\frac{4Rdv+(1+R)^2v^2}{\min\{1,2(2d)^{q/p-2}(\rho/L)^{1/p}v\}}$	$\frac{4Rdv+(1+R)^2v^2+2v(\epsilon/L)^{1/p}}{\min\{1,2(2d)^{q/p-2}(\rho/L)^{1/p}\tilde{c}v\}}$
Dyna.	$\frac{4Rd\bar{v}}{\min\{1,\underline{\gamma}[2-\bar{\gamma}(1+R)^2](\rho/L)^{2/p}\}}$	$\frac{4Rd\bar{v}+\underline{\gamma}[2-\bar{\gamma}(1+R)^2](\epsilon/L)^{2/p}}{\min\{1,\tilde{c}\underline{\gamma}[2-\bar{\gamma}(1+R)^2](\rho/L)^{2/p}\}}$

When $\alpha^k \equiv 0$, the results in Theorems 2.5.6-2.5.8 provide the rate of convergence without extrapolation. It is worth noting that both the convergence rate and tolerance region (see Table 2.1) of the inexact quasisubgradient method with extrapolation are the same as those without extrapolation. The following remark summarizes the rate of convergence in iterates for all the stepsizes.

Remark 2.5.9. *Theorem 2.5.6 indicates a sublinear convergence (see Remark 2.5.5), in particular,* $\mathcal{O}\left(1/k^s\right)$ *or* $\mathcal{O}\left(\tau^{k^s}\right)$ *for some* $0 < s < 1$ *and some* $0 < \tau < 1$. *Theorems 2.5.7 and 2.5.8 illustrate a linear convergence, in particular,* $\mathcal{O}\left(\tau^k\right)$ *for some* $0 < \tau < 1$.

2.6 Primal-Dual Inexact Quasisubgradient Method with Extrapolation

In this section, we study EiQSG in a primal-dual framework for finding a saddle point of a quasiconvex-quasiconcave function, inspired by the work of Hu et al. (2016).

A function $F(x, y)$ is said to be *quasiconvex-quasiconcave* if it is quasiconvex in x for a fixed y and quasiconcave in y for a fixed x. We say that (x^*, y^*) is a *saddle point* of F over $X \times Y$ if $(x^*, y^*) \in X \times Y$ and

$$F(x^*, y) \leq F(x^*, y^*) \leq F(x, y^*) \qquad \text{for all} \quad (x, y) \in X \times Y,$$

or equivalently,

$$\min_{x\in X}\max_{y\in Y}\; F(x,y)=F(x^*,y^*)=\max_{y\in Y}\min_{x\in X}\; F(x,y).$$

2.6.1 Method Proposal and Basic Properties

We restrict ourselves to finding a saddle point of F over $X\times Y$, where $F:\mathbb{R}^{m\times n}\to\mathbb{R}$ is a quasiconvex-quasiconcave function, and $X\subseteq\mathbb{R}^m$ and $Y\subseteq\mathbb{R}^n$ are nonempty, bounded, closed, and convex sets. We assume that there exists a saddle point of F over $X\times Y$. Then the considered saddle point problem can be expressed by the following minimax optimization problem

$$\min_{x\in X}\max_{y\in Y}\; F(x,y). \tag{2.24}$$

Notations ϵ, r^k, v^k, α^k, R, and $\mathbf{S}$ are used the same as in Section 2.2. Besides, F^* and $X^*\times Y^*$ represent the optimal value and optimal solution set of (2.24), respectively. As X and Y are bounded, then there exists $d>0$ such that $\|x\|\le d$ for all $x\in X$ and $\|y\|\le d$ for all $y\in Y$. Recall that the partial ϵ-quasisubdifferentials of F for x and y (see Subsection 1.5.2) are respectively defined by

$$\bar{\partial}^*_{x,\epsilon}F(\bar{x},\bar{y}):=\{g\in\mathbb{R}^m:\langle g,x-\bar{x}\rangle\le 0,\ \forall x \text{ with } F(x,\bar{y})<F(\bar{x},\bar{y})-\epsilon\}$$

and

$$\bar{\partial}^*_{y,\epsilon}F(\bar{x},\bar{y}):=\{g\in\mathbb{R}^n:\langle g,y-\bar{y}\rangle\ge 0,\ \forall y \text{ with } F(\bar{x},y)>F(\bar{x},\bar{y})+\epsilon\}.$$

Now, we propose the following primal-dual extrapolated inexact quasisubgradient method (Pd-EiQSG) for (2.24)

$$\begin{aligned} \hat{x}^k=x^k+\alpha^k(x^k-x^{k-1}),\quad \hat{y}^k=y^k-\alpha^k(y^k-y^{k-1}),\\ x^{k+1}=P_X\left(\hat{x}^k-v^k\widetilde{F}_x(x^k,y^k)\right),\quad y^{k+1}=P_Y\left(\hat{y}^k+v^k\widetilde{F}_y(x^k,y^k)\right), \end{aligned} \tag{2.25}$$

where $x^0 \in \mathbb{R}^m$, $x^{-1} = x^0$, $y^0 \in \mathbb{R}^n$, and $y^{-1} = y^0$ are initial points. Moreover,

$$\widetilde{F}_x(x^k, y^k) = F_x(x^k, y^k) + r^k \qquad \text{and} \qquad F_x(x^k, y^k) \in \bar{\partial}^*_{x,\epsilon} F(x^k, y^k) \cap \mathbf{S},$$

and

$$\widetilde{F}_y(x^k, y^k) = F_y(x^k, y^k) - r^k \qquad \text{and} \qquad F_y(x^k, y^k) \in \bar{\partial}^*_{y,\epsilon} F(x^k, y^k) \cap \mathbf{S}.$$

As in (2.2), the extrapolation parameter sequence $\{\alpha^k\}$ in (2.25) is again assumed to satisfy

$$0 \leq \alpha^k \leq 1 \qquad \text{and} \qquad \sum_{k=0}^{+\infty} \alpha^k < +\infty.$$

The boundedness of $\{r_k\}$ is also assumed, then there exists $R \geq 0$ such that $\|r^k\| \leq R$ for all $k \in \mathbb{N}$. For $\{v^k\}$, we focus on the diminishing and constant stepsizes:

- Diminishing stepsize: $v^k > 0$, $\lim_{k \to +\infty} v^k = 0$, and $\sum_{k=0}^{+\infty} v^k = +\infty$;
- Constant stepsize: $v^k \equiv v (> 0)$.

We need to point out that the parameters v^k, α^k, r^k, d, and ϵ can be separately selected for $\{x^k\}$ and $\{y^k\}$, and the corresponding theoretical results can still be obtained for this case. We discuss them uniformly for simplicity.

An assumption of the Hölder condition restricted to $X \times Y$ is made in analyzing the convergence of Pd-EiQSG (2.25).

Assumption 2.6.1. *Assume that F in (2.24) satisfies the Hölder condition restricted to $X \times Y$ of order $p > 0$ with modulus $L > 0$ on $\mathbb{R}^{m \times n}$, i.e.,*

$$|F(x, y) - F(\bar{x}, \bar{y})| \leq L \|(x, y) - (\bar{x}, \bar{y})\|^p$$

for all $(\bar{x}, \bar{y}) \in X \times Y$ and $(x, y) \in \mathbb{R}^{m \times n}$.

Lemma 2.6.2 is an extension of Lemma 2.2.2. With f in place of h and X^* in place of D, Lemma 2.6.2 reduces to Lemma 2.2.2.

Lemma 2.6.2. *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a quasiconvex function and $D \in \mathbb{R}^n$ be a closed and convex set. Suppose that h satisfies the Hölder condition restricted to D of order $p > 0$ with modulus $L > 0$ on $\mathbb{R}^n$, i.e.,*

$$|h(x) - h(\bar{x})| \leq L\|x - \bar{x}\|^p \qquad \text{for all} \quad \bar{x} \in D \text{ and } x \in \mathbb{R}^n.$$

Let $\bar{x} \in D$, $x \in \mathbb{R}^n$, and $\zeta \geq 0$ satisfy $h(x) > h(\bar{x}) + L\zeta^p + \epsilon$ and let $g(x) \in \bar{\partial}_\epsilon^* h(x) \cap \mathbf{S}$. Then $\langle g(x), x - \bar{x} \rangle \geq \zeta$.

Proof. From the Hölder condition, for any $y \in \mathbf{B}(\bar{x}, \zeta) := \{y : \|y - \bar{x}\| \leq \zeta\}$, one has that

$$h(y) - h(\bar{x}) \leq L \|y - \bar{x}\|^p \leq L\zeta^p < h(x) - h(\bar{x}) - \epsilon.$$

That is, $y \in \mathrm{lev}_{<h(x)-\epsilon}$ holds for all $y \in \mathbf{B}(\bar{x}, \zeta)$. Then, $\bar{x} + \zeta g(x) \in \mathrm{lev}_{<h(x)-\epsilon}$ can be obtained by virtue of $\bar{x} + \zeta g(x) \in \mathbf{B}(\bar{x}, \zeta)$. As $g(x)$ is a quasisubgradient, we have that $\langle g(x), \bar{x} + \zeta g(x) - x \rangle \leq 0$, which implies the desired estimate. □

Introducing the extrapolation steps, we derive extended basic inequalities for $\{x^k\}$ and $\{y^k\}$, respectively.

Lemma 2.6.3. *Let $\{x^k\}$ and $\{y^k\}$ be the sequences generated by Pd-EiQSG (2.25). Then, for any $k \in \mathbb{N}$ and $(x, y) \in X \times Y$, one has that*

$$\begin{aligned} \|x^{k+1} - x\|^2 \leq & \|x^k - x\|^2 + 12d^2\alpha^k + 4(1+R)dv^k\alpha^k \\ & - 2v^k \left\langle F_x(x^k, y^k), x^k - x \right\rangle + 4Rdv^k + \left[(1+R)v^k\right]^2, \end{aligned}$$

and

$$\begin{aligned} \|y^{k+1} - y\|^2 \leq & \|y^k - y\|^2 + 12d^2\alpha^k + 4(1+R)dv^k\alpha^k \\ & + 2v^k \left\langle F_y(x^k, y^k), y^k - y \right\rangle + 4Rdv^k + \left[(1+R)v^k\right]^2. \end{aligned}$$

Proof. For the first estimate, by use of the nonexpansive property of the projection operator and (2.25), we obtain that, for any $k \in \mathbb{N}$ and $x \in X$,

$$
\begin{aligned}
\|x^{k+1} - x\|^2 \leq & \|\hat{x}^k - v^k \widetilde{F}_x(x^k, y^k) - x\|^2 \\
= & \|x^k - x\|^2 + \|\hat{x}^k - x^k\|^2 + 2\langle x^k - x, \hat{x}^k - x^k\rangle - 2v^k \langle \widetilde{F}_x(x^k, y^k), \hat{x}^k - x^k\rangle \\
& - 2v^k \langle \widetilde{F}_x(x^k, y^k), x^k - x\rangle + (\|\widetilde{F}_x(x^k, y^k)\| v^k)^2.
\end{aligned}
$$

Then, it follows from the Cauchy-Schwarz inequality, boundedness of X and $\{\alpha_k\}$, and (2.25) that, for any $k \in \mathbb{N}$ and $x \in X$,

$$
\begin{aligned}
\|x^{k+1} - x\|^2 \leq & \|x^k - x\|^2 + (\|x^k - x^{k-1}\|\alpha^k)^2 + 2\|x^k - x\|\|x^k - x^{k-1}\|\alpha^k \\
& + 2(1+R)\|x^k - x^{k-1}\| v^k \alpha^k - 2v^k \langle F_x(x^k, y^k), x^k - x\rangle \\
& + 2R\|x^k - x\| v^k + (\|\widetilde{F}_x(x^k, y^k)\| v^k)^2 \\
\leq & \|x^k - x\|^2 + 12d^2\alpha^k + 4(1+R)dv^k\alpha^k \\
& - 2v^k \langle F_x(x^k, y^k), x^k - x\rangle + 4Rdv^k + \left[(1+R)v^k\right]^2,
\end{aligned}
$$

which completes the proof of the first inequality. The second inequality can be proved similarly. □

2.6.2 Convergence in Objective Values

The convergence result in objective values for the primal-dual method is presented in the following theorem. Recalling (2.3), we again consider a unified structure of the diminishing and constant stepsizes.

Theorem 2.6.4. *Let $\{x^k\}$ and $\{y^k\}$ be the sequences generated by Pd-EiQSG (2.25) with the stepsize (2.3) and Assumption 2.6.1 hold. Then*

$$
\limsup_{k \to +\infty} F(x^k, y^k) \geq F^* - L\left(\frac{v}{2}(1+R)^2 + 2Rd\right)^p - \epsilon,
$$

and

$$\liminf_{k\to+\infty} F(x^k, y^k) \leq F^* + L\left(\frac{v}{2}(1+R)^2 + 2Rd\right)^p + \epsilon.$$

Proof. For the first estimate, we assume by contradiction that

$$\limsup_{k\to+\infty} F(x^k, y^k) < F^* - L\left(\frac{v}{2}(1+R)^2 + 2Rd\right)^p - \epsilon.$$

Then, there exist $\delta > 0$ and $k_0 \in \mathbb{N}$ such that

$$F(x^k, y^k) < F^* - L\left(\frac{v}{2}(1+R)^2 + 2Rd + \delta\right)^p - \epsilon \quad \text{for all} \quad k \geq k_0.$$

By the definition of the saddle point, we have that, for any $k \geq k_0$ and $y^* \in Y^*$,

$$F(x^k, y^k) < F(x^k, y^*) - L\left(\frac{v}{2}(1+R)^2 + 2Rd + \delta\right)^p - \epsilon.$$

It follows from Lemma 2.6.2 and the quasiconvexity of $-F(\cdot, y)$ that, for any $k \geq k_0$ and $y^* \in Y^*$,

$$\left\langle -F_y(x^k, y^k), y^k - y^* \right\rangle \geq \frac{v}{2}(1+R)^2 + 2Rd + \delta. \tag{2.26}$$

As $\lim_{k\to+\infty} v^k = v$ and $\lim_{k\to+\infty} \alpha^k = 0$, there exists $k_1 \in \mathbb{N}$ such that

$$v^k \leq v + \frac{\delta}{2(1+R)^2} \quad \text{and} \quad \alpha^k \leq \frac{\delta}{4d(1+R)} \qquad \text{for all} \quad k \geq k_1. \tag{2.27}$$

Together with (2.26), (2.27), and Lemma 2.6.3, we have that, for any $k \geq k_2 := \max\{k_0, k_1\}$ and $y^* \in Y^*$,

$$\begin{aligned}\|y^{k+1} - y^*\|^2 \leq & \|y^k - y^*\|^2 + 12d^2\alpha^k + 4(1+R)dv^k\alpha^k \\ & - 2v^k\left(\frac{v}{2}(1+R)^2 + 2Rd + \delta\right) + 4Rdv^k + \left[(1+R)v^k\right]^2 \\ \leq & \|y^k - y^*\|^2 + 12d^2\alpha^k - \frac{\delta}{2}v^k,\end{aligned}$$

where, in the right-hand side of the first inequality, the second α^k and one v^k in $\left[(1+R)v^k\right]^2$ were respectively replaced by estimates in (2.27). For $n > k_2$, summing the left-hand side and right-hand side of the above inequality over $k = k_2, \cdots, n$, we obtain that

$$\|y^{n+1} - y^*\|^2 \le \|y^{k_2} - y^*\|^2 + 12d^2 \sum_{k=k_2}^{n} \alpha^k - \frac{\delta}{2} \sum_{k=k_2}^{n} v^k,$$

in contradiction when $n \to +\infty$ to the facts that $\sum_{k=0}^{+\infty} \alpha^k < +\infty$ and $\sum_{k=0}^{+\infty} v^k = +\infty$. Thus, the proof of the first inequality is completed. The proof of the second inequality can be completed similarly. □

When $v = 0$, the following estimates are the convergence result in objective values for the diminishing stepsize

$$\liminf_{k\to+\infty} F(x^k, y^k) - L(2Rd)^p - \epsilon \le F^* \le \limsup_{k\to+\infty} F(x^k, y^k) + L(2Rd)^p + \epsilon.$$

When $v^k \equiv v > 0$, the estimates in Theorem 2.6.4 are the convergence result in objective values for the constant stepsize.

2.6.3 Iteration Complexity

Now, we explore the computational complexity. For this purpose, we respectively record the minimum and maximum objective values at the K^{th} iteration as

$$F_{min}(x^K, y^K) := \min_{1\le k\le K} F(x^k, y^k)$$

and

$$F_{max}(x^K, y^K) := \max_{1\le k\le K} F(x^k, y^k).$$

We discuss the two stepsizes together again and focus on the special structure (2.4), as in Theorem 2.4.1.

Theorem 2.6.5. *Let $\delta > 0$ and $0 < \eta < \frac{2}{3}$. Let $\{x^k\}$ and $\{y^k\}$ be the sequences generated by Pd-EiQSG (2.25) with the stepsize (2.4) and extrapolation rule $\alpha^k = o(v^k)$ and Assumption 2.6.1 hold. Then there exists $\bar{K} \in \mathbb{N}$ such that*

$$F^* - F_{max}(x^K, y^K) \leq L\left(\frac{v}{2}(1+R)^2 + 2Rd + \delta\right)^p + \epsilon,$$

and

$$F_{min}(x^K, y^K) - F^* \leq L\left(\frac{v}{2}(1+R)^2 + 2Rd + \delta\right)^p + \epsilon,$$

where K is the minimum integer such that

$$(2-3\eta)\delta\left((K-\bar{K}+1)v + \frac{c[(K+1)^{1-s} - \bar{K}^{1-s}]}{1-s}\right)$$
$$> \max\left\{\text{dist}^2\left(x^{\bar{K}}, X^*\right), \text{dist}^2\left(y^{\bar{K}}, Y^*\right)\right\}.$$

Proof. For the first estimate, we assume by contradiction that, for any $1 \leq k \leq K$,

$$F^* - F(x^k, y^k) > L\left(\frac{v}{2}(1+R)^2 + 2Rd + \delta\right)^p + \epsilon.$$

By the definition of the saddle point, we have that, for any $1 \leq k \leq K$ and $y^* \in Y^*$,

$$F(x^k, y^*) - F(x^k, y^k) > L\left(\frac{v}{2}(1+R)^2 + 2Rd + \delta\right)^p + \epsilon.$$

It follows from Lemma 2.6.2 and the quasiconvexity of $-F(\cdot, y)$ that, for any $1 \leq k \leq K$ and $y^* \in Y^*$,

$$\left\langle -F_y(x^k, y^k), y^k - y^* \right\rangle \geq \frac{v}{2}(1+R)^2 + 2Rd + \delta. \tag{2.28}$$

Given $\delta > 0$ and $0 < \eta < \frac{2}{3}$, it is easy to see that there exists $\bar{K} \in \mathbb{N}$ such that

$$v^k \leq v + \frac{\eta\delta}{(1+R)^2} \quad \text{for all} \quad k \geq \bar{K}, \tag{2.29}$$

and

$$\alpha^k \leq \min\left\{ \frac{\eta\delta v^k}{12d^2} , \frac{\eta\delta}{4(1+R)d} \right\} \quad \text{for all} \quad k \geq \bar{K}. \tag{2.30}$$

Together with (2.28), (2.29), and Lemma 2.6.3, we have that, for any $\bar{K} \leq k \leq K$ and $y^* \in Y^*$,

$$\begin{aligned}
\|y^{k+1} - y^*\|^2 \leq & \|y^k - y^*\|^2 + 12d^2\alpha^k + 4(1+R)dv^k\alpha^k \\
& + 2v^k \left\langle F_y(x^k, y^k), y^k - y \right\rangle + 4Rdv^k + \left[(1+R)v^k\right]^2 \\
\leq & \|y^k - y^*\|^2 + 12d^2\alpha^k + 4(1+R)dv^k\alpha^k - 2v^k\left(\frac{v}{2}(1+R)^2 + 2Rd + \delta\right) \\
& + 4Rdv^k + (1+R)^2v^k\left(v + \frac{\eta\delta}{(1+R)^2}\right) \\
= & \|y^k - y^*\|^2 + 12d^2\alpha^k + 4(1+R)dv^k\alpha^k - 2\delta v^k + \eta\delta v^k.
\end{aligned}$$

Now by (2.30), for any $\bar{K} \leq k \leq K$ and $y^* \in Y^*$, it holds that

$$\|y^{k+1} - y^*\|^2 \leq \|y^k - y^*\|^2 - (2-3\eta)\delta v^k = \|y^k - y^*\|^2 - (2-3\eta)\delta\left(v + ck^{-s}\right).$$

Summing the above inequality over $k = \bar{K}, \cdots, K$, we obtain that

$$(2-3\eta)\delta \sum_{k=\bar{K}}^{K} \left(v + ck^{-s}\right) \leq \|y^{\bar{K}} - y^*\|^2 - \|y^{K+1} - y^*\|^2 \leq \|y^{\bar{K}} - y^*\|^2. \tag{2.31}$$

Since k^{-s} decreases as k increases, we see that

$$\sum_{k=\bar{K}}^{K} k^{-s} \geq \int_{\bar{K}}^{K+1} t^{-s} \mathrm{d}t = \frac{(K+1)^{1-s} - \bar{K}^{1-s}}{1-s}.$$

Let $y^* = P_{Y^*}(y^{\bar{K}})$. Then, it follows from (2.31) that

$$(2-3\eta)\delta\left((K - \bar{K} + 1)v + \frac{c[(K+1)^{1-s} - \bar{K}^{1-s}]}{1-s}\right) \leq \operatorname{dist}^2(y^{\bar{K}}, Y^*),$$

in contradiction to the definition of K. Thus, the proof of the first inequality is completed. The proof of the second inequality can be completed similarly. □

When $v = 0$ and $c > 0$, the following estimates provide the tolerance level for the diminishing stepsize $v^k = ck^{-s}$

$$F^* - F_{max}(x^K, y^K) \leq L\,(2Rd + \delta)^p + \epsilon$$

and

$$F_{min}(x^K, y^K) - F^* \leq L\,(2Rd + \delta)^p + \epsilon.$$

Moreover, if $\alpha^k = k^{-t}$ (for each $t > 1$), from (2.29) and (2.30), the specific forms of $\bar{K}$ and K are respectively given as

$$\bar{K} := \max\left\{ \left\lceil \left(\frac{12d^2}{\eta c\delta}\right)^{\frac{1}{t-s}} \right\rceil, \left\lceil \left(\frac{4(1+R)d}{\eta\delta}\right)^{\frac{1}{t}} \right\rceil, \left\lceil \left(\frac{c(1+R)^2}{\eta\delta}\right)^{\frac{1}{s}} \right\rceil \right\}$$

and

$$K := \left\lceil \left(\bar{K}^{1-s} + \frac{(1-s)\max\left\{\operatorname{dist}^2(x^{\bar{K}}, X^*), \operatorname{dist}^2(y^{\bar{K}}, Y^*)\right\}}{(2-3\eta)c\delta} \right)^{\frac{1}{1-s}} \right\rceil.$$

When $v > 0$ and $c = 0$, the estimates in Theorem 2.6.5 provide the tolerance level for the constant stepsize $v^k \equiv v$. Moreover, if $\alpha^k = k^{-t}$ (for each $t > 1$), from (2.29) and (2.30), the specific forms of $\bar{K}$ and K are respectively given as

$$\bar{K} : = \max\left\{ \left\lceil \left(\frac{12d^2}{\eta v\delta}\right)^{\frac{1}{t}} \right\rceil, \left\lceil \left(\frac{4(1+R)d}{\eta\delta}\right)^{\frac{1}{t}} \right\rceil \right\}$$

and

$$K : = \left\lceil \bar{K} + \frac{\max\left\{\operatorname{dist}^2(x^{\bar{K}}, X^*), \operatorname{dist}^2(y^{\bar{K}}, Y^*)\right\}}{(2-3\eta)v\delta} \right\rceil.$$

2.6.4 An Application

In convex programming, minimizing a convex function over a closed and convex set, under some mild constraint qualifications, is equivalent to finding the saddle points of a convex-concave Lagrangian function. Inspired by this property, we provide a class of optimization problems, where the strong duality holds for a modified quasiconvex-quasiconcave Lagrangian function, as an application of Pd-EiQSG (2.25).

We consider the following problem

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.t.} \quad & h_i(x) \leq 0, \; i = 1, \cdots, l \\ & x \in X, \end{aligned} \tag{2.32}$$

where $f, h_i \; (i = 1, \cdots, l) : \mathbb{R}^n \to \mathbb{R}$, and $X \subseteq \mathbb{R}^n$ is a closed and convex set.

Without the convexity of (2.32), the strong duality is not guaranteed for the classical Lagrangian function. Thus, we focus on a modified Lagrangian function of (2.32) instead, i.e.,

$$\mathcal{L}_+(x, \mu) := f(x) + \sum_{i=1}^{l} \mu_i \left(h_i(x)\right)_+ \qquad \textit{with} \quad (x, \mu) \in X \times \mathbb{R}_+^l,$$

where $(h_i(x))_+ := \max\{h_i(x), 0\}$ $(i = 1, \cdots, l)$, $\mu := (\mu_1, \cdots, \mu_l)$ is the multiplier, and $\mathbb{R}_+^l$ denotes the set of l-dimensional nonnegative real vectors.

The zero duality gap (or strong duality) for $\mathcal{L}_+(x, \mu)$ has been demonstrated in Rubinov et al. (2002) (see Proposition 2.6.6).

Proposition 2.6.6. *Let (x^*, μ^*) be a saddle point of $\mathcal{L}_+$ over $X \times \mathbb{R}_+^l$. Then x^* is an optimal solution of* (2.32).

It is clear that $\mathcal{L}_+(x, \mu)$ is linear in μ for a fixed x. Then, if $\mathcal{L}_+(x, \mu)$ is quasiconvex in x for a fixed μ, Pd-EiQSG (2.25) can be adopted on $\mathcal{L}_+(x, \mu)$, thus solves (2.32). For example, when f has some fractional form, the modified Lagrangian function can be formulated as a quasiconvex-quasiconcave function.

Suppose that $h_i(x)$ in (2.32) is a convex function, and $f(x)$ in (2.32) a quasiconvex function given by $f(x) := \frac{c(x)}{d(x)}$, where $c(x)$ is a convex function, and $d(x)$ is a concave function. We also assume that $c(x) \geq 0$ and $d(x) > 0$ for all $x \in X$. As $h_i(x) \leq 0$ is identical to $\frac{h_i(x)}{d(x)} \leq 0$ for all $x \in X$, the modified Lagrangian function of an equivalent formulation of (2.32) is written as

$$\widetilde{\mathcal{L}}_+(x, \mu) := \frac{c(x)}{d(x)} + \sum_{i=1}^{l} \mu_i \frac{(h_i(x))_+}{d(x)} \qquad \textit{with} \quad (x, \mu) \in X \times \mathbb{R}^p_+.$$

It follows from Stancu-Minasian (2012, Section 2.5) that $\widetilde{\mathcal{L}}_+(x, \cdot)$ is quasiconvex. Since the convexity or quasiconvexity is not necessary for Proposition 2.6.6, the zero duality gap property is still satisfied for $\widetilde{\mathcal{L}}_+(x, \mu)$. Therefore, a global solution of (2.32) can be obtained by applying Pd-EiQSG (2.25) to $\widetilde{\mathcal{L}}_+(x, \mu)$.

2.7 Numerical Experiments

This section specifies EiQSG (2.2) to the Cobb-Douglas production efficiency model and portfolio selection model. The testing of the Cobb-Douglas model is conducted with data randomly generated, while the experiment of the portfolio model is carried out with data from the real-world market. As we focus on the effect of the extrapolation parameter α^k, we let $\epsilon = R = 0$ in the testing. All the experiments are completed in MATLAB R2021a and macOS 11.6 on a 64-bit PC with an i5-5250U CPU and 4GB RAM.

2.7.1 Cobb-Douglas Efficiency Production Model

Given n production factors, the Cobb-Douglas production efficiency model (see Bradley & Frey Jr (1974)), maximizing profit over cost, is written as

$$
\begin{aligned}
\max_{x} \quad & \frac{s_0 \prod_{j=1}^{n} x_j^{s_j}}{\sum\limits_{j=1}^{n} t_j x_j + t_0} \\
\text{s.t.} \quad & \sum_{j=1}^{n} a_{ij} x_j \geq b_i, \;\; i = 1, \cdots, m \\
& x \geq 0,
\end{aligned}
$$

where $x := (x_1, \cdots, x_n)$ is a vector of quantities for production factors, and $s_j > 0$ with $\sum\limits_{i=1}^{n} s_j = 1$ and $t_j > 0$ $(j = 0, 1, \cdots, n)$ are profit and cost coefficients for the production factor j, respectively. The constraints include the budget balance and other general conditions in the production problem. This model has been used as a quasiconvex optimization test problem in Hu et al. (2015).

Following Hu et al. (2015), we select the coefficients from data randomly generated under the normal distribution within the following intervals

$$
s_0, t_0, t_j \in [0, 10], \quad s_j, a_{ij} \in [0, 1], \quad \text{and} \quad b_i \in [0, \frac{n}{2}].
$$

The initial point is taken as a vector of 10's.

Let the extrapolation rule be set as

$$
\alpha^k = \frac{\beta}{1 + 0.1k^2}.
$$

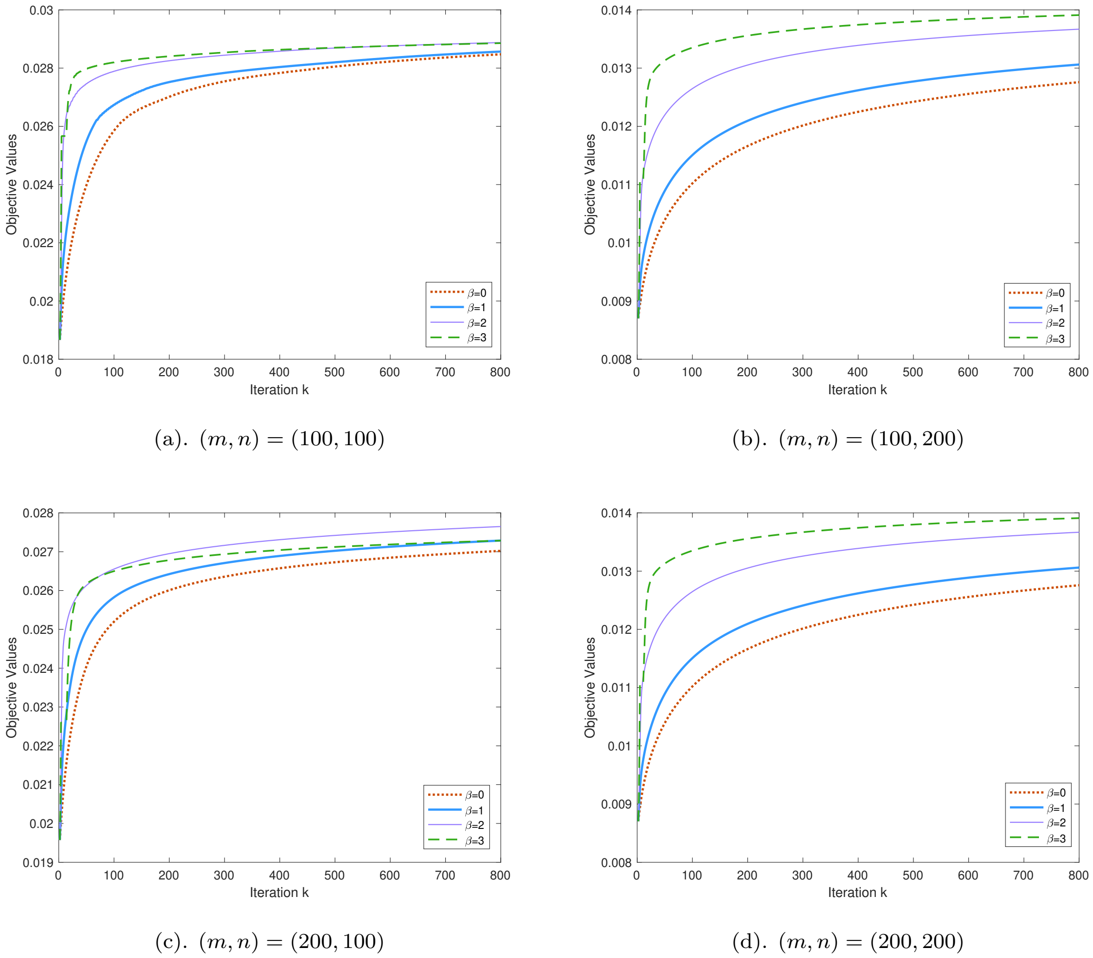
Table 2.2: Computational time for Cobb-Douglas model

Setting	$m = 100,\ n = 100$			$m = 100,\ n = 200$			$m = 200,\ n = 100$			$m = 200,\ n = 200$		
Stepsize	Dimi.	Cons.	Dyna.	Dimi.	Cons.	Dyna.	Dimi.	Cons.	Dyna.	Dimi.	Cons.	Dyna.
$\beta = 0$	26s	30s	28s	40s	44s	41s	58s	68s	63s	105s	93s	96s
$\beta = 1$	27s	27s	28s	41s	44s	43s	62s	72s	62s	103s	97s	96s
$\beta = 2$	31s	27s	27s	41s	42s	45s	63s	66s	67s	102s	97s	103s
$\beta = 3$	32s	27s	27s	46s	43s	43s	66s	68s	66s	108s	101s	104s

The diminishing, constant, and dynamic stepsizes are respectively chosen as

$$v^k = \frac{2}{1+0.1k}, \quad v^k \equiv 0.5, \quad \text{and} \quad v^k = \frac{f(x^k)-f^*}{0.01}.$$

For the dynamic stepsize, the (approximate) optimal value is estimated by the best results of the diminishing and constant stepsizes; specifically, $f^* \approx 0.036$, 0.015, 0.029, and 0.014 for the respective problem sizes. Since the projection is adopted, the resulting solutions are feasible, thus the obtained approximate optimal value is larger than the true optimal value. We consider that $p = 1$ and $L = 0.01$, where the



(a). $(m,n) = (100, 100)$

(b). $(m,n) = (100, 200)$

(c). $(m,n) = (200, 100)$

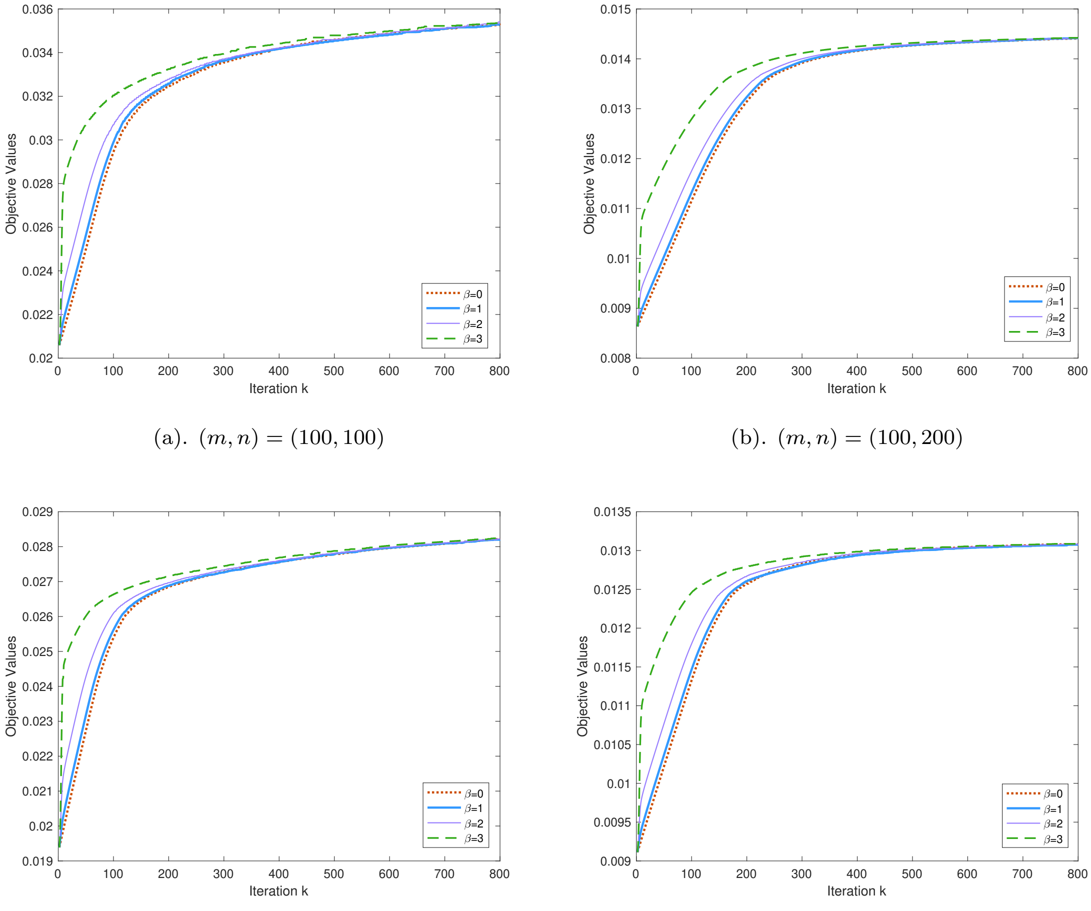
(d). $(m,n) = (200, 200)$

Figure 2.1: Convergence behavior (diminishing stepsize)

theoretical requirement (Assumption 2.6.1) is ignored. In fact, the convergence rate with p and L satisfying the theoretical requirement is very slow. Thus, to speed up the method, we neglect the assumption.

We test $\beta = 0$, 1, 2, and 3 for all the stepsizes on different data sets, where $(m, n) = (100, 100)$, $(100, 200)$, $(200, 100)$, and $(200, 200)$. The case $\beta = 0$ corresponds to the original method without extrapolation. The method terminates after 800 iterations.

The computational time for the different stepsizes and extrapolation parameters



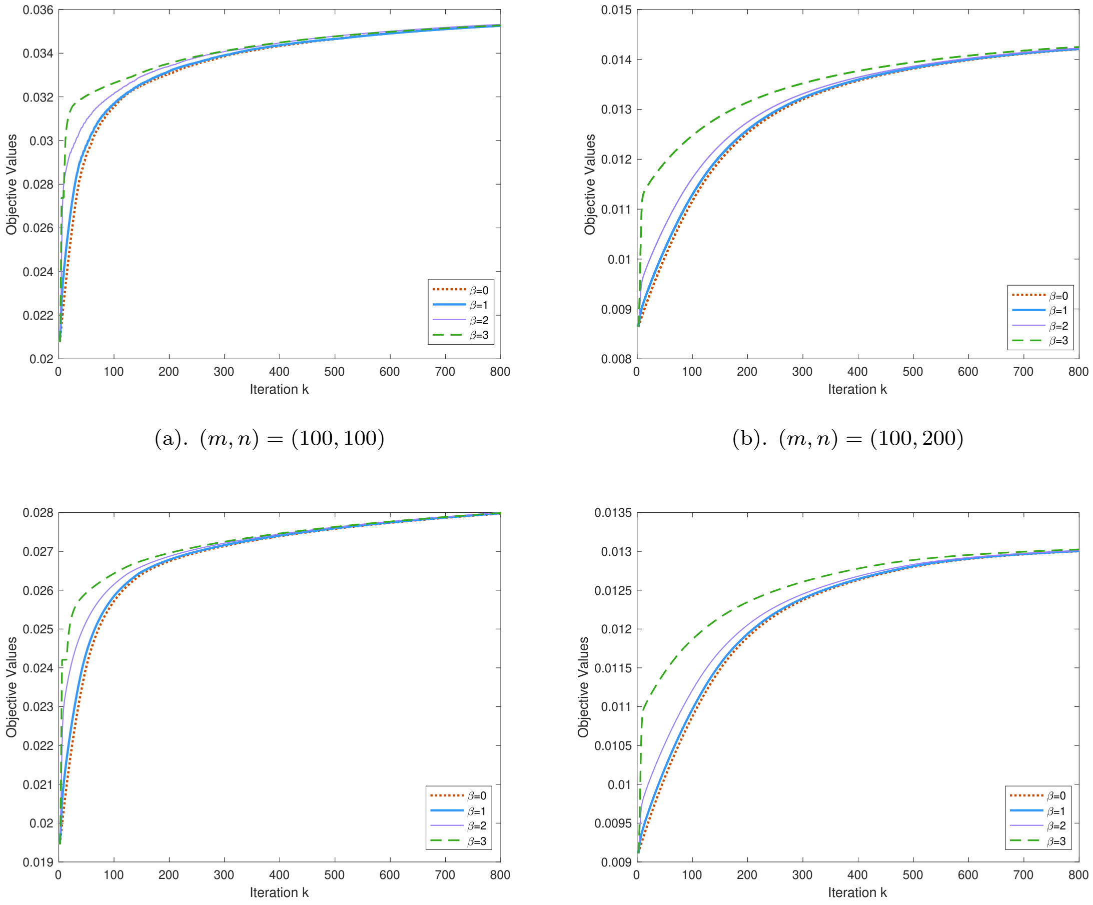
(a). $(m, n) = (100, 100)$

(b). $(m, n) = (100, 200)$

(c). $(m, n) = (200, 100)$

(d). $(m, n) = (200, 200)$

Figure 2.2: Convergence behavior (constant stepsize)



(a). $(m, n) = (100, 100)$

(b). $(m, n) = (100, 200)$

(c). $(m, n) = (200, 100)$

(d). $(m, n) = (200, 200)$

Figure 2.3: Convergence behavior (dynamic stepsize)

under the four problem sizes is presented in Table 2.2. As we can see, the choice of the extrapolation rule has no influence on the running time for all the cases.

The objective values against the number of iterations for different extrapolation rules are plotted in Figures 2.1-2.3. Figures 2.1-2.3 indicate that the extrapolation strategy accelerates the convergence of the method. In general, the method with $\beta = 3$ outperforms the others in terms of the number of iterations needed for reaching an approximate optimal solution. However, it is not necessary that the larger extrapolation parameter, the better performance is. For example, in Figure 2.1(c), the selection of $\beta = 2$ produces a superior result. This phenomenon is also observed in Pock & Sabach (2016).

2.7.2 Portfolio Selection Model

In the second experiment, we consider a portfolio selection model, which combines the classical Sharpe ratio maximization model and a generalized Sharpe ratio maximization model, and the minimum variance model.

Given n stocks with the risk-free rate $r_f \in \mathbb{R}$, expected rate of return vector $r := (r_1, \cdots, r_n)$, and covariance matrix $\Sigma \subseteq \mathbb{R}^{n\times n}$, the classical Sharpe ratio (see Sharpe (1966)) of the portfolio $w := (w_1, \cdots, w_n)$ is defined by

$$\frac{r^{\mathrm{T}}w - r_f}{\sqrt{w^{\mathrm{T}}\Sigma w}}\ .$$

The numerator $r^{\mathrm{T}}w - r_f$ represents the expected excess return, and the denominator $\sqrt{w^{\mathrm{T}}\Sigma w}$ represents the risk. On the other hand, Cai et al. (2000) proposed an l_∞ function as an alternative risk measure with the structure $\max\limits_{i=1,\cdots,n} E(|R_i w_i - r_i w_i|)$, where $R := (R_1, \cdots, R_n)$ is the rate of return vector of stocks, and $E(\cdot)$ denotes the mathematical expectation. Let $q_i := E(|R_i - r_i|)$, $i = 1 \cdots, n$. Then a generalized Sharpe ratio based on the l_∞ risk function is given by

$$\frac{r^{\mathrm{T}}w - r_f}{\max\limits_{i=1,\cdots,n} q_i w_i}\ .$$

The Sharpe ratio maximization model is to find a portfolio by maximizing the selected Sharpe ratio subject to some general constraints. Apart from it, the minimum variance model (see Luenberger (2013)), which uses the variance $w^{\mathrm{T}}\Sigma w$ to measure the risk and seeks the minimum risk, is also a popular portfolio selection model.

On the basis of the models mentioned above, we consider the following minimax portfolio selection model

$$\begin{array}{ll} \min\limits_{w} & \max\left\{\mu_1\,\dfrac{\sqrt{w^{\mathrm{T}}\Sigma w}}{r^{\mathrm{T}}w - r_f},\ \mu_2\,\dfrac{\max\limits_{i=1,\cdots,n}\ q_i w_i}{r^{\mathrm{T}}w - r_f},\ w^{\mathrm{T}}\Sigma w\right\} \\ \text{s.t.} & \mathbf{1}^{\mathrm{T}}w = 1,\ w \geq 0,\ r^{\mathrm{T}}w - r_f \geq 0, \end{array}$$

where $\mathbf{1}$ is the vector of ones and $\mu_1 > 0$ and $\mu_2 > 0$ are the normalization parameters so chosen such that the three terms are almost equal at an approximate solution. The first constraint $\mathbf{1}^{\mathrm{T}}w = 1$ represents that the budget is fully invested, the second one $w \geq 0$ means that short selling is not allowed in the investment, and $r^{\mathrm{T}}w - r_f \geq 0$ is an additional constraint to guarantee the nonnegativity of the excess return (a necessary assumption for the Sharpe ratio, see Bacon (2008)). From Stancu-Minasian (2012, Section 2.5), this model is a nonsmooth quasiconvex optimization problem.

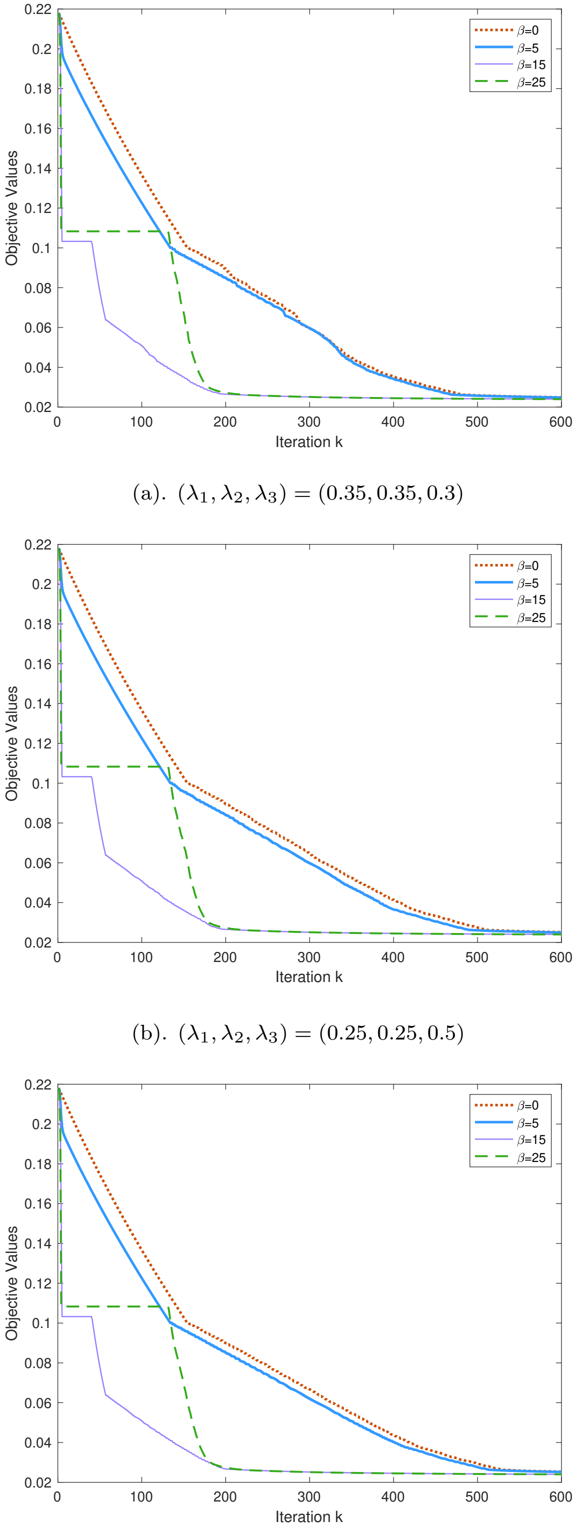
We carry out the experiment with $r_f = 0.01$ and the weekly historical data of 100 stocks from the Hang Seng stock market from 4 January 2010 to 31 December 2019. Besides, we select $\mu_1 = 0.08$ and $\mu_2 = 0.01$. The initial point is taken as $(1, 0, \cdots, 0)$.

As the results of all the stepsizes are similar, we only present the result of constant stepsize as the representative. Let the extrapolation rule and constant stepsize be respectively set as

$$\alpha^k = \frac{\beta}{k^2} \quad \text{and} \quad v^k \equiv 0.01.$$

The subgradient of the nonsmooth term is calculated as a convex combination of the gradients of three functions in the 'max' term with the corresponding convex combination parameters $\lambda_1 \geq 0, \lambda_2 \geq 0$, and $\lambda_3 \geq 0$ satisfying $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

We test $\beta = 0, 5, 15$, and 25 with $(\lambda_1, \lambda_2, \lambda_3) = (0.35, 0.35, 0.3)$, $(0.25, 0.25, 0.5)$, and $(0.15, 0.15, 0.7)$. The case $\beta = 0$ corresponds to the original method without extrapolation. The method terminates after 600 iterations.



(a). $(\lambda_1, \lambda_2, \lambda_3) = (0.35, 0.35, 0.3)$

(b). $(\lambda_1, \lambda_2, \lambda_3) = (0.25, 0.25, 0.5)$

(c). $(\lambda_1, \lambda_2, \lambda_3) = (0.15, 0.15, 0.7)$

Figure 2.4: Convergence behavior (constant stepsize)

The objective values against the number of iterations for different extrapolation and combination parameters are plotted in Figure 2.4. Similar patterns of convergence in objective values are observed for different combinations of $(\lambda_1, \lambda_2, \lambda_3)$. From Figure 2.4, the method with $\beta = 15$ performs best and achieves a good approximate optimal value within about 200 iterations, while the one with $\beta = 0$ or 5 needs over 500 iterations. We conclude that the objective value changes significantly for different extrapolation rules, but the choice of the weighted gradients among the functions in the 'max' term has a minor effect on the change for the objective value.

Chapter 3

Penalty Proximal ADMM Method with Extrapolation

3.1 Introduction

In this chapter, we consider the generalized bilinear programming (GBLP) problem proposed by Al-Khayyal (1992). Let $c_j \in \mathbb{R}^m$, $d_j \in \mathbb{R}^n$, $M_j \in \mathbb{R}^{m\times n}$ $(j = 0, 1, \cdots, p)$, $e_i \in \mathbb{R}$ $(i = 1, \cdots, p)$, $A \in \mathbb{R}^{l_1\times m}$, $B \in \mathbb{R}^{l_2\times n}$, $u \in \mathbb{R}^{l_1}$, and $v \in \mathbb{R}^{l_2}$. The GBLP problem is formulated as follows:

$$\begin{aligned} \min_{x,y} \quad & c_0^{\mathrm{T}} x + x^{\mathrm{T}} M_0 y + d_0^{\mathrm{T}} y \\ \text{s.t.} \quad & c_i^{\mathrm{T}} x + x^{\mathrm{T}} M_i y + d_i^{\mathrm{T}} y + e_i = 0,\ \ i = 1, \cdots, p \\ & Ax \le u,\ \ By \le v, \end{aligned} \tag{3.1}$$

where $x \in \mathbb{R}^m$, and $y \in \mathbb{R}^n$ are variables.

For solving (3.1), we provide an algorithm based on an ADMM scheme with extrapolation and a penalty relaxation. More precisely, we first compose an extrapolated proximal ADMM method as an inner algorithm to solve a quadratic penalty problem and then apply an update of the associated penalty parameter as an outer algorithm. The same inner-and-outer framework is also found in Lu (2014a) for an iterative hard thresholding method. The inner algorithm is constructed with extrapolation. To the best of our knowledge, the extrapolation strategy has been little adopted on ADMM-type methods except for Chen et al. (2015), where extrapolated ADMM algorithms are proposed for a class of linearly constrained convex problems. As the extrapolation step is included, a potential function instead of the augmented Lagrangian function is used to explore the convergence properties of the inner algorithm. By assuming that the related parameters satisfy a system of inequalities, we derive a descent property of the potential function, a dominant tool in our analysis, as it is the case for all nonconvex ADMM-type methods.

We establish the subsequential convergence, iteration complexity, and global convergence for the inner extrapolated ADMM method. The subsequential convergence, which means that any limit point of the iterate sequence is a stationary point, is obtained by virtue of the basic descent property. The iteration complexity is shown to be $\mathcal{O}(1/k)$, which is the same as some nonextrapolated convex and nonconvex ADMM methods in He & Yuan (2012, 2015), and Hong et al. (2016) and better than $\mathcal{O}(1/\sqrt{k})$ for extrapolated convex ADMM methods in Chen et al. (2015). Moreover, the global convergence is studied with the Kurdyka-Łojasiewicz property of the potential function. For the outer algorithm, the convergence is established under the condition that the associated penalty parameter goes to infinite and the inner method tends to be exactly convergent. Finally, numerical experiments are carried out. In numerical testing, the extrapolation step accelerates the rate of convergence to an approximate solution. When comparing the proposed algorithm with a semidefinite relaxation method for a structured QCQP problem, we observe the superiorities of our method in respect of the running time, especially when the size of the problem becomes large. However, the objective values of the two methods are close.

The rest of the chapter is organized as follows. In Section 3.2, we propose inner and outer algorithms and then introduce necessary definitions and preliminaries for the convergence analysis. After that, the attainment of a minimum of (3.1) is discussed by virtue of the asymptotic cone and asymptotic function in Section 3.3. In Sections 3.4 and 3.5, we present the convergence results of the inner and outer algorithms, respectively. Finally, the numerical testing is provided in Section 3.6.

3.2 Inner and Outer Algorithms

In this section, we compose inner and outer algorithms for a reformulation of (3.1). We also provide definitions of the stationary point and recall the Kurdyka-Łojasiewicz theory, which is essential to establish the global convergence of the inner algorithm.

3.2.1 Inner and Outer Algorithms

Invoking an auxiliary variable $z := (z_1, \cdots, z_p)$, we rewrite (3.1) as

$$\begin{aligned} \min_{x,y,z} \quad & q_0(x,y) + \delta_X(x) + \delta_Y(y) \\ \text{s.t.} \quad & q_i(x,y,z_i) = 0, \ i \in I \\ & z_i = x^{\mathrm{T}} M_i y, \ i \in I, \end{aligned} \tag{3.2}$$

where $q_0(x,y) := c_0^{\mathrm{T}} x + x^{\mathrm{T}} M_0 y + d_0^{\mathrm{T}} y$, $q_i(x,y,z_i) := z_i + c_i^{\mathrm{T}} x + d_i^{\mathrm{T}} y + e_i$ $(i \in I)$, $X := \{x : Ax \le u\}$, $Y := \{y : By \le v\}$, and $I := \{1, \cdots, p\}$.

To solve (3.2), we first apply an ADMM method with extrapolation as an inner algorithm to its quadratic penalty relaxation and then adopt an update of the penalty parameter as an outer algorithm to reach a stationary point of (3.2).

Inner Algorithm

Initially, we focus on the following quadratic penalty problem of (3.2) with the penalty parameter $\tau > 0$ fixed

$$\begin{aligned} \min_{x,y,z} \quad & \Psi_\tau(x,y,z) := q_0(x,y) + \frac{\tau}{2} \sum_{i \in I} q_i(x,y,z_i)^2 + \delta_X(x) + \delta_Y(y) \\ \text{s.t.} \quad & z_i = x^{\mathrm{T}} M_i y, \ i \in I. \end{aligned} \tag{3.3}$$

Let $\mu := (\mu_1, \cdots, \mu_p)$ be a Lagrangian multiplier. Then the Lagrangian function $\mathcal{L}_\tau$ and augmented Lagrangian function $\mathcal{L}_{\tau,\rho}$ of (3.3) are respectively given by

$$\mathcal{L}_\tau(x, y, z, \mu) := \ \Psi_\tau(x, y, z) + \sum_{i \in I} \mu_i (z_i - x^{\mathrm{T}} M_i y)$$

and

$$\mathcal{L}_{\tau,\rho}(x, y, z, \mu) := \mathcal{L}_\tau(x, y, z, \mu) + \frac{\rho}{2} \sum_{i \in I} (z_i - x^{\mathrm{T}} M_i y)^2,$$

where $\rho > 0$ is the (augmented) penalty parameter of $\mathcal{L}_\rho$. For simplicity, the smooth parts in $\mathcal{L}_\tau$ and $\mathcal{L}_{\tau,\rho}$ are respectively denoted by

$$L_\tau(x, y, z, \mu) := \ q_0(x, y) + \frac{\tau}{2} \sum_{i \in I} q_i(x, y, z_i)^2 + \sum_{i \in I} \mu_i (z_i - x^{\mathrm{T}} M_i y)$$

and

$$L_{\tau,\rho}(x, y, z, \mu) := L_\tau(x, y, z, \mu) + \frac{\rho}{2} \sum_{i \in I} (z_i - x^{\mathrm{T}} M_i y)^2.$$

Now, we propose an extrapolated proximal ADMM method for (3.3). The updating rules of the variables are given as follows:

$$\hat{x}^k = x^k + \alpha_x (x^k - x^{k-1}), \ \hat{y}^k = y^k + \alpha_y (y^k - y^{k-1}), \tag{3.4a}$$

$$\hat{z}^k = z^k + \alpha_z (z^k - z^{k-1}), \ \hat{\mu}^k = \mu^k + \alpha_\mu (\mu^k - \mu^{k-1}), \tag{3.4b}$$

$$x^{k+1} = \arg\min_{x \in X} \ \left\{ L_{\tau,\rho}(x, y^k, \hat{z}^k, \hat{\mu}^k) + \frac{\beta}{2} \|x - \hat{x}^k\|^2 \right\}, \tag{3.4c}$$

$$y^{k+1} = \arg\min_{y \in Y} \ \left\{ L_{\tau,\rho}(x^{k+1}, y, \hat{z}^k, \hat{\mu}^k) + \frac{\beta}{2} \|y - \hat{y}^k\|^2 \right\}, \tag{3.4d}$$

$$z^{k+1} = \arg\min_{z} \ \left\{ L_{\tau,\rho}(x^{k+1}, y^{k+1}, z, \hat{\mu}^k) \right\}, \tag{3.4e}$$

$$\mu_i^{k+1} = \hat{\mu}_i^k + \rho \left[z_i^{k+1} - (x^{k+1})^{\mathrm{T}} M_i y^{k+1} \right], \ i \in I, \tag{3.4f}$$

where $\alpha := (\alpha_x, \alpha_y, \alpha_z, \alpha_\mu) \geq 0$ are the extrapolation parameters, and $\beta \geq 0$ is the parameter of the proximal terms.

The use of extrapolation in (3.4c) is not standard; in particular, y^k is considered rather than $\hat{y}^k$. In (3.4c) and (3.4d), a proximal term is added to guarantee the strong convexity of the objective function, which is significant to obtain a descent property (see Lemma 3.4.3(1)). Moreover, by virtue of the optimality condition, (3.4e) produces an analytic solution (see (3.9c))

$$z_i^{k+1} = \frac{\rho(x^{k+1})^{\mathrm{T}} M_i y^{k+1} - \tau(c_i^{\mathrm{T}} x^{k+1} + d_i^{\mathrm{T}} y^{k+1} + e_i) - \hat{\mu}_i^k}{\tau + \rho}, \ i \in I.$$

On the basis of (3.4a)-(3.4f), we establish Algorithm 1 for (3.3).

Algorithm 1 Extrapolated proximal ADMM for solving (3.3) with τ fixed

```
Let $\tau$, $\rho$, $\beta$, $(\alpha_x, \alpha_y, \alpha_z, \alpha_\mu)$, and $(x^0, y^0, z^0, \mu^0)$ be given.
Initialization  $(x^{-1}, y^{-1}, z^{-1}, \mu^{-1}) = (x^0, y^0, z^0, \mu^0)$.
While  $k = 0, 1, 2, \cdots,$ do
- $\hat{x}^k = x^k + \alpha_x(x^k - x^{k-1})$, $\hat{y}^k = y^k + \alpha_y(y^k - y^{k-1})$, $\hat{z}^k = z^k + \alpha_z(z^k - z^{k-1})$, and
  $\hat{\mu}^k = \mu^k + \alpha_\mu(\mu^k - \mu^{k-1})$.
- $x^{k+1} = \arg\min_{x \in X}\{L_{\tau,\rho}(x, y^k, \hat{z}^k, \hat{\mu}^k) + \frac{\beta}{2}\|x - \hat{x}^k\|^2\}$.
- $y^{k+1} = \arg\min_{y \in Y}\{L_{\tau,\rho}(x^{k+1}, y, \hat{z}^k, \hat{\mu}^k) + \frac{\beta}{2}\|y - \hat{y}^k\|^2\}$.
- $z_i^{k+1} = \frac{\rho(x^{k+1})^{\mathrm{T}} M_i y^{k+1} - \tau(c_i^{\mathrm{T}} x^{k+1} + d_i^{\mathrm{T}} y^{k+1} + e_i) - \hat{\mu}_i^k}{\tau+\rho}$, $i \in I$.
- $\mu_i^{k+1} = \hat{\mu}^k + \rho[z_i^{k+1} - (x^{k+1})^{\mathrm{T}} M_i y^{k+1}]$, $i \in I$.
Output  $(x^{k+1}, y^{k+1}, z^{k+1}, \mu^{k+1})$.
End while
```

Outer Algorithm

Then, by specifying Algorithm 1 to (3.3) with τ updated, we construct outer iterations (Algorithm 2) for the original problem (3.2).

Algorithm 2 Penalty Algorithm 1 for solving (3.2)

```
Let the sequences $\{\tau^r\}$, $\{\rho^r\}$, and $\{\beta^r\}$, and quartet $(\alpha_x, \alpha_y, \alpha_z, \alpha_\mu)$ be given.
Initialization  $(\breve{x}^0, \breve{y}^0, \breve{z}^0, \breve{\mu}^0)$.
While  $r = 0, 1, 2, \cdots,$ do
- Let $\tau \leftarrow \tau^r$, $\rho \leftarrow \rho^r$, $\beta \leftarrow \beta^r$, and $(x^0, y^0, z^0, \mu^0) = (\breve{x}^r, \breve{y}^r, \breve{z}^r, \breve{\mu}^r)$.
- Adopt Algorithm 1 on (3.3), and write the resulting solution as $(\tilde{x}^r, \tilde{y}^r, \tilde{z}^r, \tilde{\mu}^r)$.
- Choose $(\breve{x}^{r+1}, \breve{y}^{r+1}, \breve{z}^{r+1}, \breve{\mu}^{r+1})$.
Output  $(\tilde{x}^{r+1}, \tilde{y}^{r+1}, \tilde{z}^{r+1}, \tilde{\mu}^{r+1})$.
End while
```

Notations

Let $\gamma_1 := \max\left\{\sum_{i\in I}\|c_i\|^2, \sum_{i\in I}\|d_i\|^2\right\}$ and $L^k_{\tau,\rho} := L_{\tau,\rho}(x^k, y^k, z^k, \mu^k)$. Our analysis greatly relies on a potential function $\Upsilon_{\tau,\rho,\alpha,\beta}$ and a potential sequence $\left\{\Upsilon^k_{\tau,\rho,\alpha,\beta}\right\}$, which are respectively written as

$$\begin{aligned}\Upsilon_{\tau,\rho,\alpha,\beta}(x, y, z, w, \mu) :=& \mathcal{L}_{\tau,\rho}(x, y, z, \mu) + \xi_x(\tau, \rho, \alpha, \beta)\,\|w_x\|^2 \\ &+ \xi_y(\tau, \rho, \alpha, \beta)\,\|w_y\|^2 + \xi_z(\tau, \rho, \alpha)\,\|w_z\|^2\end{aligned}$$

and

$$\begin{aligned}\Upsilon^k_{\tau,\rho,\alpha,\beta} :=& \mathcal{L}_{\tau,\rho}(x^k, y^k, z^k, \mu^k) + \xi_x(\tau, \rho, \alpha, \beta)\left\|x^k - x^{k-1}\right\|^2 \\ &+ \xi_y(\tau, \rho, \alpha, \beta)\left\|y^k - y^{k-1}\right\|^2 + \xi_z(\tau, \rho, \alpha)\left\|z^k - z^{k-1}\right\|^2,\end{aligned}$$

where $w_x \in \mathbb{R}^m$, $w_y \in \mathbb{R}^n$, $w_z \in \mathbb{R}^p$, $w := (w_x, w_y, w_z)$, and

$$\xi_x(\tau, \rho, \alpha, \beta) := \beta\alpha_x^2 + 3\tau^2\gamma_1\left(\frac{3}{2} + \frac{1}{2\rho}\right)\alpha_\mu^2$$

$$\xi_y(\tau, \rho, \alpha, \beta) := \beta\alpha_y^2 + 3\tau^2\gamma_1\left(\frac{3}{2} + \frac{1}{2\rho}\right)\alpha_\mu^2$$

$$\xi_z(\tau, \rho, \alpha) := (\tau + \rho)\alpha_z^2 + 3\tau^2\left(\frac{3}{2} + \frac{1}{2\rho}\right)\alpha_\mu^2.$$

3.2.2 Definitions and Preliminaries

We now introduce the stationary point of (3.2) and (3.3) and review the Kurdyka-Łojasiewicz theory for further analysis.

Stationary Point

Definition 3.2.1. *We say that* $(x^\star, y^\star, z^\star, \mu^\star, \nu^\star)$ *is a stationary point of* (3.2) *if the following holds*

$$-\Big(c_0 + M_0 y^\star - \sum_{i\in I}\mu_i^\star M_i y^\star + \sum_{i\in I}\nu_i^\star c_i\Big) \in N_X(x^\star) \tag{3.5a}$$

$$-\Big(d_0 + M_0^{\mathrm{T}} x^\star - \sum_{i\in I}\mu_i^\star M_i^{\mathrm{T}} x^\star + \sum_{i\in I}\nu_i^\star d_i\Big) \in N_Y(y^\star) \tag{3.5b}$$

$$\mu^\star + \nu^\star = 0 \tag{3.5c}$$

$$z_i^\star = (x^\star)^{\mathrm{T}} M_i y^\star,\ \ i \in I \tag{3.5d}$$

$$z_i^\star + c_i^{\mathrm{T}} x^\star + d_i^{\mathrm{T}} y^\star + e_i = 0,\ \ i \in I. \tag{3.5e}$$

Definition 3.2.2. *We say that* (x^*, y^*, z^*, μ^*) *is a stationary point of* (3.3) *if the following holds*

$$-\nabla_x L_\tau(x^*, y^*, z^*, \mu^*) \in N_X(x^*) \tag{3.6a}$$

$$-\nabla_y L_\tau(x^*, y^*, z^*, \mu^*) \in N_Y(y^*) \tag{3.6b}$$

$$\tau(z_i^* + c_i^{\mathrm{T}} x^* + d_i^{\mathrm{T}} y^* + e_i) + \mu_i^* = 0,\ \ i \in I \tag{3.6c}$$

$$z_i^* = (x^*)^{\mathrm{T}} M_i y^*,\ \ i \in I. \tag{3.6d}$$

In both definitions, the primal-dual limiting stationary point (i.e., the limiting stationary point of the Lagrangian problem) is considered rather than the primal one. For example, the system (3.6a)-(3.6d) interprets the condition $0 \in \partial^L \mathcal{L}_\tau(x^*, y^*, z^*, \mu^*)$ (see Proposition 1.5.4). For (3.2) and (3.3), the limiting subdifferential is identical to the Fréchet subdifferential due to the convexity of $\delta_X(x)$ and $\delta_Y(y)$.

Kurdyka-Łojasiewicz (KŁ) Theorey

The Kurdyka-Łojasiewicz (KŁ) theory (see Attouch et al. (2010) and Bolte et al. (2014)) plays an important role in analyzing the global convergence of Algorithm 1.

For $\eta \in (0, +\infty]$, let Φ_η be the class of continuous and concave functions $\phi : [0, \eta) \to [0, +\infty)$ satisfying (i) $\phi(0) = 0$; (ii) ϕ is continuously differentiable on $(0, \eta)$ and continuous at 0; (iii) $\phi'(x) > 0$ for all $x \in (0, \eta)$. The *KŁ property* and *KŁ function* are defined as follows.

Definition 3.2.3. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper and lower semi-continuous function and $\bar{x} \in \mathrm{dom}\partial^L f := \{x \in \mathbb{R}^n : \partial^L f(x) \neq \emptyset\}$. If there exist $\eta \in (0, +\infty]$, a neighborhood U of $\bar{x}$, and a function $\phi \in \Phi_\eta$ such that*

$$\phi'(f(x) - f(\bar{x}))\mathrm{dist}(0, \partial^L f(x)) \geq 1$$

for all $x \in U \cap \{x \in \mathbb{R}^n : f(\bar{x}) < f(x) < f(\bar{x}) + \eta\}$, then f is said to satisfy the Kurdyka-Łojasiewicz (KŁ) property at $\bar{x}$. Furthermore, we say that f is a KŁ function if f satisfies the KŁ property at every point in $\mathrm{dom}\partial^L f$.

For more general use of KŁ functions, Bolte et al. (2014, Lemma 6) provided the following *uniformized KŁ property*.

Proposition 3.2.4. *Let $\Xi \subseteq \mathbb{R}^n$ be a compact set and $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper and lower semi-continuous function. Suppose that f is constant on Ξ and satisfies the KŁ property at every point in Ξ. Then there exist $\varepsilon > 0$, $\eta > 0$, and a function $\phi \in \Phi_\eta$ such that*

$$\phi'(f(x) - f(\bar{x}))\mathrm{dist}(0, \partial^L f(x)) \geq 1$$

for all $\bar{x} \in \Xi$ and $x \in \{x \in \mathbb{R}^n : \mathrm{dist}(x, \Xi) < \varepsilon\} \cap \{x \in \mathbb{R}^n : f(\bar{x}) < f(x) < f(\bar{x}) + \eta\}$.

KŁ functions emerge in various applications and involve many classes of functions, among which the *semi-algebraic function* is a common instance (see Attouch et al. (2010, Section 4.3)).

Definition 3.2.5. *Let $S \subseteq \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. We say that S is a semi-algebraic set if it can be written as a finite union of sets with the structure*

$$\{x \in \mathbb{R}^n : g_i(x) = 0,\ h_i(x) < 0,\ i = 1, \cdots, r\},$$

where $g_i(x)$, $h_i(x)$ $(i = 1, \cdots, r) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are polynomial functions, and r is a positive finite integer. Furthermore, we say that f is a semi-algebraic function if its graph, i.e., $\{(x, t) \in \mathbb{R}^{n+1} : f(x) = t\}$, is a semi-algebraic subset of $\mathbb{R}^{n+1}$.

Proposition 3.2.6. *Let $S \subseteq \mathbb{R}^n$ be a semi-algebraic set and $g, h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be semi-algebraic functions. Then $g+h$, $g \cdot h$, and δ_S are all semi-algebraic functions.*

Proposition 3.2.7. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper and lower semi-continuous function. If f is a semi-algebraic function, then it is also a KŁ function.*

3.3 Existence of a Minimum of GBLP

Before investigating the convergence properties of Algorithms 1 and 2, we discuss the attainment of a minimum of (3.1) first. In doing so, we need to make use of the asymptotic cone and asymptotic function (for systematical learning, see Auslender & Teboulle (2006, Section 2)).

3.3.1 Asymptotic Cone and Asymptotic Function

Here, we introduce the concept of *asymptotic cone* and *asymptotic function.*

Definition 3.3.1. *For any nonempty set $S \subseteq \mathbb{R}^n$, let*

$$S_\infty := \left\{ d \in \mathbb{R}^n : \exists\ t_k \to +\infty \text{ and } \frac{x_k}{t_k} \to d \text{ with } x_k \in S \text{ as } k \to +\infty \right\}.$$

We say that S_∞ is the asymptotic cone of S.

Definition 3.3.2. *For any proper function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, there exists a unique function $f_\infty : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$ such that* $\mathrm{epi} f_\infty = (\mathrm{epi} f)_\infty$. *We say that f_∞ is the asymptotic function of f.*

Critical propositions, which contribute to analyzing the existence of a minimum, are stated as follows.

Proposition 3.3.3. *Let $S \subseteq \mathbb{R}^n$. Then S is bounded if and only if $S_\infty = \{0\}$.*

Proposition 3.3.4. *Let $S_i \subseteq \mathbb{R}^n$, $i \in \mathcal{I}$, where $\mathcal{I}$ is an arbitrary index set. Then*

$$(\cap_{i\in\mathcal{I}} S_i)_\infty \subseteq \cap_{i\in\mathcal{I}} (S_i)_\infty.$$

Proposition 3.3.5. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper function and $\alpha \in \mathbb{R}$ satisfy $lev_{\leq\alpha} f \neq \emptyset$. Then $(lev_{\leq\alpha} f)_\infty \subseteq lev_{\leq 0} f_\infty$, i.e.,*

$$\{x \in \mathbb{R}^n : f(x) \leq \alpha\}_\infty \subseteq \{d \in \mathbb{R}^n : f_\infty(d) \leq 0\}.$$

Noting that a bilinear function can be reformulated as a quadratic function, we present the explicit asymptotic function of a quadratic function for further analysis. An example for the asymptotic cone of a polyhedral convex set is also introduced, as such a set covers the inequality constraints in (3.1).

Example 3.3.6. *Let $f(x) := \frac{1}{2}x^{\mathrm{T}}Qx + r^{\mathrm{T}}x + s$, where $Q \in \mathbb{R}^{n\times n}$, $r \in \mathbb{R}^n$, and $s \in \mathbb{R}$.*

(1). If Q is positive semidefinite, then

$$f_\infty(d) = \begin{cases} r^{\mathrm{T}}d, & \text{if } Qd = 0, \\ +\infty, & \text{if } Qd \neq 0. \end{cases}$$

(2). If Q is not positive semidefinite, then

$$f_\infty(d) = \begin{cases} +\infty, & \text{if } d^{\mathrm{T}}Qd > 0, \\ -\infty, & \text{if } d^{\mathrm{T}}Qd \leq 0. \end{cases}$$

Example 3.3.7. *Let* $S := \{x \in \mathbb{R}^n : Qx \le r\}$*, where* $Q \in \mathbb{R}^{m\times n}$ *and* $r \in \mathbb{R}^m$*. Then* $S_\infty = \{d \in \mathbb{R}^n : Qd \le 0\}$*.*

3.3.2 Existence of a Minimum

Based on the above propositions and examples, we provide a sufficient condition of the existence of a minimum of (3.1).

Theorem 3.3.8. *Let the feasible set of* (3.1) *be nonempty and*

$$\mathcal{S} := \left\{(x, y) \in \mathbb{R}^m \times \mathbb{R}^n : c_i^{\mathrm{T}}x + d_i^{\mathrm{T}}y = 0, i \in \mathcal{E};\ x^{\mathrm{T}}M_iy = 0, i \in \mathcal{N};\ Ax \le 0;\ By \le 0\right\},$$

where $\mathcal{E} := \{i \in I : M_i = 0\}$*, and* $\mathcal{N} := \{i \in I : M_i \ne 0\}$*. Then the optimal solution set of* (3.1) *is nonempty and compact if one of the following holds.*

(1). $M_0 = 0$ *and* $\left\{(x, y) \in \mathcal{S} : c_0^{\mathrm{T}}x + d_0^{\mathrm{T}}y \le 0\right\} = \{(0, 0)\}$*.*

(2). $M_0 \ne 0$ *and* $\left\{(x, y) \in \mathcal{S} : x^{\mathrm{T}}M_0y \le 0\right\} = \{(0, 0)\}$*.*

Proof. For convenience, we write

$$\begin{aligned}
&\bar{q}_0(x, y) := x^{\mathrm{T}}M_0y + c_0^{\mathrm{T}}x + d_0^{\mathrm{T}}y;\\
&\bar{q}_i(x, y) := x^{\mathrm{T}}M_iy + c_i^{\mathrm{T}}x + d_i^{\mathrm{T}}y + e_i,\ i \in I;\\
&\mathcal{Q}_i := \{(x, y) : \bar{q}_i(x, y) = 0\},\ i \in I;\\
&\mathcal{Q} := \cap_{i\in I}\mathcal{Q}_i,\ \mathcal{R} := \{(x, y) : Ax \le u,\ By \le v\};\\
&h(x, y) := \bar{q}_0(x, y) + \delta_{\mathcal{Q}} + \delta_{\mathcal{R}}(x, y).
\end{aligned}$$

In fact, the proof can be completed by a well-known result: if a proper and lower semicontinues function $f(x)$ is level bounded, then the optimal solution set of $\min_x f(x)$ is nonempty and compact (see Rockafellar & Wets (2009, Theorem 1.9)). As the boundedness of a set can be expressed by its asymptotic cone (see Proposition 3.3.3), next, we evaluate $(\mathrm{lev}_{\le\alpha}h)_\infty$ for each $\alpha \in \mathbb{R}$ satisfying $\mathrm{lev}_{\le\alpha}h \ne \emptyset$.

Noting $\bar{q}_i(x,y) = \frac{1}{2}(x^{\mathrm{T}}\ y^{\mathrm{T}}) \begin{pmatrix} 0 & M_i \\ M_i^{\mathrm{T}} & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + (c_i^{\mathrm{T}}\ d_i^{\mathrm{T}}) \begin{pmatrix} x \\ y \end{pmatrix} + e_i$, we see from Example 3.3.6 that, for any $j \in \{0\} \cup I$,

$$(\bar{q}_j)_\infty(x,y) = \begin{cases} c_j^{\mathrm{T}} x + d_j^{\mathrm{T}} y, & \text{if } M_j = 0, \\ +\infty, & \text{if } M_j \neq 0,\ x^{\mathrm{T}} M_j y > 0, \\ -\infty, & \text{if } M_j \neq 0,\ x^{\mathrm{T}} M_j y \leq 0. \end{cases}$$

Then, it follows from Propositions 3.3.4 and 3.3.5 that, for any $i \in I$,

$$\begin{aligned} (\mathcal{Q}_i)_\infty &\subseteq \{(x,y) : (q_i)_\infty(x,y) \leq 0,\ (-q_i)_\infty(x,y) \leq 0\} \\ &= \begin{cases} \left\{(x,y) : c_i^{\mathrm{T}} x + d_i^{\mathrm{T}} y = 0\right\}, & \text{if } M_i = 0, \\ \left\{(x,y) : x^{\mathrm{T}} M_i y = 0\right\}, & \text{if } M_i \neq 0, \end{cases} \end{aligned} \tag{3.7}$$

and from Example 3.3.7 that

$$\mathcal{R}_\infty = \{(x,y) : Ax \leq 0,\ By \leq 0\}. \tag{3.8}$$

Now, we obtain the estimate of $(\mathrm{lev}_{\leq \alpha} h)_\infty$, that is,

$$\begin{aligned} (\mathrm{lev}_{\leq\alpha} h)_\infty &= (\mathrm{lev}_{\leq\alpha} \bar{q}_0 \cap \mathcal{Q} \cap \mathcal{R})_\infty \\ &\subseteq (\mathrm{lev}_{\leq\alpha} \bar{q}_0)_\infty \cap_{i \in I} (\mathcal{Q}_i)_\infty \cap \mathcal{R}_\infty \\ &\subseteq \left\{(x,y) \in \mathcal{S} : (\bar{q}_0)_\infty(x,y) \leq 0\right\}, \end{aligned}$$

where the first inclusion comes from Proposition 3.3.4, and the second holds due to Proposition 3.3.5, (3.7), and (3.8). Under the conditions (1) and (2), one has that $(\mathrm{lev}_{\leq\alpha} h)_\infty \subseteq \{(0,0)\}$. On the other hand, we have that $\{(0,0)\} \subseteq (\mathrm{lev}_{\leq\alpha} h)_\infty$ from the definition of the asymptotic cone. Then, by virtue of Proposition 3.3.3, h is level bounded, which completes the proof. □

3.4 Convergence Analysis of Inner Algorithm

This section explores the convergence properties of Algorithm 1 and starts with listing the first-order optimality conditions of (3.4c)-(3.4e) as follows:

$$\left\langle \nabla_x L_{\tau,\rho}(x^{k+1}, y^k, \hat{z}^k, \hat{\mu}^k) + \beta(x^{k+1} - \hat{x}^k), x - x^{k+1} \right\rangle \geq 0 \quad \text{for all} \quad x \in X, \tag{3.9a}$$

$$\left\langle \nabla_y L_{\tau,\rho}(x^{k+1}, y^{k+1}, \hat{z}^k, \hat{\mu}^k) + \beta(y^{k+1} - \hat{y}^k), y - y^{k+1} \right\rangle \geq 0 \quad \text{for all} \quad y \in Y, \tag{3.9b}$$

$$\tau q_i(x^{k+1}, y^{k+1}, z_i^{k+1}) + \hat{\mu}_i^k + \rho[z_i^{k+1} - (x^{k+1})^{\mathrm{T}} M_i y^{k+1}] = 0, \ i \in I. \tag{3.9c}$$

The estimates (3.4f) and (3.9c) yield the following equality, which will be frequently used in the subsequent proofs

$$\mu_i^{k+1} = -\tau q_i(x^{k+1}, y^{k+1}, z_i^{k+1}), \ i \in I. \tag{3.10}$$

Assumption 3.4.1 is made throughout this section for establishing an important descent property of the potential function.

Assumption 3.4.1.

(1). Assume that the sequences $\left\{x^k\right\}$ *and* $\left\{y^k\right\}$*, generated by Algorithm 1, are bounded. Then, for any* $k \in \mathbb{N}$*, there exists* $\underline{q}_0 \in \mathbb{R}$ *such that* $q_0(x^k, y^k) \geq \underline{q}_0$ *and exists* $\gamma_2 > 0$ *such that* $\sum\limits_{i\in I} \left\|(x^k)^{\mathrm{T}} M_i\right\|^2 \leq \gamma_2$ *and* $\sum\limits_{i\in I} \left\|M_i y^k\right\|^2 \leq \gamma_2$.

(2). Assume that the parameters ρ*,* β*, and* $(\alpha_x, \alpha_y, \alpha_z, \alpha_\mu)$ *in Algorithm 1 satisfy one of the following.*

(2a). $(\alpha_x, \alpha_y, \alpha_z) \neq 0$ *and*

$$\begin{cases} \left(\dfrac{3}{4} - \max\left\{\alpha_x^2, \alpha_y^2\right\}\right)\beta \\ \qquad - \dfrac{(1+\rho)\gamma_2}{2} - \dfrac{\tau^2\gamma_1}{2}\left[\dfrac{1}{\tau} + \dfrac{9}{\rho} + 3\alpha_\mu^2\left(3 + \dfrac{1}{\rho}\right)\right] > 0 \\ \dfrac{\tau+\rho-1}{2} - \dfrac{9\tau^2}{2\rho} - (\tau+\rho)\alpha_z^2 - 3\tau^2\alpha_\mu^2\left(\dfrac{3}{2} + \dfrac{1}{2\rho}\right) > 0. \end{cases}$$

(2b). $(\alpha_x, \alpha_y, \alpha_z) = 0$, $\alpha_\mu \neq 0$, *and*

$$\begin{cases} \beta - \dfrac{\gamma_2}{2} - \dfrac{3\tau^2\gamma_1}{2}\left[\dfrac{3}{\rho} + \alpha_\mu^2\left(3 + \dfrac{1}{\rho}\right)\right] > 0 \\ \dfrac{\tau + \rho - 1}{2} - \dfrac{9\tau^2}{2\rho} - 3\tau^2\alpha_\mu^2\left(\dfrac{3}{2} + \dfrac{1}{2\rho}\right) > 0. \end{cases}$$

(2c). $(\alpha_x, \alpha_y, \alpha_z, \alpha_\mu) = 0$ *and* $\rho > \max\left\{\frac{3\tau^2\gamma_1}{\beta},\ 2\tau\right\}$.

Parameters satisfying Assumption 3.4.1(2a) or (2b) always exist. Specifically, when ρ and β are sufficiently large, those inequalities are met.

3.4.1 Basic Properties

Lemmas presented in this subsection indicate some significant properties of the iterates generated by Algorithm 1.

The first lemma states that the successive difference of dual iterates can be bounded by that of primal ones.

Lemma 3.4.2. *Let $\{x^k\}$, $\{y^k\}$, $\{z^k\}$, and $\{\mu^k\}$ be the sequences generated by Algorithm 1. Then, for any $k \in \mathbb{N}$, one has that*

$$\|\mu^{k+1} - \mu^k\|^2 \leq 3\tau^2\left(\gamma_1\|x^{k+1} - x^k\|^2 + \gamma_1\|y^{k+1} - y^k\|^2 + \|z^{k+1} - z^k\|^2\right).$$

Proof. It follows from (3.10) that, for any $k \in \mathbb{N}$ and $i \in I$,

$$|\mu_i^{k+1} - \mu_i^k| = \tau|q_i(x^{k+1}, y^{k+1}, z_i^{k+1}) - q_i(x^k, y^k, z_i^k)|.$$

Then, for any $k \in \mathbb{N}$ and $i \in I$, it holds that

$$\begin{aligned} |\mu_i^{k+1} - \mu_i^k|^2 &\leq \tau^2(\|c_i\|\|x^{k+1} - x^k\| + \|d_i\|\|y^{k+1} - y^k\| + |z_i^{k+1} - z_i^k|)^2 \\ &\leq 3\tau^2(\|c_i\|^2\|x^{k+1} - x^k\|^2 + \|d_i\|^2\|y^{k+1} - y^k\|^2 + |z_i^{k+1} - z_i^k|^2), \end{aligned}$$

where the estimates come from the Cauchy-Schwarz inequality and Young's inequality, respectively. Summing the left-hand side and right-hand side of the above inequality over $i \in I$, we obtain the desired estimate. Thus, the proof is completed. $\square$

With Lemma 3.4.2, we can obtain the descent of $\left\{\Upsilon^k_{\tau,\rho,\alpha,\beta}\right\}$. This kind of property is a critical condition to establish the subsequential convergence and global convergence for nonconvex ADMM-type algorithms. In what follows, we provide the convergence of $\left\{\Upsilon^k_{\tau,\rho,\alpha,\beta}\right\}$ by virtue of its descent and lower boundedness properties.

Lemma 3.4.3. *Let $\left\{x^k\right\}$, $\left\{y^k\right\}$, $\left\{z^k\right\}$, and $\left\{\mu^k\right\}$ be the sequences generated by Algorithm 1 and Assumption 3.4.1 hold. Then*

(1). $\left\{\Upsilon^k_{\tau,\rho,\alpha,\beta}\right\}$ is decreasing, and in particular, there exists $\xi > 0$, for any $k \in \mathbb{N}$, one has that

$$\Upsilon^{k+1}_{\tau,\rho,\alpha,\beta} - \Upsilon^k_{\tau,\rho,\alpha,\beta} \leq -\,\xi\left(\|x^{k+1}-x^k\|^2 + \|y^{k+1}-y^k\|^2 + \|z^{k+1}-z^k\|^2\right);$$

(2). $\left\{\Upsilon^k_{\tau,\rho,\alpha,\beta}\right\}$ is lower bounded, and in particular, for any $k \in \mathbb{N}$, one has that

$$\Upsilon^k_{\tau,\rho,\alpha,\beta} \geq L^k_{\tau,\rho} \geq \underline{q}_0.$$

Therefore, the potential sequence $\left\{\Upsilon^k_{\tau,\rho,\alpha,\beta}\right\}$ is convergent.

Proof. (1). Our proof begins with bounding the successive difference of $\{L^k_{\tau,\rho}\}$. To this end, we split the successive difference into four parts, that is, for any $k \in \mathbb{N}$,

$$L^{k+1}_{\tau,\rho} - L^k_{\tau,\rho} = E^{k+1}_1 + E^{k+1}_2 + E^{k+1}_3 + E^{k+1}_4,$$

where

$$\begin{aligned}
E^{k+1}_1 &:= L^{k+1}_{\tau,\rho} - L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^{k+1}, \mu^k),\\
E^{k+1}_2 &:= L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^{k+1}, \mu^k) - L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^k, \mu^k),\\
E^{k+1}_3 &:= L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^k, \mu^k) - L_{\tau,\rho}(x^{k+1}, y^k, z^k, \mu^k),\\
E^{k+1}_4 &:= L_{\tau,\rho}(x^{k+1}, y^k, z^k, \mu^k) - L^k_{\tau,\rho}.
\end{aligned}$$

We consider three cases: $(\alpha_x, \alpha_y, \alpha_z) \neq 0$, $(\alpha_x, \alpha_y, \alpha_z) = 0$ and $\alpha_\mu \neq 0$, and $(\alpha_x, \alpha_y, \alpha_z, \alpha_\mu) = 0$.

Case (i). Let $(\alpha_x, \alpha_y, \alpha_z) \neq 0$. Now, we estimate the upper bounds of E_1^{k+1}–E_4^{k+1} one by one. For E_1^{k+1}, we have that

$$\begin{aligned} E_1^{k+1} &= \sum_{i \in I} (\mu_i^{k+1} - \mu_i^k)[z_i^{k+1} - (x^{k+1})^{\mathrm{T}} M_i y^{k+1}] \\ &= \frac{1}{\rho}\left\langle \mu^{k+1} - \mu^k, \mu^{k+1} - \mu^k \right\rangle + \frac{1}{\rho}\left\langle \mu^{k+1} - \mu^k, \mu^k - \hat{\mu}^k \right\rangle \\ &\leq \frac{3}{2\rho}\|\mu^{k+1} - \mu^k\|^2 + \frac{1}{2\rho}\|\mu^k - \hat{\mu}^k\|^2, \end{aligned}$$

where the second and third estimates hold due to (3.4f) and the Young's inequality, respectively.

For E_2^{k+1}, we have that

$$\begin{aligned} E_2^{k+1} = &\left(L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^{k+1}, \hat{\mu}^k) - L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^k, \hat{\mu}^k)\right) \\ &+ \left(L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^{k+1}, \mu^k) - L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^{k+1}, \hat{\mu}^k)\right) \\ &- \left(L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^k, \mu^k) - L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^k, \hat{\mu}^k)\right) \\ \leq &- \left\langle \nabla_z L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^{k+1}, \hat{\mu}^k),\ z^k - z^{k+1} \right\rangle - \frac{\tau + \rho}{2}\left\| z^{k+1} - z^k \right\|^2 \\ &+ \sum_{i \in I} (\mu_i^k - \hat{\mu}_i^k)[z_i^{k+1} - (x^{k+1})^{\mathrm{T}} M_i y^{k+1}] - \sum_{i \in I} (\mu_i^k - \hat{\mu}_i^k)[z_i^k - (x^{k+1})^{\mathrm{T}} M_i y^{k+1}] \\ \leq &- \frac{\tau + \rho}{2}\left\| z^{k+1} - z^k \right\|^2 + \left\langle \mu^k - \hat{\mu}^k, z^{k+1} - z^k \right\rangle \\ \leq &- \frac{\tau + \rho}{2}\left\| z^{k+1} - z^k \right\|^2 + \frac{1}{2}\left\| \mu^k - \hat{\mu}^k \right\|^2 + \frac{1}{2}\left\| z^{k+1} - z^k \right\|^2, \end{aligned}$$

where the second to fourth estimates hold due to the strong convexity of $L_{\tau,\rho}$ with respect to z, optimality condition of (3.4e), and Young's inequality, respectively.

When evaluating E_3^{k+1}, we make use of the following relation

$$\|y^{k+1} - \hat{y}^k\|^2 + \|y^k - \hat{y}^k\|^2 \geq \frac{1}{2}\|y^{k+1} - y^k\|^2, \tag{3.11}$$

which comes from the triangle inequality and Lemma 4.1(ii) in Huang & Yang (2003). For E_3^{k+1}, we have that

$$\begin{aligned}
E_3^{k+1} =& \Big[\big(L_{\tau,\rho}(x^{k+1}, y^{k+1}, \hat{z}^k, \hat{\mu}^k) + \frac{\beta}{2}\|y^{k+1} - \hat{y}^k\|^2\big) \\
&- \big(L_{\tau,\rho}(x^{k+1}, y^k, \hat{z}^k, \hat{\mu}^k) + \frac{\beta}{2}\|y^k - \hat{y}^k\|^2\big)\Big] - \frac{\beta}{2}\|y^{k+1} - \hat{y}^k\|^2 + \frac{\beta}{2}\|y^k - \hat{y}^k\|^2 \\
&+ \big(L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^k, \mu^k) - L_{\tau,\rho}(x^{k+1}, y^{k+1}, \hat{z}^k, \hat{\mu}^k)\big) \\
&- \big(L_{\tau,\rho}(x^{k+1}, y^k, z^k, \mu^k) - L_{\tau,\rho}(x^{k+1}, y^k, \hat{z}^k, \hat{\mu}^k)\big) \\
\leq& - \frac{\beta}{2}\|y^{k+1} - y^k\|^2 - \frac{\beta}{2}\|y^{k+1} - \hat{y}^k\|^2 + \frac{\beta}{2}\|y^k - \hat{y}^k\|^2 \\
&+ \tau\sum_{i\in I} d_i^{\mathrm{T}}(y^{k+1} - y^k)(z_i^k - \hat{z}_i^k) + \sum_{i\in I}(x^{k+1})^{\mathrm{T}} M_i(y^k - y^{k+1})(\mu_i^k - \hat{\mu}_i^k) \\
&+ \rho\sum_{i\in I}(x^{k+1})^{\mathrm{T}} M_i(y^k - y^{k+1})(z_i^k - \hat{z}_i^k) \\
\leq& - \frac{3\beta}{4}\|y^{k+1} - y^k\|^2 + \beta\|y^k - \hat{y}^k\|^2 + \frac{\tau\gamma_1}{2}\left\|y^{k+1} - y^k\right\|^2 + \frac{\tau}{2}\left\|z^k - \hat{z}^k\right\|^2 \\
&+ \frac{\gamma_2}{2}\left\|y^{k+1} - y^k\right\|^2 + \frac{1}{2}\left\|\mu^k - \hat{\mu}^k\right\|^2 + \frac{\rho\gamma_2}{2}\left\|y^{k+1} - y^k\right\|^2 + \frac{\rho}{2}\left\|z^k - \hat{z}^k\right\|^2 \\
=& - \left(\frac{3\beta}{4} - \frac{\tau\gamma_1 + \gamma_2 + \rho\gamma_2}{2}\right)\|y^{k+1} - y^k\|^2 + \beta\|y^k - \hat{y}^k\|^2 \\
&+ \frac{\tau + \rho}{2}\left\|z^k - \hat{z}^k\right\|^2 + \frac{1}{2}\left\|\mu^k - \hat{\mu}^k\right\|^2,
\end{aligned}$$

where the second inequality is derived by the strong convexity of the objective function in (3.4d) with respect to y and optimality condition (3.9b), and the third follows from (3.11), Assumption 3.4.1(1), and the Young's inequality, respectively.

Likewise, for E_4^{k+1}, we have that

$$\begin{aligned} E_4^{k+1} \leq &-\left(\frac{3\beta}{4} - \frac{\tau\gamma_1 + \gamma_2 + \rho\gamma_2}{2}\right)\|x^{k+1} - x^k\|^2 + \beta\|x^k - \hat{x}^k\|^2 \\ &+ \frac{\tau+\rho}{2}\left\|z^k - \hat{z}^k\right\|^2 + \frac{1}{2}\left\|\mu^k - \hat{\mu}^k\right\|^2. \end{aligned}$$

Summing up E_i^{k+1}, $i = 1, 2, 3, 4$, and invoking (3.4a), (3.4b), and Lemma 3.4.2, we obtain that, for any $k \in \mathbb{N}$,

$$\begin{aligned} L_{\tau,\rho}^{k+1} - L_{\tau,\rho}^{k} \leq &- \bar{\xi}_x(\tau,\rho,\beta)\|x^{k+1} - x^k\|^2 + \xi_x(\tau,\rho,\alpha,\beta)\|x^k - x^{k-1}\|^2 \\ &- \bar{\xi}_y(\tau,\rho,\beta)\|y^{k+1} - y^k\|^2 + \xi_y(\tau,\rho,\alpha,\beta)\|y^k - y^{k-1}\|^2 \\ &- \bar{\xi}_z(\tau,\rho)\|z^{k+1} - z^k\|^2 + \xi_z(\tau,\rho,\alpha)\|z^k - z^{k-1}\|^2, \end{aligned}$$

where $\bar{\xi}_x(\tau,\rho,\beta) = \bar{\xi}_y(\tau,\rho,\beta) = \frac{3\beta}{4} - \frac{\tau\gamma_1+\gamma_2+\rho\gamma_2}{2} - \frac{9\tau^2\gamma_1}{2\rho}$ and $\bar{\xi}_z(\tau,\rho) = \frac{\tau+\rho-1}{2} - \frac{9\tau^2}{2\rho}$ (recall that $\xi_x(\tau,\rho,\alpha,\beta)$, $\xi_y(\tau,\rho,\alpha,\beta)$, and $\xi_z(\tau,\rho,\alpha)$ are the coefficients of $\Upsilon_{\tau,\rho,\alpha,\beta}$). Then, we obtain the bound of the successive difference of $\left\{\Upsilon_{\tau,\rho,\alpha,\beta}^k\right\}$; that is, for any $k \in \mathbb{N}$,

$$\begin{aligned} \Upsilon_{\tau,\rho,\alpha,\beta}^{k+1} - \Upsilon_{\tau,\rho,\alpha,\beta}^{k} \leq &- (\bar{\xi}_x(\tau,\rho,\beta) - \xi_x(\tau,\rho,\alpha,\beta))\|x^{k+1} - x^k\|^2 \\ &- (\bar{\xi}_y(\tau,\rho,\beta) - \xi_y(\tau,\rho,\alpha,\beta))\|y^{k+1} - y^k\|^2 \\ &- (\bar{\xi}_z(\tau,\rho) - \xi_z(\tau,\rho,\alpha))\|z^{k+1} - z^k\|^2. \end{aligned}$$

Clearly, $\{\Upsilon_{\tau,\rho,\alpha,\beta}^k\}$ is decreasing with Assumption 3.4.1(2a), which makes the coefficients $\bar{\xi}_x(\rho,\beta) - \xi_x(\rho,\alpha,\beta)$, $\bar{\xi}_y(\tau,\rho,\beta) - \xi_y(\tau,\rho,\alpha,\beta)$, and $\bar{\xi}_z(\tau,\rho) - \xi_z(\tau,\rho,\alpha)$ all positive. Therefore, the statement (1) is true for this case.

Case (ii). Let $(\alpha_x, \alpha_y, \alpha_z) = 0$ and $\alpha_\mu \neq 0$. Similar to Case (i), the upper bounds of E_1^{k+1}–E_4^{k+1} are respectively estimated as

$$\begin{aligned}
E_1^{k+1} &\leq \frac{3}{2\rho}\|\mu^{k+1}-\mu^k\|^2+\frac{1}{2\rho}\|\mu^k-\hat{\mu}^k\|^2\\
E_2^{k+1} &\leq -\frac{\tau+\rho-1}{2}\left\|z^{k+1}-z^k\right\|^2+\frac{1}{2}\left\|\mu^k-\hat{\mu}^k\right\|^2\\
E_3^{k+1} &\leq -\left(\beta-\frac{\gamma_2}{2}\right)\|y^{k+1}-y^k\|^2+\frac{1}{2}\left\|\mu^k-\hat{\mu}^k\right\|^2\\
E_4^{k+1} &\leq -\left(\beta-\frac{\gamma_2}{2}\right)\|x^{k+1}-x^k\|^2+\frac{1}{2}\left\|\mu^k-\hat{\mu}^k\right\|^2.
\end{aligned}$$

With Assumption 3.4.1(2b), the rest of the proof for this case can be completed similarly as for Case (i).

Case (iii). Let $(\alpha_x,\alpha_y,\alpha_z,\alpha_\mu)=0$. Similar to Case (i), the upper bounds of E_1^{k+1}–E_4^{k+1} are respectively estimated as

$$\begin{aligned}
E_1^{k+1} &\leq \frac{1}{\rho}\|\mu^{k+1}-\mu^k\|^2\\
E_2^{k+1} &\leq -\frac{\tau+\rho}{2}\left\|z^{k+1}-z^k\right\|^2\\
E_3^{k+1} &\leq -\beta\|y^{k+1}-y^k\|^2\\
E_4^{k+1} &\leq -\beta\|x^{k+1}-x^k\|^2.
\end{aligned}$$

With Assumption 3.4.1(2c), the rest of the proof for this case can be completed similarly as for Case (i).

(2). As $\Upsilon_{\tau,\rho,\alpha,\beta}^k \geq L_{\tau,\rho}^k$ $(k\in\mathbb{N})$ is evident from the definition, we only need to focus on the lower bound of $\left\{L_{\tau,\rho}^k\right\}$.

It follows from (3.10) that, for any $k \in \mathbb{N}$,

$$\begin{aligned} L_{\tau,\rho}^k &= q_0(x^k, y^k) + \frac{\tau}{2}\sum_{i\in I} q_i(x^k, y^k, z_i^k)^2 \\ &\quad - \tau \sum_{i\in I} q_i(x^k, y^k, z_i^k)\left[z_i^k - (x^k)^{\mathrm{T}} M_i y^k\right] + \frac{\rho}{2}\sum_{i\in I}\left[z_i^k - (x^k)^{\mathrm{T}} M_i y^k\right]^2 \\ &= q_0(x^k, y^k) + \frac{\rho-\tau}{2}\sum_{i\in I}\left[z_i^k - (x^k)^{\mathrm{T}} M_i y^k\right]^2 \\ &\quad + \frac{\tau}{2}\sum_{i\in I}\left[q_i(x^k, y^k, z_i^k) - \left(z_i^k - (x^k)^{\mathrm{T}} M_i y^k\right)\right]^2. \end{aligned} \tag{3.12}$$

Noting that $\rho > \tau$ is guaranteed by Assumption 3.4.1(2), we obtain $L_{\tau,\rho}^k \geq \underline{q}_0$ for all $k \in \mathbb{N}$ from Assumption 3.4.1(1). Thus, the proof is completed. □

We end this subsection by the boundedness of iterates (see Lemma 3.4.4) and the bound of iterative subgradients (see Lemma 3.4.5).

Lemma 3.4.4. *Let $\{x^k\}$, $\{y^k\}$, $\{z^k\}$, and $\{\mu^k\}$ be the sequences generated by Algorithm 1 and Assumption 3.4.1 hold. Then $\{z^k\}$ and $\{\mu^k\}$ are bounded.*

Proof. Initially, from Lemma 3.4.3, $L_{\tau,\rho}^k < +\infty$ holds for all $k \in \mathbb{N}$. Then, as $\rho > \tau$ (from Assumption 3.4.1(2)), we see from (3.12) that $|q_i(x^k, y^k, z_i^k)| < +\infty$ holds for all $k \in \mathbb{N}$ and $i \in I$. Finally, by virtue of (3.10) and the boundedness of $\{x^k\}$ and $\{y^k\}$, one has that $\{z^k\}$ and $\{\mu^k\}$ are bounded. □

Lemma 3.4.5. *Let $\{x^k\}$, $\{y^k\}$, $\{z^k\}$, and $\{\mu^k\}$ be the sequences generated by Algorithm 1 and Assumption 3.4.1 hold. Then, for any $k \in \mathbb{N}$, there exists $\zeta > 0$ such that*

$$\begin{aligned} &\operatorname{dist}\left(0, \partial^L \mathcal{L}_{\tau,\rho}(x^{k+1}, y^{k+1}, z^{k+1}, \mu^{k+1})\right) \\ \leq\ &\zeta\big(\|x^{k+1} - x^k\| + \|y^{k+1} - y^k\| + \|z^{k+1} - z^k\| + \|\mu^{k+1} - \mu^k\| \\ &+ \|x^k - x^{k-1}\| + \|y^k - y^{k-1}\| + \|z^k - z^{k-1}\| + \|\mu^k - \mu^{k-1}\|\big). \end{aligned}$$

Proof. From Proposition 1.5.4 and (3.4f), the left-hand side of the estimate can be bounded above by four parts, that is, for any $k \in \mathbb{N}$,

$$\begin{aligned}
&\operatorname{dist}\left(0, \partial^L \mathcal{L}_{\tau,\rho}(x^{k+1}, y^{k+1}, z^{k+1}, \mu^{k+1})\right) \\
\leq\ &\operatorname{dist}\left(-\nabla_x L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^{k+1}, \mu^{k+1}), N_X(x^{k+1})\right) \\
&+ \operatorname{dist}\left(-\nabla_y L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^{k+1}, \mu^{k+1}), N_Y(y^{k+1})\right) \\
&+ \left\|\nabla_z L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^{k+1}, \mu^{k+1})\right\| + \frac{1}{\rho}\left\|\mu^{k+1} - \hat{\mu}^k\right\|.
\end{aligned}$$

We first consider the bound of $\operatorname{dist}\left(-\nabla_x L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^{k+1}, \mu^{k+1}), N_X(x^{k+1})\right)$. For this propose, we rewrite the optimality condition (3.9a) as

$$-\nabla_x L_{\tau,\rho}(x^{k+1}, y^k, \hat{z}^k, \hat{\mu}^k) - \beta(x^{k+1} - \hat{x}^k) \in N_X(x^{k+1}).$$

On the other hand, $\nabla_x L_{\tau,\rho}$ is Lipschitz continuous on any bounded set. Then, for any $k \in \mathbb{N}$, there exists $\zeta_1 > 0$ such that

$$\begin{aligned}
&\left|\nabla_x L_{\tau,\rho}(x^{k+1}, y^k, \hat{z}^k, \hat{\mu}^k) + \beta(x^{k+1} - \hat{x}^k) - \nabla_x L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^{k+1}, \mu^{k+1})\right| \\
\leq\ &\zeta_1\left(\|x^{k+1} - \hat{x}^k\| + \|y^{k+1} - y^k\| + \|z^{k+1} - \hat{z}^k\| + \|\mu^{k+1} - \hat{\mu}^k\|\right).
\end{aligned}$$

From the above two estimates, there exists $\zeta_2 > 0$, for any $k \in \mathbb{N}$, one has that

$$\begin{aligned}
&\operatorname{dist}\left(-\nabla_x L_{\tau,\rho}(x^{k+1}, y^{k+1}, z^{k+1}, \mu^{k+1}), N_X(x^{k+1})\right) \\
\leq\ &\zeta_1\left(\|x^{k+1} - \hat{x}^k\| + \|y^{k+1} - y^k\| + \|z^{k+1} - \hat{z}^k\| + \|\mu^{k+1} - \hat{\mu}^k\|\right) \\
\leq\ &\zeta_1\left(\|x^{k+1} - x^k\| + \|y^{k+1} - y^k\| + \|z^{k+1} - z^k\| + \|\mu^{k+1} - \mu^k\|\right) \\
&+ \zeta_2\left(\|x^k - x^{k-1}\| + \|y^k - y^{k-1}\| + \|z^k - z^{k-1}\| + \|\mu^k - \mu^{k-1}\|\right).
\end{aligned}$$

Repeating the above process for the rest terms, we obtain similar inequalities. Thus, the proof is completed. □

3.4.2 Subsequential Convergence

In this subsection, we establish the subsequential convergence to a stationary point and the convergence of some critical sequences.

The following theorem presents the convergence for the successive difference of iterates and subsequential convergence of the method.

Theorem 3.4.6. *Let $\{x^k\}$, $\{y^k\}$, $\{z^k\}$, and $\{\mu^k\}$ be the sequences generated by Algorithm 1 and Assumption 3.4.1 hold. Then*

(1). $\lim\limits_{k\to+\infty} \|x^{k+1}-x^k\| + \|y^{k+1}-y^k\| + \|z^{k+1}-z^k\| + \|\mu^{k+1}-\mu^k\| = 0$*;*

(2). any limit point of $\{(x^k, y^k, z^k, \mu^k)\}$ *is a stationary point of* (3.3)*.*

Proof. (1). From Lemma 3.4.3, we have that

$$\lim_{k\to+\infty} \|x^{k+1}-x^k\| = \lim_{k\to+\infty} \|y^{k+1}-y^k\| = \lim_{k\to+\infty} \|z^{k+1}-z^k\| = 0,$$

hence immediately obtain that $\lim\limits_{k\to+\infty} \|\mu^{k+1}-\mu^k\| = 0$ by virtue of Lemma 3.4.2.

(2). The existence of a limit point is guaranteed by the boundedness of the sequence (see Lemma 3.4.4). Let (x^*, y^*, z^*, μ^*) be an arbitrary limit point of the sequence $\{x^k, y^k, z^k, \mu^k\}$ and $\mathcal{K}$ be the corresponding infinite subsequence such that $\lim\limits_{k\in\mathcal{K}}(x^k, y^k, z^k, \mu^k) = (x^*, y^*, z^*, \mu^*)$.

Clearly, the following relation holds due to (3.4a), (3.4b), and the statement (1)

$$\lim_{k\to+\infty} \|x^{k+1}-\hat{x}^k\| + \|y^{k+1}-\hat{y}^k\| + \|z^{k+1}-\hat{z}^k\| + \|\mu^{k+1}-\hat{\mu}^k\| = 0. \tag{3.13}$$

Then, combining (3.4f), we see that

$$\lim_{k\to+\infty} z_i^k - (x^k)^{\mathrm{T}} M_i y^k = \lim_{k\to+\infty} \frac{\mu_i^{k+1}-\hat{\mu}_i^k}{\rho} = 0,\ \ i \in I. \tag{3.14}$$

Now, we show that (x^*, y^*, z^*, μ^*) satisfies (3.6a)-(3.6d). For (3.6a), as $\nabla_x L_{\tau,\rho}$ is Lipschitz continuous on any bounded set, then, by virtue of (3.13), (3.14), and the statement (1), we have that

$$\lim_{k\in\mathcal{K}} \nabla_x L_{\tau,\rho}(x^{k+1}, y^k, \hat{z}^k, \hat{\mu}^k) + \beta\left(x^{k+1} - \hat{x}^k\right) = \nabla_x L_\tau(x^*, y^*, z^*, \mu^*).$$

Therefore, taking the limit of (3.9a) on $\mathcal{K}$, we obtain that

$$\langle \nabla_x L_\tau(x^*, y^*, z^*, \mu^*), x - x^* \rangle \geq 0 \quad \text{for all} \quad x \in X,$$

which is identical to (3.6a). The proof of (3.6b) is similar to (3.6a). Furthermore, (3.6c) and (3.6d) can be straightly verified by (3.10) and (3.14), respectively. Thus, the proof is completed. □

The global convergence properties in objective values and in iterates to the set of stationary points are also explored on the basis of Theorem 3.4.6.

Proposition 3.4.7. *Let $\{x^k\}$, $\{y^k\}$, $\{z^k\}$, and $\{\mu^k\}$ be the sequences generated by Algorithm 1 and Assumption 3.4.1 hold. Then*

(1). $\lim_{k\to+\infty} \Psi_\tau(x^k, y^k, z^k) = \Psi_\tau(x^*, y^*, z^*)$, where (x^*, y^*, z^*) represents an arbitrary limit point of $\{(x^k, y^k, z^k)\}$;

(2). $\lim_{k\to+\infty} \operatorname{dist}\left((x^k, y^k, z^k, \mu^k), \Omega^*\right) = 0$, where Ω^* represents the set of stationary points of (3.3).

Proof. (1). The statement (1) comes from Lemma 3.4.3, Theorem 3.4.6(1), (3.14), and the lower semi-continuity of Ψ_τ. More precisely, there exists an infinite subsequence $\mathcal{K}$ such that

$$\lim_{k\to+\infty} \Upsilon^k_{\tau,\rho,\alpha,\beta} = \lim_{k\to+\infty} \Psi_\tau(x^k, y^k, z^k) = \lim_{k\in\mathcal{K}} \Psi_\tau(x^k, y^k, z^k) = \Psi_\tau(x^*, y^*, z^*).$$

(2). We assume by contradiction that the distance is not convergent to 0. Then there exist $\bar{\epsilon} > 0$ and an infinite subsequence $\mathcal{K}_0$ such that

$$\operatorname{dist}\left((x^k, y^k, z^k, \mu^k), \Omega^*\right) \geq \bar{\epsilon} \quad \text{for all} \quad k \in \mathcal{K}_0.$$

As $\left\{(x^k, y^k, z^k, \mu^k)\right\}_{k \in \mathcal{K}_0}$ is bounded (see Lemma 3.4.4), this sequence has a limit point. Let $(\dot{x}, \dot{y}, \dot{z}, \dot{\mu})$ be an arbitrary limit point of $\left\{(x^k, y^k, z^k, \mu^k)\right\}_{k \in \mathcal{K}_0}$. Then there exists an infinite subsequence $\mathcal{K}_1 \subseteq \mathcal{K}_0$ such that

$$\operatorname{dist}\left((\dot{x}, \dot{y}, \dot{z}, \dot{\mu}), \Omega^*\right) = \lim_{k \in \mathcal{K}_1} \operatorname{dist}\left((x^k, y^k, z^k, \mu^k), \Omega^*\right) \geq \bar{\epsilon} > 0,$$

in contradiction to the fact that $(\dot{x}, \dot{y}, \dot{z}, \dot{\mu}) \in \Omega^*$ (see Theorem 3.4.6(2)). □

3.4.3 Iteration Complexity

In analyzing the computational complexity, we select $\operatorname{dist}^2\left(0, \partial^L \mathcal{L}_{\tau,\rho}(x^k, y^k, z^k, \mu^k)\right)$ to measure the progress of the algorithm.

Theorem 3.4.8. *Let $\{x^k\}$, $\{y^k\}$, $\{z^k\}$, and $\{\mu^k\}$ be the sequences generated by Algorithm 1, Assumption 3.4.1 hold, and $\delta > 0$. Then there exists $C > 0$ such that*

$$\min_{1 \leq k \leq K} \operatorname{dist}^2\left(0, \partial^L \mathcal{L}_{\tau,\rho}(x^k, y^k, z^k, \mu^k)\right) \leq \delta, \tag{3.15}$$

where $K := \left\lceil \frac{C(\Upsilon^0_{\tau,\rho,\alpha,\beta} + \Upsilon^1_{\tau,\rho,\alpha,\beta} - 2\underline{q}_0)}{\delta} \right\rceil + 1$.

Proof. For any $k \geq 2$, there exist C_1, C_2, $C > 0$ such that

$$\begin{aligned}
&\operatorname{dist}^2\left(0, \partial^L \mathcal{L}_{\tau,\rho}(x^k, y^k, z^k, \mu^k)\right) \\
\leq\ & C_1\big(\|x^k - x^{k-1}\|^2 + \|y^k - y^{k-1}\|^2 + \|z^k - z^{k-1}\|^2 + \|\mu^k - \mu^{k-1}\|^2 \\
&\quad + \|x^{k-1} - x^{k-2}\|^2 + \|y^{k-1} - y^{k-2}\|^2 + \|z^{k-1} - z^{k-2}\|^2 + \|\mu^{k-1} - \mu^{k-2}\|^2\big) \\
\leq\ & C_2\big(\|x^k - x^{k-1}\|^2 + \|y^k - y^{k-1}\|^2 + \|z^k - z^{k-1}\|^2 \\
&\quad + \|x^{k-1} - x^{k-2}\|^2 + \|y^{k-1} - y^{k-2}\|^2 + \|z^{k-1} - z^{k-2}\|^2\big) \\
\leq\ & C(\Upsilon^{k-2}_{\tau,\rho,\alpha,\beta} - \Upsilon^{k}_{\tau,\rho,\alpha,\beta}),
\end{aligned}$$

where the existence of C_1, C_2, and C comes from Lemmas 3.4.5, 3.4.2, and 3.4.3, respectively. Summing the left-hand side and right-hand side of the above inequality over $k = 2, \cdots, K$, we obtain that

$$\begin{aligned}
&\sum_{k=2}^{K} \operatorname{dist}^2(0, \partial^L \mathcal{L}_{\tau,\rho}(x^k, y^k, z^k, \mu^k)) \\
&\leq C(\Upsilon^0_{\tau,\rho,\alpha,\beta} + \Upsilon^1_{\tau,\rho,\alpha,\beta} - \Upsilon^{K-1}_{\tau,\rho,\alpha,\beta} - \Upsilon^{K}_{\tau,\rho,\alpha,\beta}) \\
&\leq C(\Upsilon^0_{\tau,\rho,\alpha,\beta} + \Upsilon^1_{\tau,\rho,\alpha,\beta} - 2\underline{q}_0),
\end{aligned}$$

where the second inequality holds due to Lemma 3.4.3(2). Then, we see that

$$\min_{1\leq k\leq K} \operatorname{dist}^2\left(0, \partial^L \mathcal{L}_{\tau,\rho}(x^k, y^k, z^k, \mu^k)\right) \leq \frac{C(\Upsilon^0_{\tau,\rho,\alpha,\beta} + \Upsilon^1_{\tau,\rho,\alpha,\beta} - 2\underline{q}_0)}{K-1}.$$

Thus, the proof is completed. □

As indicated in Theorem 3.4.8, the iteration complexity of Algorithm 1 is $\mathcal{O}(1/k)$. This complexity is the same as existing research on convex and nonconvex ADMM-type algorithms without extrapolation (see He & Yuan (2012, 2015) and Hong et al. (2016)) and better than the rate $\mathcal{O}(1/\sqrt{k})$ in the work of Chen et al. (2015) on ADMM methods with extrapolation for a linearly constrained convex problem.

3.4.4 Global Convergence

The KŁ theory makes a major contribution to our global convergence analysis. As the potential function is a KŁ function, we can establish the global convergence of Algorithm 1 using the uniformized KŁ property (Proposition 3.2.4).

Theorem 3.4.9. *Let $\{x^k\}$, $\{y^k\}$, $\{z^k\}$, and $\{\mu^k\}$ be the sequences generated by Algorithm 1 and Assumption 3.4.1 hold. Then $\{(x^k, y^k, z^k, \mu^k)\}$ is convergent.*

Proof. Noting the descent and convergence of $\{\Upsilon^k_{\tau,\rho,\alpha,\beta}\}$ (see Lemma 3.4.3), we write $\Upsilon^*_{\tau,\rho,\alpha,\beta} := \lim_{k\to+\infty} \Upsilon^k_{\tau,\rho,\alpha,\beta}$ and consider two cases: $\Upsilon^k_{\tau,\rho,\alpha,\beta} = \Upsilon^*_{\tau,\rho,\alpha,\beta}$ for some $k \in \mathbb{N}$, and $\Upsilon^k_{\tau,\rho,\alpha,\beta} > \Upsilon^*_{\tau,\rho,\alpha,\beta}$ for all $k \in \mathbb{N}$. If there is $k \in \mathbb{N}$ such that $\Upsilon^k_{\tau,\rho,\alpha,\beta} = \Upsilon^*_{\tau,\rho,\alpha,\beta}$, then $\{(x^k, y^k, z^k, \mu^k)\}$ is convergent finitely from Lemmas 3.4.2 and 3.4.3, thus the statement holds. Now, we consider the case that $\Upsilon^k_{\tau,\rho,\alpha,\beta} > \Upsilon^*_{\tau,\rho,\alpha,\beta}$ for all $k \in \mathbb{N}$. Let $w^k := (x^k - x^{k-1}, y^k - y^{k-1}, z^k - z^{k-1})$ and $\overline{\Omega}^*$ denote the set of limit points of $\{(x^k, y^k, z^k, w^k, \mu^k)\}$.

Initially, we verify that the uniformized KŁ property (Proposition 3.2.4) is satisfied for $\Upsilon_{\tau,\rho,\alpha,\beta}$. As $\{(x^k, y^k, z^k, \mu^k)\}$ is bounded (see Lemma 3.4.4), then $\overline{\Omega}^*$ is nonempty and bounded, thus compact by definition. It is also clear that $\Upsilon_{\tau,\rho,\alpha,\beta}$ is constant on $\overline{\Omega}^*$. Specifically, for any limit point $(x^*, y^*, z^*, w^*, \mu^*) \in \overline{\Omega}^*$, there exists an infinite subsequence $\mathcal{K}$ such that

$$\Upsilon_{\tau,\rho,\alpha,\beta}(x^*, y^*, z^*, w^*, \mu^*) = \lim_{k \in \mathcal{K}} \Upsilon^k_{\tau,\rho,\alpha,\beta} = \lim_{k \to +\infty} \Upsilon^k_{\tau,\rho,\alpha,\beta} (= \Upsilon^*_{\tau,\rho,\alpha,\beta}),$$

where the equalities follow from the lower semi-continuity of $\Upsilon_{\tau,\rho,\alpha,\beta}$ and Lemma 3.4.3, respectively. From Propositions 3.2.6 and 3.2.7, $\Upsilon_{\tau,\rho,\alpha,\beta}$ is a KŁ function. Thus, $\Upsilon_{\tau,\rho,\alpha,\beta}$ satisfies the uniformized KŁ property on $\overline{\Omega}^*$. Also, $\lim_{k \to +\infty} \Upsilon^k_{\tau,\rho,\alpha,\beta} = \Upsilon^*_{\tau,\rho,\alpha,\beta}$ and $\lim_{k \to +\infty} \operatorname{dist}((x^k, y^k, z^k, w^k, \mu^k), \overline{\Omega}^*) = 0$ hold due to the definition. Then, by virtue of Proposition 3.2.4, there exist $k_0 \geq 3$ and a concave function ϕ, for any $k \geq k_0$, one has that $\phi(\Upsilon^k_{\tau,\rho,\alpha,\beta} - \Upsilon^*_{\tau,\rho,\alpha,\beta}) > \phi(\Upsilon^{k+1}_{\tau,\rho,\alpha,\beta} - \Upsilon^*_{\tau,\rho,\alpha,\beta})$ and

$$\phi'(\Upsilon^k_{\tau,\rho,\alpha,\beta} - \Upsilon^*_{\tau,\rho,\alpha,\beta}) \operatorname{dist}\left(0, \partial^L \Upsilon_{\tau,\rho,\alpha,\beta}(x^k, y^k, z^k, w^k, \mu^k)\right) \geq 1. \tag{3.16}$$

Next, we show the convergence by making use of (3.16). Let

$$\sigma^k := \phi(\Upsilon^k_{\tau,\rho,\alpha,\beta} - \Upsilon^*_{\tau,\rho,\alpha,\beta}) - \phi(\Upsilon^{k+1}_{\tau,\rho,\alpha,\beta} - \Upsilon^*_{\tau,\rho,\alpha,\beta}).$$

Then, for any $k \geq k_0$, there exist $\theta_1 > 0$ and $\theta_2 > 0$ such that

$$
\begin{aligned}
&\theta_1\big(\|x^k - x^{k-1}\| + \|y^k - y^{k-1}\| + \|z^k - z^{k-1}\| + \|\mu^k - \mu^{k-1}\| \\
&\quad + \|x^{k-1} - x^{k-2}\| + \|y^{k-1} - y^{k-2}\| + \|z^{k-1} - z^{k-2}\| + \|\mu^{k-1} - \mu^{k-2}\|\big)\sigma^k \\
&\geq \operatorname{dist}\left(0, \partial^L \Upsilon_{\tau,\rho,\alpha,\beta}(x^k, y^k, z^k, w^k, \mu^k)\right)\sigma^k \\
&\geq \operatorname{dist}\left(0, \partial^L \Upsilon_{\tau,\rho,\alpha,\beta}(x^k, y^k, z^k, w^k, \mu^k)\right)\left(\Upsilon^k_{\tau,\rho,\alpha,\beta} - \Upsilon^{k+1}_{\tau,\rho,\alpha,\beta}\right)\phi'\left(\Upsilon^k_{\tau,\rho,\alpha,\beta} - \Upsilon^*_{\tau,\rho,\alpha,\beta}\right) \\
&\geq \Upsilon^k_{\tau,\rho,\alpha,\beta} - \Upsilon^{k+1}_{\tau,\rho,\alpha,\beta} \\
&\geq \theta_2\left(\|x^{k+1} - x^k\|^2 + \|y^{k+1} - y^k\|^2 + \|z^{k+1} - z^k\|^2\right) \\
&\geq \frac{\theta_2}{3}\left(\|x^{k+1} - x^k\| + \|y^{k+1} - y^k\| + \|z^{k+1} - z^k\|\right)^2,
\end{aligned}
\tag{3.17}
$$

where the first and fourth estimates are implied by Lemmas 3.4.5 and 3.4.3, respectively. The second and last inequalities follow from the concavity of ϕ and Young's inequality, respectively, and the third relation holds due to (3.16). We also estimate the upper bound of the left-hand side of (3.17). Let κ_0 be an arbitrary positive number. Then, for any $k \geq 3$, there exists $\theta_3 > 0$ such that

$$
\begin{aligned}
&\theta_1\big(\|x^k - x^{k-1}\| + \|y^k - y^{k-1}\| + \|z^k - z^{k-1}\| + \|\mu^k - \mu^{k-1}\| \\
&\quad + \|x^{k-1} - x^{k-2}\| + \|y^{k-1} - y^{k-2}\| + \|z^{k-1} - z^{k-2}\| + \|\mu^{k-1} - \mu^{k-2}\|\big)\sigma^k \\
&\leq \frac{1}{4}\bigg[\frac{\theta_1\sigma^k}{\kappa_0} + \kappa_0\big(\|x^k - x^{k-1}\| + \|y^k - y^{k-1}\| + \|z^k - z^{k-1}\| + \|\mu^k - \mu^{k-1}\| \\
&\quad + \|x^{k-1} - x^{k-2}\| + \|y^{k-1} - y^{k-2}\| + \|z^{k-1} - z^{k-2}\| + \|\mu^{k-1} - \mu^{k-2}\|\big)\bigg]^2 \qquad (3.18) \\
&\leq \frac{1}{4}\bigg[\frac{\theta_1\sigma^k}{\kappa_0} + \kappa_0(1+\theta_3)\big(\|x^k - x^{k-1}\| + \|y^k - y^{k-1}\| + \|z^k - z^{k-1}\| \\
&\quad + \|x^{k-1} - x^{k-2}\| + \|y^{k-1} - y^{k-2}\| + \|z^{k-1} - z^{k-2}\|\big)\bigg]^2,
\end{aligned}
$$

where the estimates come from the Young's inequality and Lemma 3.4.2, respectively.

Let $\kappa := \sqrt{\frac{3}{4\theta_2}}(1+\theta_3)\kappa_0$ and $\theta := \frac{3\theta_1}{4\theta_2}(1+\theta_3)$. From (3.17) and (3.18), for any $k \geq k_0$, it holds that

$$\begin{aligned}
&\|x^{k+1}-x^k\| + \|y^{k+1}-y^k\| + \|z^{k+1}-z^k\| \\
\leq\ & \kappa\big(\|x^k - x^{k-1}\| + \|y^k - y^{k-1}\| + \|z^k - z^{k-1}\| \\
&+ \|x^{k-1} - x^{k-2}\| + \|y^{k-1} - y^{k-2}\| + \|z^{k-1} - z^{k-2}\|\big) + \frac{\theta}{\kappa}\sigma^k.
\end{aligned}$$

Rearranging terms, we see that, for any $k \geq k_0$,

$$\begin{aligned}
&(1-2\kappa)\big(\|x^{k+1}-x^k\| + \|y^{k+1}-y^k\| + \|z^{k+1}-z^k\|\big) \\
\leq\ & \kappa\big(\|x^k - x^{k-1}\| + \|x^{k-1} - x^{k-2}\| - 2\|x^{k+1} - x^k\| \\
&+ \|y^k - y^{k-1}\| + \|y^{k-1} - y^{k-2}\| - 2\|y^{k+1} - y^k\| \\
&+ \|z^k - z^{k-1}\| + \|z^{k-1} - z^{k-2}\| - 2\|z^{k+1} - z^k\|\big) + \frac{\theta}{\kappa}\sigma^k.
\end{aligned}$$

For $n \geq k_0$, summing the above inequality over $k = k_0, \cdots, n$, we obtain that

$$\begin{aligned}
&(1-2\kappa)\left(\sum_{k=k_0}^{n}\|x^{k+1}-x^k\| + \sum_{k=k_0}^{n}\|y^{k+1}-y^k\| + \sum_{k=k_0}^{n}\|z^{k+1}-z^k\|\right) \\
\leq\ & \kappa\big(2\|x^{k_0} - x^{k_0-1}\| + \|x^{k_0-1} - x^{k_0-2}\| - 2\|x^{n+1} - x^n\| - \|x^n - x^{n-1}\| \\
&+ 2\|y^{k_0} - y^{k_0-1}\| + \|y^{k_0-1} - y^{k_0-2}\| - 2\|y^{n+1} - y^n\| - \|y^n - y^{n-1}\| \\
&+ 2\|z^{k_0} - z^{k_0-1}\| + \|z^{k_0-1} - z^{k_0-2}\| - 2\|z^{n+1} - z^n\| - \|z^n - z^{n-1}\|\big) \\
&+ \frac{\theta}{\kappa}\left[\phi(\Upsilon^{k_0}_{\tau,\rho,\alpha,\beta} - \Upsilon^*_{\tau,\rho,\alpha,\beta}) - \phi(\Upsilon^{n+1}_{\tau,\rho,\alpha,\beta} - \Upsilon^*_{\tau,\rho,\alpha,\beta})\right].
\end{aligned}$$

When $0 < \kappa < \frac{1}{2}$ and $n \to +\infty$, the above inequality together with Lemma 3.4.2 implies that

$$\sum_{k=k_0}^{+\infty}\|x^{k+1}-x^k\| + \|y^{k+1}-y^k\| + \|z^{k+1}-z^k\| + \|\mu^{k+1}-\mu^k\| < +\infty.$$

Thus, the proof is completed. □

3.5 Convergence Analysis of Outer Algorithm

The convergence of the outer iterations is established under the condition that the penalty parameter τ tends to be infinite and the inner algorithm tends to be exactly convergent.

Theorem 3.5.1. *Let* $\{\tilde{x}^r\}$, $\{\tilde{y}^r\}$, $\{\tilde{z}^r\}$, *and* $\{\tilde{\mu}^r\}$ *be the sequences generated by Algorithm 2 and* $\tilde{\nu}_i^r := \tau^r(\tilde{z}_i^r + c_i^{\mathrm{T}}\tilde{x}^r + d_i^{\mathrm{T}}\tilde{y}^r + e_i)$, $i \in I$. *Suppose that* $\lim\limits_{r\to+\infty} \tau^r = +\infty$ *and* $\lim\limits_{r\to+\infty} \operatorname{dist}\left(0, \partial^L \mathcal{L}_{\tau^r}(\tilde{x}^r, \tilde{y}^r, \tilde{z}^r, \tilde{\mu}^r)\right) = 0$. *If* $\{(\tilde{x}^r, \tilde{y}^r, \tilde{z}^r, \tilde{\mu}^r, \tilde{\nu}^r)\}$ *has a limit point, then any limit point is a stationary point of* (3.2).

Proof. Suppose that $\{(\tilde{x}^r, \tilde{y}^r, \tilde{z}^r, \tilde{\mu}^r, \tilde{\nu}^r)\}$ has a limit point. Let $(x^\star, y^\star, z^\star, \mu^\star, \nu^\star)$ be an arbitrary limit point and $\mathcal{R}$ be the corresponding infinite subsequence such that $\lim\limits_{r\in\mathcal{R}} (\tilde{x}^r, \tilde{y}^r, \tilde{z}^r, \tilde{\mu}^r, \tilde{\nu}^r) = (x^\star, y^\star, z^\star, \mu^\star, \nu^\star)$.

It follows from Proposition 1.5.4 and $\lim\limits_{r\to+\infty} \operatorname{dist}\left(0, \partial^L \mathcal{L}_{\tau^r}(\tilde{x}^r, \tilde{y}^r, \tilde{z}^r, \tilde{\mu}^r)\right) = 0$ that

$$\lim_{r\to+\infty} \operatorname{dist}\left(-\nabla_x L_{\tau^r}(\tilde{x}^r, \tilde{y}^r, \tilde{z}^r, \tilde{\mu}^r), N_X(\tilde{x}^r)\right) = 0 \tag{3.19a}$$

$$\lim_{r\to+\infty} \operatorname{dist}\left(-\nabla_y L_{\tau^r}(\tilde{x}^r, \tilde{y}^r, \tilde{z}^r, \tilde{\mu}^r), N_Y(\tilde{y}^r)\right) = 0 \tag{3.19b}$$

$$\lim_{r\to+\infty} \tau^r(\tilde{z}_i^r + c_i^{\mathrm{T}}\tilde{x}^r + d_i^{\mathrm{T}}\tilde{y}^r + e_i) + \tilde{\mu}_i^r = 0,\ i \in I \tag{3.19c}$$

$$\lim_{r\to+\infty} \tilde{z}_i^r - (\tilde{x}^r)^{\mathrm{T}} M_i \tilde{y}^r = 0,\ i \in I. \tag{3.19d}$$

Now, we show that $(x^\star, y^\star, z^\star, \mu^\star, \nu^\star)$ satisfies (3.5a)-(3.5e). Initially, since $\tilde{\nu}_i^r$ is given by $\tilde{\nu}_i^r := \tau^r(\tilde{z}_i^r + c_i^{\mathrm{T}}\tilde{x}^r + d_i^{\mathrm{T}}\tilde{y}^r + e_i)$, (3.5a)-(3.5d) can be straightly verified by (3.19a)-(3.19d), respectively. For (3.5e), combining (3.19c) and $\lim\limits_{r\to+\infty} \tau^r = +\infty$, we see that

$$z_i^\star + c_i^{\mathrm{T}} x^\star + d_i^{\mathrm{T}} y^\star + e_i = \lim_{r\in\mathcal{R}}\ \tilde{z}_i^r + c_i^{\mathrm{T}}\tilde{x}^r + d_i^{\mathrm{T}}\tilde{y}^r + e_i = \lim_{r\in\mathcal{R}} -\frac{\tilde{\mu}_i^r}{\tau^r} = 0,\ i \in I.$$

Thus, the proof is completed. □

As the inner algorithm is convergent to a primal-dual stationary point, the convergence of Algorithm 2 is also obtained in a primal-dual framework. This is different from the classical convergence result of the quadratic penalty method (see Nocedal & Wright (2006, Theorem 17.2)).

At the end of the section, we discuss how to make the penalty problem (3.3) less ill-conditioned by adjusting β^r in Algorithm 2.

Remark 3.5.2. *With τ^r increasing, the Hessian matrices in (3.4c) and (3.4d) possibly become ill-conditioned, leading to computational difficulties. The respective Hessian matrices in (3.4c) and (3.4d) are*

$$\mathcal{H}_x^k = \tau^r \sum_{i \in I} c_i c_i^{\mathrm{T}} + \rho^r \sum_{i \in I} M_i y^k (y^k)^{\mathrm{T}} M_i^{\mathrm{T}} + \beta^r \mathbf{E}_m$$

and

$$\mathcal{H}_y^k = \tau^r \sum_{i \in I} d_i d_i^{\mathrm{T}} + \rho^r \sum_{i \in I} M_i^{\mathrm{T}} x^{k+1} (x^{k+1})^{\mathrm{T}} M_i + \beta^r \mathbf{E}_n,$$

where $\mathbf{E}_m$ and $\mathbf{E}_n$ are the identity matrices of size m and size n, respectively.

For two symmetric matrices G and H, it follows from the Weyl's inequality (see Weyl (1912) and Fomin et al. (2005)) that

$$\begin{cases} \lambda_{max}(G+H) \leq \lambda_{max}(G) + \lambda_{max}(H) \\ \lambda_{min}(G+H) \geq \lambda_{min}(G) + \lambda_{min}(H), \end{cases}$$

where $\lambda_{\max}(\cdot)/\lambda_{\min}(\cdot)$ denotes the largest/smallest eigenvalue. Then, the condition numbers $\mathcal{K}(\mathcal{H}_x^k)$ and $\mathcal{K}(\mathcal{H}_y^k)$ respectively satisfy that

$$\mathcal{K}(\mathcal{H}_x^k) \leq \frac{\tau^r \lambda_{max}(\sum_{i \in I} c_i c_i^{\mathrm{T}}) + \rho^r \lambda_{max}(\sum_{i \in I} M_i y^k (y^k)^{\mathrm{T}} M_i^{\mathrm{T}}) + \beta^r}{\tau^r \lambda_{min}(\sum_{i \in I} c_i c_i^{\mathrm{T}}) + \rho^r \lambda_{min}(\sum_{i \in I} M_i y^k (y^k)^{\mathrm{T}} M_i^{\mathrm{T}}) + \beta^r}$$

and

$$\mathcal{K}(\mathcal{H}_y^k) \leq \frac{\tau^r \lambda_{max}(\sum_{i \in I} d_i d_i^{\mathrm{T}}) + \rho^r \lambda_{max}(\sum_{i \in I} M_i^{\mathrm{T}} x^{k+1} (x^{k+1})^{\mathrm{T}} M_i) + \beta^r}{\tau^r \lambda_{min}(\sum_{i \in I} d_i d_i^{\mathrm{T}}) + \rho^r \lambda_{min}(\sum_{i \in I} M_i^{\mathrm{T}} x^{k+1} (x^{k+1})^{\mathrm{T}} M_i) + \beta^r}.$$

Thus, the ill-conditioning difficulty can be partly overcome if we choose β^r such that $\frac{\tau^r}{\beta^r}$ and $\frac{\rho^r}{\beta^r}$ fall into an appropriate range.

3.6 Numerical Experiments

This section provides numerical studies of the proposed algorithms. To see the effectiveness of extrapolation, we initially conduct an experiment on the inner algorithm with different extrapolation steps. Then, the outer algorithm is specified to a linear sum-of-ratios problem and a structured quadratically constrained quadratic programming problem, respectively. All the experiments are completed in MATLAB R2021a and macOS 11.6 on a 64-bit PC with an i5-5250U CPU and 4GB RAM.

Stopping criteria of the inner and outer algorithms are respectively designed as follows:

$$\begin{aligned}\text{Accuracy} := \max \Bigg\{ & \frac{\left|\Psi_{\tau^r}(x^{k+1}, y^{k+1}, z^{k+1}) - \Psi_{\tau^r}(x^k, y^k, z^k)\right|}{|\Psi_{\tau^r}(x^k, y^k, z^k)| + 1}\ , \\ & \sum_{i \in I} \left(z_i^{k+1} - (x^{k+1})^{\mathrm{T}} M_i y^{k+1}\right)^2 \Bigg\} \leq \mathrm{Tol}_{in}\end{aligned} \tag{3.20}$$

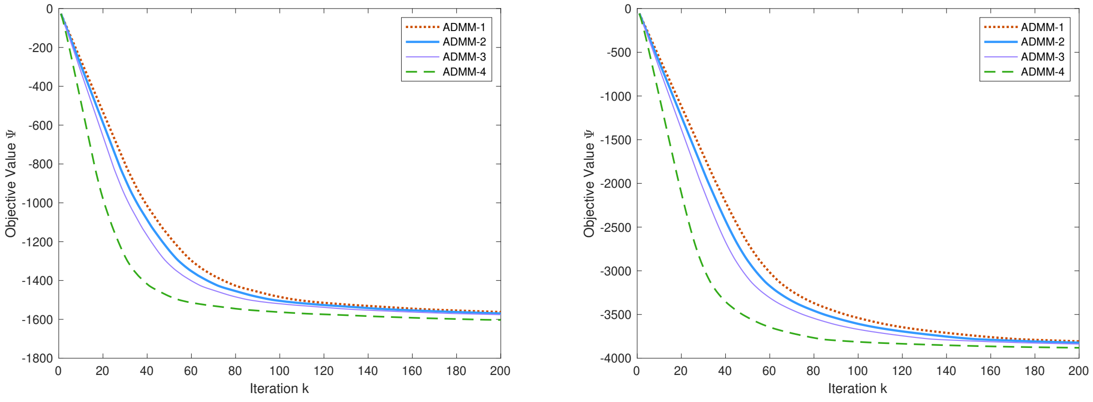
and

$$\text{Feasibility} := \sum_{i \in I} q_i(x^{k+1}, y^{k+1}, z_i^{k+1})^2 \leq \mathrm{Tol}_{out}. \tag{3.21}$$

For the outer algorithm, the penalty parameter τ^r is initialized as some given $\tilde{\tau}$ and updated as $10\tau^r$ at the next iteration, that is, $\tau^r = 10^r \tilde{\tau}$. When adopting the inner algorithm at the r^{th} outer iteration, we determine ρ^r and β^r by heuristics. Given $\tilde{\rho}$ and $\tilde{c}$, we start with $\rho^r = 10^r \tilde{\rho}$ and $\beta^r = \frac{\tau^r}{\tilde{c}}$. Then, if the descent of $\left\{\Upsilon^k_{\tau^r, \rho^r, \alpha, \beta^r}\right\}$ is met, we keep them unchanged; otherwise, we increase them at the next (inner) iteration until $\Upsilon^{k+1}_{\tau^r, \rho^r, \alpha, \beta^r} \leq \Upsilon^k_{\tau^r, \rho^r, \alpha, \beta^r}$. Remarkably, in this way, the parameters selected may not satisfy Assumption 3.4.1 (2). The following extrapolation parameters are tested in all the experiments

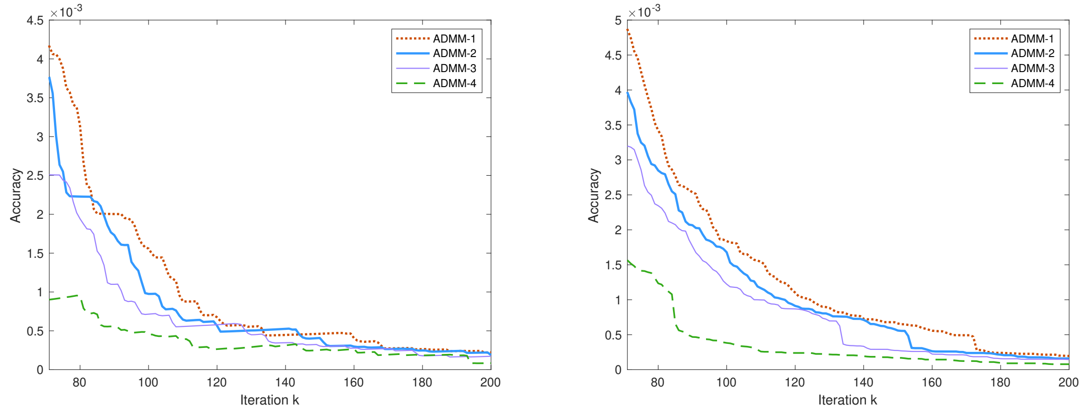
$$\begin{aligned} &\text{ADMM-1} : \alpha = (\alpha_x, \alpha_y, \alpha_z, \alpha_\mu) = (0, 0, 0, 0) \\ &\text{ADMM-2} : \alpha = (\alpha_x, \alpha_y, \alpha_z, \alpha_\mu) = (0.1, 0.1, 0.1, 0.2) \\ &\text{ADMM-3} : \alpha = (\alpha_x, \alpha_y, \alpha_z, \alpha_\mu) = (0.2, 0.2, 0.2, 0.5) \\ &\text{ADMM-4} : \alpha = (\alpha_x, \alpha_y, \alpha_z, \alpha_\mu) = (0.5, 0.5, 0.5, 1). \end{aligned} \tag{3.22}$$

3.6.1 Effect of Extrapolation Strategy



(a). $(m, n, p) = (120, 80, 20)$ (b). $(m, n, p) = (500, 200, 100)$

Figure 3.1: Convergence behavior (objective value)



(a). $(m, n, p) = (120, 80, 20)$ (b). $(m, n, p) = (500, 200, 100)$

Figure 3.2: Convergence behavior (Accuracy)

To test the effect of extrapolation, we carry out the first experiment on Algorithm 1 for the quadratic approximation (3.3) with the extrapolation parameters in (3.22). We select the coefficients from data randomly generated under the uniform distribution within the following intervals

$$c_i \in [-2, 2], \quad d_i, M_i \in [-3, 3], \quad \text{and} \quad e_i \in [-1, 1].$$

The inequality constraint sets are given by

$$X = \{x : -10 \le x \le 10\} \quad \text{and} \quad Y = \{y : -10 \le y \le 10\}.$$

The initial point is taken as the origin. The parameters in Algorithm 1 are set as $\tau = 10^4$, $\tilde{\rho} = 10^2$, and $\tilde{c} = 10^3$. Two problem sizes $(m, n, p) = (120, 80, 20)$ and $(500, 200, 100)$ are considered in the testing.

Figure 3.1 (resp. Figure 3.2) plots the objective value (resp. 'Accuracy' (see (3.20))) against the number of iterations for the four extrapolation rules. We observe that Algorithm 1 with 'ADMM-4' converges faster than others in objective values, and it is always the first to terminate for any stopping tolerance. The extrapolation with large parameters performs better in this experiment.

3.6.2 Linear Sum-of-Ratios Problem

In the second experiment, we evaluate the performance of the outer algorithm for a linear sum-of-ratios problem with 'ADMM-1' to 'ADMM-4' in (3.22).

A linear sum-of-ratios problem (see Benson (2007) and Jiao & Liu (2015)) has the structure

$$\begin{aligned} \min_x \quad & \sum_{i=1}^{n} \frac{c_i^{\mathrm{T}} x + p_i}{d_i^{\mathrm{T}} x + q_i} \\ \text{s.t.} \quad & Ax \le b, \end{aligned}$$

where $c_i^{\mathrm{T}} x + p_i \ge 0$ and $d_i^{\mathrm{T}} x + q_i > 0$ $(i = 1, \cdots, n)$ are assumed for all the feasible solutions. A linear fractional function is strongly associated with a bilinear function.

Algorithm 3 Penalty extrapolated proximal ADMM for solving (3.24)

```
Let the sequences $\{\tau^r\}$, $\{\rho^r\}$, and $\{\beta^r\}$, and quartet $(\alpha_x, \alpha_y, \alpha_z, \alpha_\mu)$ be given.
While $r = 0, 1, 2, \cdots,$ do
  Initialization $(x^0, y^0, z^0, \mu^0) = (x^{-1}, y^{-1}, z^{-1}, \mu^{-1})$.
  While $k = 0, 1, 2, \cdots,$ do
  - $\hat{x}^k = x^k + \alpha_x(x^k - x^{k-1})$, $\hat{y}^k = y^k + \alpha_y(y^k - y^{k-1})$, $\hat{z}^k = z^k + \alpha_z(z^k - z^{k-1})$, and
    $\hat{\mu}^k = \mu^k + \alpha_\mu(\mu^k - \mu^{k-1})$.
  - $x^{k+1} = \arg\min_{x \in X}\{L_{\tau^r,\rho^r}(x, y^k, \hat{z}^k, \hat{\mu}^k) + \frac{\beta^r}{2}\|x - \hat{x}^k\|^2\}$.
  - $y_i^{k+1} = \frac{(\hat{\mu}_i^k + \rho^r \hat{z}_i^k)(x^{k+1})^{\mathrm{T}} d_i - 1 - \tau^r q_i(\hat{z}_i^k - c_i^{\mathrm{T}} x^{k+1} - p_i) + \beta^r \hat{y}_i^k}{\tau^r q_i^2 + \rho((x^{k+1})^{\mathrm{T}} d_i)^2 + \beta^r}, \; i = 1, \cdots, n.$
  - $z_i^{k+1} = \frac{\tau^r(c_i^{\mathrm{T}} x^{k+1} - q_i y_i^{k+1} + p_i) - \hat{\mu}_i^k + \rho^r (x^{k+1})^{\mathrm{T}} d_i y_i^{k+1}}{\tau^r + \rho^r}, \; i = 1, \cdots, n.$
  - $\mu_i^{k+1} = \hat{\mu}^k + \rho^r[z_i^{k+1} - (x^{k+1})^{\mathrm{T}} d_i y_i^{k+1}], \; i = 1, \cdots, n.$
  End while
End while
```

Letting $y_i = \frac{c_i^{\mathrm{T}} x + p_i}{d_i^{\mathrm{T}} x + q_i}$, we obtain a bilinear constraint $x^{\mathrm{T}} d_i y_i - c_i^{\mathrm{T}} x + q_i y_i - p_i = 0$. In this way, any linear sum-of-ratios problem can be reformulated as a GBLP problem.

In this experiment, we consider the following linear sum-of-ratios problem

$$\begin{aligned} \min_{x} \quad & \sum_{i=1}^{n} \frac{c_i^{\mathrm{T}} x + p_i}{d_i^{\mathrm{T}} x + q_i} \\ \text{s.t.} \quad & c_i^{\mathrm{T}} x + p_i \geq 0, \; d_i^{\mathrm{T}} x + q_i \geq 0, \; i = 1, \cdots, n \\ & Ax \leq b, \; \underline{x} \leq x \leq \overline{x}, \end{aligned} \tag{3.23}$$

where $x, \underline{x}, \overline{x}, c_i, d_i \in \mathbb{R}^m$, $p_i, q_i \in \mathbb{R}$, $A \in \mathbb{R}^{l \times m}$, and $b \in \mathbb{R}^l$. Remarkably, $c_i^{\mathrm{T}} x + p_i \geq 0$ and $d_i^{\mathrm{T}} x + q_i \geq 0$ are additional constraints to guarantee the assumption of (3.23), and the bound constraint $\underline{x} \leq x \leq \overline{x}$ is added to make the optimal value finite.

With auxiliary variables $y := (y_1, \cdots, y_n)$ and $z := (z_1, \cdots, z_n)$, we transform

(3.23) into a GBLP structure

$$\begin{aligned} \min_{x,y,z} \quad & \sum_{i=1}^{n} y_i \\ \text{s.t.} \quad & z_i - c_i^{\mathrm{T}} x + q_i y_i - p_i = 0,\ i = 1, \cdots, n \\ & z_i = x^{\mathrm{T}} d_i y_i,\ i = 1, \cdots, n \\ & c_i^{\mathrm{T}} x + p_i \geq 0,\ d_i^{\mathrm{T}} x + q_i \geq 0,\ i = 1, \cdots, n \\ & Ax \leq b,\ \underline{x} \leq x \leq \overline{x}. \end{aligned} \tag{3.24}$$

Let $X := \left\{x : c_i^{\mathrm{T}} x + p_i \geq 0,\ d_i^{\mathrm{T}} x + q_i \geq 0,\ i = 1, \cdots, n;\ Ax \leq b;\ \underline{x} \leq x \leq \overline{x}\right\}$ and

$$\begin{aligned} L_{\tau^r, \rho^r}(x, y, z, \mu) := \sum_{i=1}^{n} \Big[& y_i + \frac{\tau^r}{2}(z_i - c_i^{\mathrm{T}} x + q_i y_i - p_i)^2 \\ & + \mu_i(z_i - x^{\mathrm{T}} d_i y_i) + \frac{\rho^r}{2}(z_i - x^{\mathrm{T}} d_i y_i)^2 \Big]. \end{aligned}$$

Then Algorithm 3 can be straightly applied to (3.24).

We select the initial point x^0 and coefficients c_i, d_i, and A from data randomly generated under the uniform distribution within $[-2, 2]$. The bound coefficients are given by $\underline{x} = -10$ and $\overline{x} = 10$. To make the feasible set nonempty, we produce the rest coefficients based on x^0. Specifically, let $p_i = -c_i^{\mathrm{T}} x^0 + \epsilon$, $q_i = -d_i^{\mathrm{T}} x^0 + \epsilon$, and $b = Ax^0 + \epsilon$, where ϵ is randomly generated under the uniform distribution within $[0, 2]$. The initial points of y, z, and μ are taken as $y_i^0 = \frac{c_i^{\mathrm{T}} x^0 + p_i}{d_i^{\mathrm{T}} x^0 + q_i}$, $z_i^0 = (x^0)^{\mathrm{T}} d_i y_i^0$, and $\mu^0 = 0$, respectively. Three problem sizes $(m, n, l) = (200, 100, 50)$, $(400, 100, 50)$, and $(400, 200, 100)$ are considered in the testing, and we experiment on Algorithm 3 with two data sets for each problem size. The tolerance of the inner algorithm is chosen as $\mathrm{Tol}_{in} = 10^{-9}$, and the tolerance of the outer algorithm is chosen as $\mathrm{Tol}_{out} = 10^{-6}$. The parameters in Algorithm 3 are set as $\tilde{\tau} = 10^8$, $\tilde{\rho} = 10^9$, and $\tilde{c} = 10^7$.

Table 3.1: Computation result for linear sum-of-ratios problem

Algorithm	Problem Size			Data Set A			Data Set B		
	m	n	l	Fval	Time	Feasibility	Fval	Time	Feasibility
ADMM-1	200	100	50	0.4844	6.80	6.18E-08	0.8398	11.43	3.38E-07
ADMM-2				0.4844	1.25	3.24E-08	0.8399	12.22	4.39E-07
ADMM-3				0.4844	1.05	2.18E-08	0.8397	17.04	8.03E-07
ADMM-4				0.4844	5.81	4.47E-08	0.8397	5.69	4.63E-08
ADMM-1	400	100	50	0.0858	6.61	3.37E-08	0.4114	9.87	5.26E-08
ADMM-2				0.0858	2.79	2.56E-08	0.4114	4.49	4.76E-08
ADMM-3				0.0858	2.64	2.33E-08	0.4114	9.67	3.98E-08
ADMM-4				0.0858	5.10	2.86E-08	0.4114	6.00	3.61E-08
ADMM-1	400	200	100	0.5373	209.36	6.09E-08	0.3431	320.80	5.08E-07
ADMM-2				0.5472	270.85	1.57E-07	0.3431	301.40	1.70E-07
ADMM-3				0.5373	233.04	1.79E-07	0.3431	186.45	5.93E-08
ADMM-4				0.5371	63.97	9.70E-08	0.3430	104.85	7.08E-08

Computation results are presented in Table 3.1. The objective value (‘Fval’ in the table) and running time are measures of efficiency, and ‘Feasibility’ (see (3.21)) indicates the validness of the resulting solutions. In Table 3.1, there are few differences in objective values for the four extrapolation rules. In respect of the running time, Algorithm 3 with ‘ADMM-4’ outperforms the others on average. The use of extrapolation speeds up the algorithm, but it is not necessary that the larger extrapolation parameters, the better performance is. This phenomenon is also observed in Pock & Sabach (2016).

3.6.3 QCQP Problem

Eventually, we compare the proposed method with a semidefinite relaxation (SDR) method for a specially structured nonconvex quadratically constrained quadratic programming (QCQP) problem.

The QCQP problem, which plays a significant part in numerical applications (see Linderoth (2005), Luo et al. (2010), and Anstreicher (2012)), is given by

$$\begin{aligned} \min_{x} \quad & x^{\mathrm{T}} A_0 x \\ \text{s.t.} \quad & x^{\mathrm{T}} A_i x \trianglerighteq_i b_i,\ i = 1, \cdots, m, \end{aligned}$$

where "$\trianglerighteq_i$" ($i = 1, \cdots, m$) represents either "$\geq$", "$\leq$", or "$=$". It is worth noting that a quadratic term $x^{\mathrm{T}} A_i x$ is identical to a bilinear term $x^{\mathrm{T}} A_i y$ plus a linear constraint $x = y$. Therefore, we can convert the QCQP problem into a GBLP framework

$$\begin{aligned} \min_{x,y} \quad & x^{\mathrm{T}} A_0 y \\ \text{s.t.} \quad & x^{\mathrm{T}} A_i y \trianglerighteq_i b_i,\ i = 1, \cdots, m \\ & x = y. \end{aligned}$$

In this experiment, we consider the following nonconvex QCQP problem

$$\begin{aligned} \min_{x} \quad & x^{\mathrm{T}} A_0 x \\ \text{s.t.} \quad & x^{\mathrm{T}} A_i x \geq 1,\ i = 1, \cdots, p, \end{aligned} \tag{3.25}$$

where $x \in \mathbb{R}^n$ and $A_0, A_i \in \mathbb{R}^{n \times n}$ ($i = 1, \cdots, p$). Also, $A_0 \succeq 0$ and $A_i \succ 0$; that is, A_0 is symmetric and positive semidefinite, and A_i is symmetric and positive definite.

The SDR method is one of the most popular methods for solving QCQP problems, which focuses on a convex relaxation rather than the original problem (see Luo et al. (2010)). In particular, from

$$x^{\mathrm{T}} A_j x = \mathrm{Trace}(x^{\mathrm{T}} A_j x) = \mathrm{Trace}(A_j x x^{\mathrm{T}}),\ j = 0, 1, \cdots, p,$$

we obtain an equivalent formulation of (3.25)

$$\begin{aligned} \min_{X} \quad & \mathrm{Trace}(A_0 X) \\ \text{s.t.} \quad & \mathrm{Trace}(A_i X) \geq 1,\ i = 1, \cdots, p \\ & X \succeq 0,\ \mathrm{Rank}(X) = 1, \end{aligned}$$

where $X := xx^{\mathrm{T}} \in \mathbb{R}^{n\times n}$. The SDR method is to remove the nonconvex constraint $\mathrm{Rank}(X) = 1$ and solve the following convex semidefinite programming problem

$$\begin{aligned} \min_{X} \quad & \mathrm{Trace}(A_0X) \\ \text{s.t.} \quad & \mathrm{Trace}(A_iX) \geq 1, \ i = 1, \cdots, p \\ & X \succeq 0. \end{aligned} \tag{3.26}$$

Then, an optimal or approximate solution of (3.25) can be obtained by extracting the optimal solution X^* of (3.26). In fact, if X^* is rank-one, the extracted solution is an optimal solution. However, if not, the best result for this method is to reach a feasible solution as an approximate solution. For $\mathrm{Rank}(X^*) \neq 1$, an applicable extraction method is to use the eigendecomposition. Initially, take $\bar{x} = \sqrt{\lambda_{max}} q_{max}$, where λ_{max} is the largest eigenvalue of X^*, and q_{max} is the corresponding eigenvector. Then, the point $\bar{x}^* = \frac{\bar{x}}{\sqrt{\min_i \bar{x}^{\mathrm{T}} A_i \bar{x}}}$ is a feasible solution of (3.25). Luo et al. (2010) recommended the CVX toolbox (see Grant et al. (2020)) to solve (3.26). But here, we choose the SDTP3 package (see Toh et al. (1999)), as it is faster for this problem.

On the other hand, as mentioned above, (3.25) can be rewritten as

$$\begin{aligned} \min_{x,y,s,z} \quad & x^{\mathrm{T}} A_0 y \\ \text{s.t.} \quad & x = y, \ z_i - s_i - 1 = 0, \ i = 1, \cdots, p \\ & z_i = x^{\mathrm{T}} A_i y, \ i = 1, \cdots, p \\ & s_i \geq 0, \ i = 1, \cdots, p, \end{aligned} \tag{3.27}$$

where $y := (y_1, \cdots, y_n)$, $s := (s_1, \cdots, s_p)$, and $z := (z_1, \cdots, z_p)$ are auxiliary variables. We treat that (3.27) is bilinear with respect to x and (y, s) and let $\mathbf{E}$ be an identity matrix. Then Algorithm 4 can be straightly applied to (3.27).

In the testing, the coefficients are given by $A_0 = \mathbf{E}$ and $A_i = \bar{A}_i^{\mathrm{T}} \bar{A}_i + \mathbf{E}$, where $\bar{A}_i$ is randomly generated under the uniform distribution within $[-3, 3]$. We select

Algorithm 4 Penalty extrapolated proximal ADMM for solving (3.27)

```
Let the sequences $\{\tau^r\}$, $\{\rho^r\}$, and $\{\beta^r\}$, and the point $(\alpha_x, \alpha_y, \alpha_s, \alpha_z, \alpha_\mu)$ with $\alpha_s = \alpha_z$ be given.
While $r = 0, 1, 2, \cdots,$ do
  Initialization $(x^0, y^0, s^0, z^0, \mu^0) = (x^{-1}, y^{-1}, s^{-1}, z^{-1}, \mu^{-1})$.
  While $k = 0, 1, 2, \cdots,$ do
    - $\hat{x}^k = x^k + \alpha_x(x^k - x^{k-1})$, $\hat{y}^k = y^k + \alpha_y(y^k - y^{k-1})$, $\hat{s}^k = s^k + \alpha_y(s^k - s^{k-1})$, $\hat{z}^k = z^k + \alpha_z(z^k - z^{k-1})$, and $\hat{\mu}^k = \mu^k + \alpha_\mu(\mu^k - \mu^{k-1})$.
    - $x^{k+1} = \Big[\rho^r \sum_{i=1}^p A_i y^k (y^k)^{\mathrm{T}} A_i^{\mathrm{T}} + (\tau^r + \beta^r)\mathbf{E}\Big]^{-1} \Big[\tau^r y^k + \sum_{i=1}^p (\hat{\mu}_i^k + \rho^r \hat{z}_i^k) A_i y^k - A_0 y^k + \beta^r \hat{x}^k\Big]$.
    - $y^{k+1} = \Big[\rho^r \sum_{i=1}^p A_i^{\mathrm{T}} x^{k+1} (x^{k+1})^{\mathrm{T}} A_i + (\tau^r + \beta^r)\mathbf{E}\Big]^{-1} \Big[\tau^r x^{k+1} + \sum_{i=1}^p (\hat{\mu}_i^k + \rho^r \hat{z}_i^k) A_i^{\mathrm{T}} x^{k+1} - A_0^{\mathrm{T}} x^{k+1} + \beta^r \hat{y}^k\Big]$.
    - $s_i^{k+1} = \max\left\{\frac{\tau^r(\hat{z}_i^k - 1) + \beta^r \hat{s}_i^k}{\tau^r + \beta^r}, 0\right\}, \ i = 1, \cdots, p$.
    - $z_i^{k+1} = \frac{\tau^r(s_i^{k+1} + 1) - \hat{\mu}_i^k + \rho^r (x^{k+1})^{\mathrm{T}} A_i y^{k+1}}{\tau^r + \rho^r}, \ i = 1, \cdots, p$.
    - $\mu_i^{k+1} = \hat{\mu}^k + \rho^r[z_i^{k+1} - (x^{k+1})^{\mathrm{T}} A_i y^{k+1}], \ i = 1, \cdots, p$.
  End while
End while
```

the initial point of x as $x^0 = \frac{\bar{x}^0}{\sqrt{\min_i (\bar{x}^0)^{\mathrm{T}} A_i \bar{x}^0}}$, where $\bar{x}^0$ is randomly generated under the uniform distribution within $[-1, 1]$. The initial points of y, s, and μ are taken as $y^0 = x^0$, $s_i^0 = (x^0)^{\mathrm{T}} A_i y^0 - 1$, and $\mu^0 = 0$, respectively. The parameters of Algorithm 4 are set as $\tilde{\tau} = 10^6$, $\tilde{\rho} = 10^3$, and $\tilde{c} = 10^5$. As the SDR method produces feasible solutions, we consider $\mathrm{Tol}_{out} = 10^{-7}$ when adopting Algorithm 4, where the obtained solutions can be seen to be feasible.

Table 3.2 lists the optimal value and running time of the SDR method and Algorithm 4 under different problem sizes (n, p). Since all the extrapolated methods (see (3.22)) perform closely, we only present the result of Algorithm 4 with 'ADMM-1' as the representative. As shown in Table 3.2, the two methods are comparable in terms of the objective values. However, Algorithm 4 has a better performance on the running time for most data sets, especially for large-scale problems.

Table 3.2: Computation result for QCQP problem

Problem Size		SDR		Algorithm 4	
n	p	Fval	Time	Fval	Time
200	20	0.0355	1.29	0.0345	0.13
200	50	0.0089	2.03	0.0104	0.20
200	100	0.0047	3.67	0.0046	0.45
200	200	0.0019	13.12	0.0021	1.23
500	20	0.0684	2.92	0.0575	0.61
500	50	0.0136	5.22	0.0126	1.09
500	100	0.0049	13.95	0.0050	1.84
500	200	0.0022	48.13	0.0022	3.71
1000	20	0.0703	8.28	0.0587	4.99
1000	50	0.0128	14.79	0.0123	5.67
1000	100	0.0046	42.78.	0.0050	9.08
1000	200	0.0022	151.54	0.0023	30.76

Chapter 4

Sparse Minimax Portfolio and Sharpe Ratio Models

4.1 Introduction

In this chapter, we consider sparse portfolio models regularized by the l_p $(0 < p \leq 1)$ norm, where a risk measure (see Young (1998)) is used.

We first introduce an l_p-sparse $(0 < p \leq 1)$ minimax model and obtain a descent property of the l_p norm of the optimal portfolio with respect to the regularization parameter. That is to say, for this model, the regularization parameter can be a controller to adjust the level of sparsity and the space for short selling. In numerical studies, the l_1-sparse minimax model and l_p-sparse $(0 < p < 1)$ minimax model are examined separately due to their different natures on computation. Since the $l_{1/2}$ regularizer performs best among l_p $(0 < p \leq 1)$ regularizers (see Chartrand (2007) and Hu et al. (2017)), we take the $l_{1/2}$-sparse minimax model as the representative for $0 < p < 1$. The benchmarks are l_1-sparse and $l_{1/2}$-sparse mean-variance models and the equal-weighted rule. When comparing different sparse portfolio models, we observe their out-of-sample performance at the same level of sparsity. We find that, when the level of sparsity is extremely high, the $l_{1/2}$-sparse minimax model is more competitive than the l_1-sparse minimax model. However, as the resulting portfolios become less sparse, both models are comparable. The corresponding differences are not evident for the l_1-sparse and $l_{1/2}$-sparse mean-variance models.

Afterward, we construct a generalized l_1-sparse Sharpe ratio model based on the minimax risk measure in Young (1998). To avoid a zero denominator, we modify the original minimax risk measure by a pre-selected parameter to keep the denominator positive. When solving the proposed model, we transform it into a bilinear formulation and then adopt Algorithm 2 introduced in Chapter 3. Apart from it, we also design a parametric algorithm for this model, a global method extending the algorithm proposed by Konno & Kuno (1990). In numerical experiments, both algorithms are applied. The parametric method performs better on average in terms of the out-of-sample performance and stability. However, these superiorities are not significant, and Algorithm 2 has an advantage of the running time.

Now, we present the definitions and notations in this chapter. The l_p *regularizer* or l_p *norm* $(0 < p \leq 1)$ of the vector $x := (x_1, \cdots, x_n)$ is defined by

$$\|x\|_p := \left(\sum_{j=1}^{N} |x_j|^p \right)^{\frac{1}{p}}.$$

For all the models in Sections 4.2 and 4.3, we consider N securities. The portfolio under consideration is denoted by $w := (w_1, \cdots, w_N)$, where w_j $(j = 1, 2, \ldots, N)$ represents the percentage of the budget invested in security j. The security j is said to be *active* if $w_j \neq 0$. For simplicity, let $\mathbf{1} := (1, \cdots, 1) \in \mathbb{R}^N$.

The rest of the chapter is organized as follows. Initially, several sparse mean-variance models with l_p $(0 < p \leq 1)$ norm are briefly introduced in Section 4.2. Then, in Section 4.3, we establish an l_p-sparse minimax model and a generalized l_1-sparse minimax Sharpe ratio model and develop an extended parametric method for the second model. The chapter is ended by numerical studies for different sparse portfolio models in Section 4.4.

4.2 Sparse Mean-variance Models With l_p Norm

Given N securities with the expected rate of return vector $r \in \mathbb{R}^N$ and covariance matrix $\Sigma \subseteq \mathbb{R}^{N\times N}$, the *mean-variance model* (Markowitz, 1952) is formulated as

$$\begin{aligned} \min_{w} \quad & \frac{1}{2} w^{\mathrm{T}} \Sigma w \\ \text{s.t.} \quad & r^{\mathrm{T}} w = G,\ \mathbf{1}^{\mathrm{T}} w = 1. \end{aligned}$$

where the terms $w^{\mathrm{T}}\Sigma w$ and $r^{\mathrm{T}}w$ interpret the risk and return of w, respectively. The parameter $G \geq 0$ represents the desired rate of return, and the constraint $\mathbf{1}^{\mathrm{T}}w = 1$ means that the budget is fully invested. Moreover, $w \geq 0$ needs to be included if short selling is prohibited in the investment.

Based on the mean-variance model, Brodie et al. (2009) established the following l_1-regularized mean-variance model with the regularization parameter $\tau \geq 0$

$$\begin{aligned} \min_{w} \quad & \frac{1}{2} w^{\mathrm{T}} \Sigma w + \tau \left\| w \right\|_1 \\ \text{s.t.} \quad & r^{\mathrm{T}} w = G,\ \mathbf{1}^{\mathrm{T}} w = 1. \end{aligned}$$

Notably, as $\|w\|_1 = 1$ holds for all $w \geq 0$, the no-shorting constraint is not allowed when using the l_1 norm.

In this chapter, we consider a more general l_p-sparse $(0 < p \leq 1)$ mean-variance model, i.e.,

$$\begin{aligned} \min_{w} \quad & \frac{1}{2} w^{\mathrm{T}} \Sigma w + \tau \|w\|_p^p \\ \text{s.t.} \quad & r^{\mathrm{T}} w \geq G \\ & \mathbf{1}^{\mathrm{T}} w = 1,\ w \geq \alpha, \end{aligned} \tag{4.1}$$

which will be frequently mentioned in numerical studies as one of the benchmarks. In this model, the choice of τ influences the sparsity and short selling of the optimal portfolios (see Proposition 4.2.1). Similar results have also been provided in Brodie et al. (2009) for their l_1-regularized model.

Proposition 4.2.1. *Let w_τ be an optimal solution of (4.1) corresponding to a specific τ and w_τ^- be the componentwise negative part of w_τ.*

(1). For any $0 < p \leq 1$, one has that

$$(\tau_1 - \tau_2)\left(\|w_{\tau_2}\|_p^p - \|w_{\tau_1}\|_p^p\right) \geq 0.$$

(2). Furthermore, when $p = 1$, one has that

$$(\tau_1 - \tau_2)\left(\left\|w_{\tau_2}^-\right\|_1 - \left\|w_{\tau_1}^-\right\|_1\right) \geq 0.$$

Proof. (1). Indeed, we have that

$$\begin{aligned}
& \frac{1}{2}w_{\tau_1}^{\mathrm{T}}\Sigma w_{\tau_1} + \tau_1\left\|w_{\tau_1}\right\|_p^p \\
\leq\ & \frac{1}{2}w_{\tau_2}^{\mathrm{T}}\Sigma w_{\tau_2} + \tau_1\left\|w_{\tau_2}\right\|_p^p \\
=\ & \frac{1}{2}w_{\tau_2}^{\mathrm{T}}\Sigma w_{\tau_2} + \tau_2\left\|w_{\tau_2}\right\|_p^p + (\tau_1 - \tau_2)\left\|w_{\tau_2}\right\|_p^p \\
\leq\ & \frac{1}{2}w_{\tau_1}^{\mathrm{T}}\Sigma w_{\tau_1} + \tau_2\left\|w_{\tau_1}\right\|_p^p + (\tau_1 - \tau_2)\left\|w_{\tau_2}\right\|_p^p \\
=\ & \frac{1}{2}w_{\tau_1}^{\mathrm{T}}\Sigma w_{\tau_1} + \tau_1\left\|w_{\tau_1}\right\|_p^p + (\tau_1 - \tau_2)\left(\left\|w_{\tau_2}\right\|_p^p - \left\|w_{\tau_1}\right\|_p^p\right),
\end{aligned}$$

where the inequalities hold due to the minimization for w_{τ_1} and w_{τ_2}, respectively.

(2). Let w_τ^+ denote the componentwise positive part of w_τ. From the relations $\mathbf{1}^{\mathrm{T}}w_\tau = 1$, $\mathbf{1}^{\mathrm{T}}w_\tau = \|w_\tau^+\|_1 - \|w_\tau^-\|_1$, and $\|w_\tau\|_1 = \|w_\tau^+\|_1 + \|w_\tau^-\|_1$, we obtain that $\|w_\tau\|_1 = 1 + 2\|w_\tau^-\|_1$. Then,

$$\|w_{\tau_2}\|_1 - \|w_{\tau_1}\|_1 = 2\left(\left\|w_{\tau_2}^-\right\|_1 - \left\|w_{\tau_1}^-\right\|_1\right).$$

Therefore, the statement (2) can be proved by the statement (1). □

4.3 Sparse Minimax Models with l_p Norm

This section introduces the minimax portfolio model proposed by Young (1998) and then establishes an l_p-sparse $(0 < p \leq 1)$ minimax model and a generalized l_1-sparse minimax Sharpe ratio model. Besides, a parametric method is developed for finding a global solution of the l_1-sparse minimax Sharpe ratio model.

We observe N securities over T time periods and use R_{jt} to denote the rate of return of security j in time period t. Let $r \in \mathbb{R}^N$ be the expected rate of return vector and $\bar{r}_t := (R_{1t}, \cdots, R_{Nt})$, $t = 1, \cdots, T$.

In Young (1998), the author was concerned with the minimum return among all the time periods, i.e.,

$$\min_{t=1,\cdots,T} \bar{r}_t^{\mathrm{T}} w,$$

which has a piecewise linear structure. By maximizing the above minimum return, the *minimax model* is constructed as follows:

$$\begin{aligned} \max_{w} \quad & \min_{t=1,\cdots,T} \bar{r}_t^{\mathrm{T}} w \\ \text{s.t.} \quad & r^{\mathrm{T}} w \geq G \\ & \mathbf{1}^{\mathrm{T}} w = 1, \; w \geq 0, \end{aligned}$$

where the parameter $G \geq 0$ represents the minimum level of rate of return. With an auxiliary variable $M_p := \min_{t=1,\cdots,T} \bar{r}_t^{\mathrm{T}} w$, the minimax model can be equivalently written as the following linear programming problem

$$\begin{aligned} \max_{w, M_p} \quad & M_p \\ \text{s.t.} \quad & \bar{r}_t^{\mathrm{T}} w \geq M_p, \; t = 1, \cdots, T \\ & r^{\mathrm{T}} w \geq G, \; \mathbf{1}^{\mathrm{T}} w = 1, \; w \geq 0. \end{aligned}$$

As the minimax model is risk-averse, the risk measure in this model is $-M_p$ rather than M_p. For simplicity, we call it the *minimax risk measure* hereafter.

4.3.1 l_p-Sparse Minimax Models

By adding the l_p $(0 < p \leq 1)$ norm in the minimax portfolio model, we obtain the following *l_p-sparse minimax model*

$$\begin{aligned} \min_{w,M_p} \quad & -M_p + \tau\|w\|_p^p \\ \text{s.t.} \quad & \bar{r}_t^{\mathrm{T}} w \geq M_p,\ t = 1, \cdots, T \\ & r^{\mathrm{T}} w \geq G,\ \mathbf{1}^{\mathrm{T}} w = 1,\ w \geq \alpha, \end{aligned} \tag{4.2}$$

where α is the lower bound of the portfolio, and $\tau \geq 0$ is a tunable regularization parameter.

In (4.2), a more general bound constraint $w \geq \alpha$ is adopted instead of $w \geq 0$. When $p = 1$, we restrict ourselves to the case of $\alpha < 0$, which means the limited short selling is allowed in the investment; otherwise, (4.2) will reduce to the original minimax model in that $\|w\|_1 = 1$ if $\alpha \geq 0$. Furthermore, different from the mean-variance model, the minimax model requires a finite lower bound of the portfolio, as the model may generate an infinite optimal value when $\alpha = -\infty$. But for a finite α, the feasible region is bounded, then the corresponding optimal value is finite. Hence, we set $\alpha > -\infty$ to guarantee the validness of (4.2).

A descent property is also satisfied for (4.2) (see Proposition 4.3.1). We omit the proof since it is similar to that of Proposition 4.2.1.

Proposition 4.3.1. *Let w_τ be an optimal portfolio produced by (4.2) corresponding to a specific τ and w_τ^- be the componentwise negative part of w_τ.*

(1). For any $0 < p \leq 1$, one has that

$$(\tau_1 - \tau_2)\left(\|w_{\tau_2}\|_p^p - \|w_{\tau_1}\|_p^p\right) \geq 0.$$

(2). Furthermore, when $p = 1$, one has that

$$(\tau_1 - \tau_2)\left(\left\|w_{\tau_2}^-\right\|_1 - \left\|w_{\tau_1}^-\right\|_1\right) \geq 0.$$

Proposition 4.3.1(1) indicates that a larger τ leads to a higher level of sparsity (see Figures 4.3(a) and 4.4(a)). And Proposition 4.3.1(2) demonstrates that, with a smaller τ, the optimal portfolio produced by (4.2) includes more short selling stocks; furthermore, a nonnegative portfolio may be obtained when τ is sufficiently large (see Figures 4.3(b) and 4.4(b)).

4.3.2 l_1-Sparse Minimax Sharpe Ratio Model

In this subsection, we extend the classical Sharpe ratio by a modified minimax risk measure, propose an l_1-sparse minimax Sharpe ratio model using the extended minimax Sharpe ratio, and design a global algorithm for the proposed model. To this end, we first introduce the background of the Sharpe ratio.

The classical *Sharpe ratio* of a portfolio w is defined by the following fractional function

$$\frac{r^{\mathrm{T}}w - r_f}{\sqrt{w^{\mathrm{T}}\Sigma w}} ,$$

where r_f is the risk-free rate, and the numerator represents the expected excess return of w. In the denominator, the volatility or standard deviation $\sqrt{w^{\mathrm{T}}\Sigma w}$ is used to measure the risk of w.

It is worth noting that a nonnegative expected excess return needs to be assumed when using the Sharpe ratio. Bacon (2008) pointed out that the negative expected excess return makes the Sharpe ratio difficult to interpret. In fact, a larger Sharpe ratio corresponds to a higher rank of the portfolio. However, a negative expected excess return generates a negative Sharpe ratio, thus results in a perverse ranking. Specifically, a higher not lower level of risk is preferable when the expected excess return is negative.

l_1-sparse Minimax Sharpe Ratio Model

Inspired by the quotient structure of the Sharpe ratio, we are interested in constructing a generalized Sharpe ratio by using the minimax risk measure $-M_p$ to replace the volatility $\sqrt{w^{\mathrm{T}}\Sigma w}$ in the denominator. As discussed above, the nonnegativity assumption of the expected excess return is still necessary for the new Sharpe ratio. For a similar reason, the denominator of the new Sharpe ratio needs to be positive. But unfortunately, $-M_p$ is negative in general. Therefore, it is infeasible to straightly adopt the original minimax risk measure. To overcome this difficulty, we revise the minimax risk measure by guaranteeing the denominator positive with a pre-selected parameter λ. In particular, we consider $\lambda - M_p$ instead of $-M_p$, where λ is a constant such that $\lambda - M_p > 0$ for all the concerned portfolios. The validness of this revision is evident. The duty of the risk measure is to rank the risk of portfolios; in this sense, $\lambda - M_p$ is consistent with $-M_p$.

In what follows, we study a generalized l_1*-sparse minimax Sharpe ratio model* based on the revised minimax risk measure, i.e.,

$$\begin{array}{rl} \min\limits_{w, M_p} & -\dfrac{r^{\mathrm{T}}w - r_f}{\lambda - M_p} + \tau\|w\|_1 \\ \text{s.t.} & \bar{r}_t^{\mathrm{T}} w \geq M_p,\ t = 1, \cdots, T \\ & \mathbf{1}^{\mathrm{T}} w = 1,\ w \geq \alpha, \end{array} \tag{4.3}$$

where r_f is the risk-free rate, and λ is a parameter such that $\lambda - M_p > 0$ over the constraint set. Moreover, we assume that $r^{\mathrm{T}}w - r_f \geq 0$ is satisfied for all the feasible portfolios. The relations in Proposition 4.3.1 still hold for (4.3).

The choice of λ influences the final result. Figure 4.1 states a return-risk space, where point $A(\lambda_1)$ represents the excess rate of return of portfolio A and its corresponding minimax risk revised by λ_1, and the other points are defined similarly. In this space, the value of the Sharpe ratio is expressed by the slope of the point. As we can see from Figure 4.1, with the selection of λ_1, portfolio A is better than portfolio B since the gradient of $A(\lambda_1)$ is steeper. While in the situation with λ_2, the result is the opposite. However, we need to emphasize that, as long as λ is selected such that $\lambda - M_p > 0$, the revised minimax risk measure remains the essence of the original one, thus the corresponding result is always reasonable.

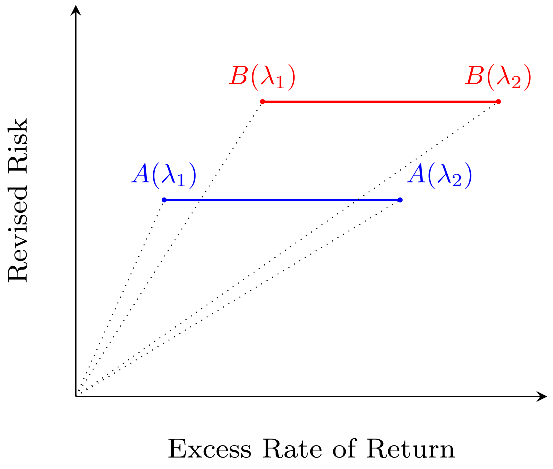


Figure 4.1: Choice of λ

The parametric algorithm

Here, we investigate a global algorithm for the l_1-sparse minimax Sharpe ratio model (4.3). To this end, we study a generalization of the parametric algorithm proposed by Konno & Kuno (1990), which is to minimize the sum of a differentiable convex function and a linear fractional function subject to linear inequality constraints. Specifically, we aim to develop a parametric algorithm for the following generalized linear fractional programming problem

$$\begin{aligned} \min_x \quad & g(x) - \frac{c_1^{\mathrm{T}} x + c_{10}}{c_2^{\mathrm{T}} x + c_{20}} \\ \text{s.t.} \quad & x \in X =: \{x \in \mathbb{R}^n : A_1 x \geq b_1,\ A_2 x = b_2\}, \end{aligned} \tag{4.4}$$

where $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a convex but not necessarily differentiable function. Moreover, c_1, $c_2 \in \mathbb{R}^n$, c_{10}, $c_{20} \in \mathbb{R}$, $A_1 \in \mathbb{R}^{p\times n}$, $A_2 \in \mathbb{R}^{q\times n}$, $b_1 \in \mathbb{R}^p$, and $b_2 \in \mathbb{R}^q$ are given coefficients. We assume that the feasible region X is nonempty and bounded and that

$$c_1^{\mathrm{T}} + c_{10} \geq 0 \quad \text{and} \quad c_2^{\mathrm{T}} + c_{20} > 0 \qquad \text{for all} \quad x \in X.$$

We first consider a master problem of (4.4)

$$\begin{aligned} \min_{x,\xi} \quad & g(x) - 2\xi\sqrt{c_1^{\mathrm{T}}x + c_{10}} + \xi^2(c_2^{\mathrm{T}}x + c_{20}) \\ \text{s.t.} \quad & x \in X,\ \xi \geq 0, \end{aligned} \tag{4.5}$$

where $\xi \in \mathbb{R}$ is an auxiliary single variable. Proposition 4.3.2 (Konno & Kuno (1990), Theorem 4.3) is a fundamental property for designing the parametric algorithm, where the relation between (4.4) and (4.5) is presented.

Proposition 4.3.2. *Let (x^*, ξ^*) be an optimal solution of (4.5). Then x^* is an optimal solution of (4.4).*

Let $P(\xi)$ be the optimal value of the following convex optimization problem, an associated problem of (4.5) with a specific $\xi \geq 0$,

$$\begin{aligned} \min_{x} \quad & g(x) - 2\xi\sqrt{c_1^{\mathrm{T}}x + c_{10}} + \xi^2(c_2^{\mathrm{T}}x + c_{20}) \\ \text{s.t.} \quad & x \in X. \end{aligned} \tag{4.6}$$

According to Proposition 4.3.2, an optimal solution of (4.4) can be obtained by solving (4.6) with $\xi = \xi^*$, where ξ^* is a nonnegative number such that $\mathrm{P}(\xi^*) \leq \mathrm{P}(\xi)$ for all $\xi \geq 0$. Therefore, the main idea of the parametric algorithm is to solve (4.6) over all $\xi \geq 0$, then the solution corresponding to the smallest optimal value of (4.6) produces an optimal solution of (4.4). However, it is hard to solve (4.6) when $\xi \to +\infty$. As a matter of fact, for some classes of parametric programming problems, all the problems with sufficiently large ξ, say $\xi \geq \xi_{max}$, share the same optimal solutions, say x^*_{max}. Following this property, we can focus on $[0, \xi_{max}]$ instead of $[0, +\infty)$. Therefore, we need to prove that this property is satisfied for (4.5).

Proposition 4.3.3. *There exists $\xi_{max} \in \mathbb{R}$ such that x^*_{max} is an optimal solution of* (4.6) *for any $\xi \geq \xi_{max}$, where $x^*_{max} \in S^* := \arg\min\{g(x) : x \in S^*_1\}$,*

$$S^*_1 := \arg\max\{c_1^{\mathrm{T}} x : x \in S^*_2\}, \quad \textit{and} \quad S^*_2 := \arg\min\{c_2^{\mathrm{T}} x : x \in X\}.$$

If S^*_2 is a singleton, then $x^*_{max} = \arg\min\{c_2^{\mathrm{T}} x : x \in X\}$.

Proof. We prove the following equivalent statement: there exists $\xi_{max} \in \mathbb{R}$ such that $F(x, \xi) \geq F(x^*_{max}, \xi)$ for all $\xi \geq \xi_{max}$ and $x \in X \backslash S^*$. Let $\gamma_1(x) := c_2^{\mathrm{T}} x - c_2^{\mathrm{T}} x^*_{max}$, $\gamma_2(x) := \sqrt{c_1^{\mathrm{T}} x + c_{10}} - \sqrt{c_1^{\mathrm{T}} x^*_{max} + c_{10}}$, and $\gamma_3(x) := g(x) - g(x^*_{max})$. Clearly,

$$F(x, \xi) - F(x^*_{max}, \xi) = \gamma_1(x)\xi^2 - 2\gamma_2(x)\xi + \gamma_3(x).$$

We consider three cases: $x \in S^*_1 \backslash S^*$, $x \in S^*_2 \backslash S^*_1$, and $x \in X \backslash S^*_2$.

Case (i). Let $x \in S^*_1 \backslash S^*$. We have that $\gamma_1(x) = \gamma_2(x) = 0$ and $\gamma_3(x) > 0$. Then $F(x, \xi) \geq F(x^*_{max}, \xi)$ holds for all $\xi_{max} \in \mathbb{R}$.

Case (ii). Let $x \in S^*_2 \backslash S^*_1$. We have that $\gamma_1(x) = 0$ and $\gamma_2(x) < 0$. Then ξ_{max} satisfying $F(x, \xi) \geq F(x^*_{max}, \xi)$ exists in that $\gamma_3(x)$ is bounded.

Case (iii). Let $x \in X \backslash S^*_2$. We have that $\gamma_1(x) > 0$ and

$$\frac{F(x, \xi) - F(x^*_{max}, \xi)}{\gamma_1(x)} = \left(\xi - \frac{\gamma_2(x)}{\gamma_1(x)}\right)^2 + \frac{\gamma_3(x)}{\gamma_1(x)} - \left(\frac{\gamma_2(x)}{\gamma_1(x)}\right)^2.$$

Then it follows from the boundedness of $\frac{\gamma_2(x)}{\gamma_1(x)}$ and $\frac{\gamma_3(x)}{\gamma_1(x)}$ that there exists ξ_{max} such that $F(x, \xi) \geq F(x^*_{max}, \xi)$ for all $x \in X \backslash S^*_2$ and $\xi \geq \xi_{max}$.

From all cases, the statement is true. □

Although Konno & Kuno (1990) provided an approach to finding x^*_{max} for a generalized linear multiplicative programming problem, we need to point out that their criteria may fail when the associated problem has more than one solution.

Now, we present how to locate ξ_{max} using x^*_{max}. When $\xi \geq \xi_{max}$, as x^*_{max} is an optimal solution of (4.6) and the linearity constraint qualification (LCQ) is satisfied, there exist multipliers $\lambda := (\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^p$ and $\mu := (\mu_1, \ldots, \mu_q) \in \mathbb{R}^q$ such that

$$\begin{cases} 0 \in \partial g(x^*_{max}) - \dfrac{c_1}{\sqrt{c_1^{\mathrm{T}} x^*_{max} + c_{10}}}\xi + c_2\xi^2 - A_1^{\mathrm{T}}\lambda - A_2^{\mathrm{T}}\mu \\ \lambda_i(A_1 x^*_{max} - b_1)_i = 0, \;\; \lambda_i \geq 0, \;\; i = 1, \ldots, p, \end{cases}$$

where $(A_1 x^*_{max} - b_1)_i$ represents the i^{th} element of the vector $A_1 x^*_{max} - b_1$. Let $\bar{\lambda}$ and $\bar{A}_1$ be the sub-vector and sub-matrix of λ and A_1 corresponding to active constraints in $A_1 x^*_{max} \geq b_1$ (i.e., the inequality is strict). Then, the above system becomes

$$A_0^{\mathrm{T}}\nu \in \partial g(x^*_{max}) - \frac{c_1}{\sqrt{c_1^{\mathrm{T}} x^*_{max} + c_{10}}}\xi + c_2\xi^2,$$

where $A_0 = \begin{bmatrix} \bar{A}_1 \\ A_2 \end{bmatrix}$ and $\nu = \begin{bmatrix} \bar{\lambda} \\ \mu \end{bmatrix}$. As a result, ξ_{max} can be estimated through the following system

$$\begin{cases} A_0^{\mathrm{T}}\nu \in \partial g(x^*_{max}) - \dfrac{c_1}{\sqrt{c_1^{\mathrm{T}} x^*_{max} + c_{10}}}\xi + c_2\xi^2 \\ \bar{\lambda} \geq 0, \end{cases} \tag{4.7}$$

Solving (4.7) can be expensive and complicated. But fortunately, when A_0 is a matrix of full-rank square, the process can be much simplified. With this assumption, (4.7) can be rearranged as

$$\begin{cases} \nu \in Q_0 - q_1\xi + q_2\xi^2 \\ \bar{\lambda} \geq 0, \end{cases}$$

where $Q_0 := \{(A_0^{\mathrm{T}})^{-1}\} \times \partial g(x_{max}^*)$, $q_1 := \frac{(A_0^{\mathrm{T}})^{-1}c_1}{\sqrt{c_1^{\mathrm{T}}x_{max}^*+c_{10}}}$, and $q_2 := (A_0^{\mathrm{T}})^{-1}c_2$. Remarkably, Q_0 can be viewed as a vector, where the elements are sets rather than numbers. Let $Q_0^{(\lambda)}, q_1^{(\lambda)}$, and $q_2^{(\lambda)}$ be the sub-vectors of Q_0, q_1, and q_2 corresponding to λ. Then, the existence of λ implies that

$$q_0^{(\lambda)} - q_1^{(\lambda)}\xi + q_2^{(\lambda)}\xi^2 \geq 0 \quad \text{and} \quad q_0^{(\lambda)} = \max\{Q_0^{(\lambda)}\}, \tag{4.8}$$

where $q_0^{(\lambda)} = \max\{Q_0^{(\lambda)}\}$ means that $q_0^{(\lambda)}$ is a vector consisting of the maximums of all sets in $Q_0^{(\lambda)}$. Noting that the existence of λ is absolute, we have that $q_2^{(\lambda)} \geq 0$. That is, the solution of (4.8) can be derived explicitly.

As discussed above, we conclude the parametric algorithm in two steps.

Step 1. *Find x_{max}^* by solving problems in Proposition 4.3.3 (see S^*, S_1^*, and S_2^*) and ξ_{max} by solving (4.7) or (4.8).*

Step 2. *If $\xi_{max} \leq 0$, then x_{max}^* is the global solution of (4.4); otherwise, solve (4.6) over $\xi \in [0, \xi_{max}]$, then the solution x^* corresponding to the smallest optimal value is a global solution of (4.4).*

Discretization is a practical method to search the minimum optimal value of (4.6) over $[0, \xi_{max}]$. More precisely, we first divide the interval $[0, \xi_{max}]$ into many subdivisions and then solve (4.6) with ξ at every breakpoint. If all the subdivisions are narrow enough, the resulting solution would be sufficiently close to a global solution of (4.4).

We end the part by some remarks for the parametric algorithm.

Remark 4.3.4.

(1). Konno & Kuno (1990) adopted the generalized inverse matrix to solve (4.7). However, this method possibly leads to a wrong ξ_{max} when A_0 is not a matrix of full-rank square.

(2). The assumption that A_0 is a matrix of full-rank square is not very strict. For instance, it is satisfied if S_2^* in Proposition 4.3.3 is a singleton and the Linear independence constraint qualification (LICQ) holds at x_{max}^*.

(3). Every step of locating ξ_{max} is sufficient and necessary; thus, ξ_{max} derived by the above process is exact for the problem.

4.4 Numerical Experiments

This section evaluates the numerical performance of the l_p-sparse $(0 < p \leq 1)$ minimax model (4.2) and l_1-sparse minimax Sharpe ratio model (4.3) using the weekly historical data of 1200 stocks from the Hang Seng, Shanghai Composite, and NASDAQ stock markets (400 stocks from each) from 1 January 2005 to 31 December 2019. The rate of return R_{jt} is calculated by the formula $R_{jt} = \frac{p_{j,t+1} - p_{jt}}{p_{jt}}$, where p_{jt} represents the price of stock j in week t. The expected rate of return is estimated by the average rate of return, i.e., $r = (\frac{1}{T}\sum_{t=1}^{T} R_{1t}, \cdots, \frac{1}{T}\sum_{t=1}^{T} R_{Nt})$.

For (4.2), the l_1-sparse minimax model is a convex optimization problem, while the model becomes nonconvex when $0 < p < 1$. Therefore, we treat them separately in the testing. As the $l_{1/2}$ norm has been shown to be the best in the literature (see Chartrand (2007) and Hu et al. (2017)), we use the $l_{1/2}$-sparse minimax model to typify the case $0 < p < 1$. Benchmarks are the equal-weighted rule and the l_1-sparse and $l_{1/2}$-sparse mean-variance models (4.1). The equal-weighted rule, taking $w = (1/N, \cdots, 1/N)$, has been illustrated to outperform many portfolio selection models (see DeMiguel et al. (2007)).

We now present the computation for all the concerned models one by one. All the experiments are completed in MATLAB R2020a and Windows 10 on a 64-bit PC with an i7-4790 CPU and 32GB RAM.

For the l_1-sparse minimax model, we first translate it into a smooth formulation and then adopt the optimization toolbox (function ‘linprog’) in Matlab to solve the equivalent problem. With an auxiliary variable $u := |w|$, the model (4.2) with $p = 1$ can be equivalently written as a linear programming problem

$$\begin{aligned} \min_{u,w,M_p} \quad & -M_p + \tau \cdot \mathbf{1}^{\mathrm{T}} u \\ \text{s.t.} \quad & -u \leq w \leq u \\ & \bar{r}_t^{\mathrm{T}} w \geq M_p, \ t = 1, \cdots, T \\ & r^{\mathrm{T}} w \geq G, \ \mathbf{1}^{\mathrm{T}} w = 1, \ w \geq \alpha. \end{aligned}$$

This is a parametric linear programming problem with respect to τ. A similar transformation can be applied to the l_1-sparse mean-variance model. According to Best & Grauer (1991) and Berkelaar et al. (1997), there exists a finite set of breakpoints $0 \leq \tau_1 < \cdots < \tau_K < +\infty$ such that the optimal solution set keeps unchanged on any (open) interval between two successive breakpoints, which is also observed in the figures of Experiment 2.

The computation of the l_1-sparse mean-variance model is conducted by the CVX toolbox (see Grant et al. (2020)). Furthermore, the iterative reweighted minimization method provided by Lu (2014b) is utilized to solve the $l_{1/2}$-sparse minimax and mean-variance models. The l_1-sparse minimax Sharpe ratio model is computed by the parametric algorithm in Subsection 4.3.2 and Algorithm 2 in Chapter 3, respectively. To this end, we need to transform the model (4.3) into a generalized bilinear framework. By virtue of auxiliary variables $u := |w|$ and $z := \frac{r^{\mathrm{T}} w - r_f}{\lambda - M_p}$, we obtain an

Table 4.1: Computational time for different sparse models

l_1-MM	$l_{1/2}$-MM	l_1-MV	$l_{1/2}$-MV	l_1-SR(p)	l_1-SR(a)
0.56s	2.90s	24.35s	85.91s	225.33s	117.21s

equivalent problem of (4.3), i.e.,

$$\begin{aligned}
\min_{u,w,z,M_p} \quad & -z+\tau\cdot\mathbf{1}^{\mathrm{T}}u \\
\text{s.t.} \quad & -u\leq w\leq u \\
& z\cdot M_p+r^{\mathrm{T}}w-\lambda z=r_f \\
& \bar{r}_t^{\mathrm{T}}w\geq M_p,\ t=1,\cdots,T \\
& \mathbf{1}^{\mathrm{T}}w=1,\ w\geq\alpha,
\end{aligned}$$

which is bilinear with respect to (u,w,z) and M_p.

Preliminarily, we test the computational time for all models with a specific τ (see Table 4.1). In Table 4.1, 'l_1-SR(p)' and 'l_1-SR(a)' represent the result of the parametric algorithm and Algorithm 2. For reliability, the selected value of τ for each model corresponds to 12–14 active stocks.

Here, we begin to examine the sparse minimax models using data from the real market. Initially, we obverse the out-of-sample rate of return of the l_1-sparse minimax model with different τ in the first experiment. Then, in Experiment 2, we investigate the effect of the regularization parameter τ on the sparsity and short selling of the optimal portfolio. Finally, using the descent tendency observed in the second experiment, we conduct the last experiment, comparing the performance of sparse minimax models and sparse mean-variance models at the same level of sparsity.

Experiment 1. Return with Different Regularization Parameters

In this experiment, we test the out-of-sample rate of return of the l_1-sparse minimax model (4.2) with different τ and compare it with the equal-weighted rule. We set the required rate of return G to be the average rate of return of all the stocks. Each time period is taken as one week, and the number of periods is set as $T=11$. The lower bound α is fixed at -0.2; that is, the short selling for each stock is limited to under 20%.

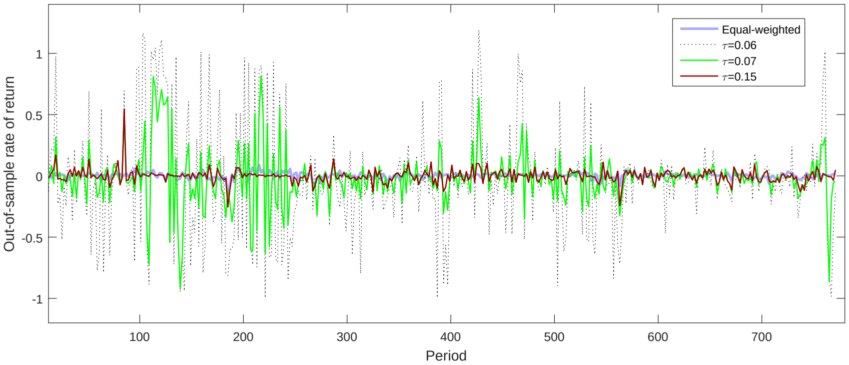


Figure 4.2: Rate of return with different τ

We apply the rolling window process. In particular, for the current time period, data from the previous 11 time periods (since $T = 11$) is used to determine the coefficients of the model (i.e., $\bar{r}_t$, r, and G). Then, an optimal portfolio can be obtained by solving the l_1-sparse minimax model (4.2) with these $\bar{r}_t$, r, and G. The out-of-sample rate of return is calculated by the obtained optimal portfolio and the rate of return of the current time period. For example, the out-of-sample rate of return in period 12 is estimated using the associated coefficients produced by data from period 1–11 and the rate of return of all stocks in period 12, and the same procedure is repeated in the sequential periods.

Figure 4.2 plots the out-of-sample rates of return of the equal-weighted rule and the l_1-sparse minimax model with $\tau = 0.06$, 0.07, and 0.15, respectively. There are many similarities between the four curves in terms of the trend. Specifically, those rates of return increase or decrease at the same time in most periods. For the three l_1-sparse minimax models, a small τ leads to evident fluctuations while a larger one produces fewer variations. This tendency is partly due to the effect of τ on the level of short selling (see Proposition 4.3.1(2)). And as claimed in Luenberger (2013), short selling is considered quite risky, thus causes fluctuations.

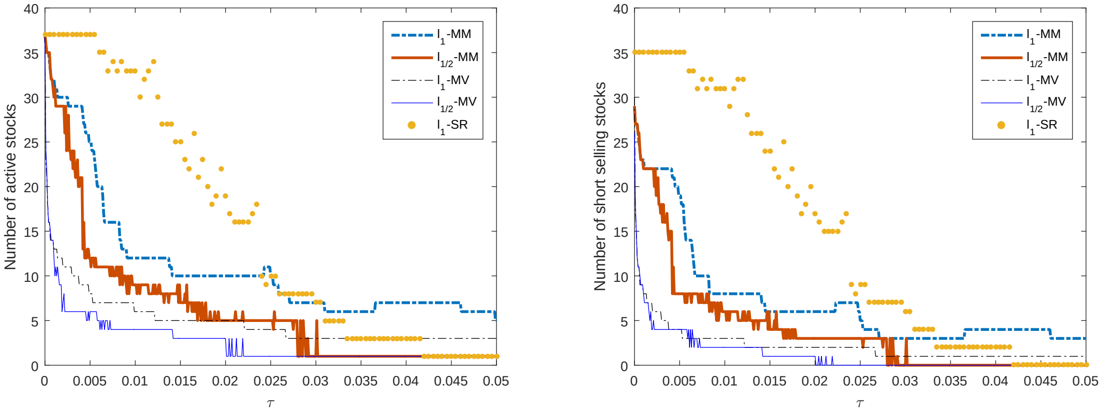
Experiment 2. Descent Trend on Sparsity and Short Selling

In Experiment 2, we select 37 blue chips from the Hang Seng market to illustrate the variation tendency of sparsity and short selling of the l_1-sparse minimax model (4.2), $l_{1/2}$-sparse minimax model (4.2), and l_1-sparse minimax Sharpe ratio model (4.3), together with two benchmark models: the l_1-sparse and $l_{1/2}$-sparse mean-variance models (4.1). For this purpose, we obverse their sparsity and short selling with τ going through 0–0.05. We keep the basic setting the same as in Experiment 1. To see the influence of the bound parameter α, we also experiment $\alpha = -0.5$.

Figures 4.3(a) and 4.4(a) show that, for all the five models, the number of active (nonzero) stocks in the optimal portfolio (i.e., the level of sparsity) decreases as the value of τ increases, which can be explained by Proposition 4.3.1(1) on theoretical considerations. The monotonicity of the l_1-sparse minimax Sharpe ratio model is less exact compared with that of the other two l_1-sparse models. As τ goes up, the curve representing the $l_{1/2}$-sparse minimax (resp. mean-variance) model reduces more dramatically than that of the l_1-sparse minimax (resp. mean-variance) model. This descent property is quite practical and critical for the following experiment. More precisely, we can target optimal portfolios in which the number of active stocks is required within a specific range by taking τ over a smaller interval.

Figures 4.3(b) and 4.4(b) demonstrate a similar descent trend in terms of the short selling, which coincides with Proposition 4.3.1(2) for the l_1-sparse models. When $0 < p < 1$, the relation $\|x\|_p^p = \|x^+\|_p^p + \|x^-\|_p^p$ partly explains the consistent tendency between the sparsity and short selling. From Figures 4.3 and 4.4, a more sparse portfolio, at the same time, is a portfolio with a smaller number of negative-weighted stocks, and a quite sparse portfolio may not include any short selling. Furthermore, we find that the selection of α does not influence the descent tendency, and an extremely sparse portfolio is again attained with $\alpha = 0.5$. The only difference is that the extremely sparse portfolio is obtained at a smaller τ for the model with a larger α. It is also noteworthy that, in all the figures, graphs are piecewise constant due to the parametric construction of sparse models, as analyzed in Subsection 4.3.1.

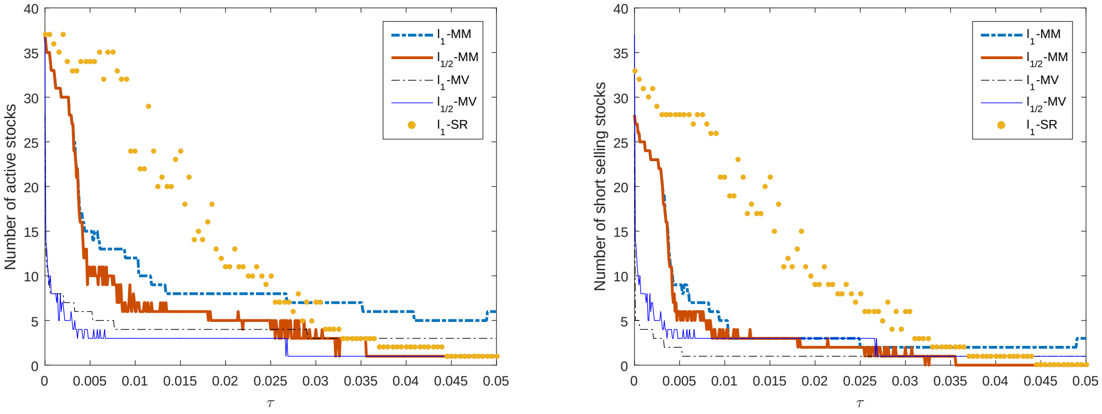
In fact, the 1200-stock case shares the same descent trend. But for the problem including 1200 stocks, three sparse minimax models vary in a relatively large range, say 0–0.07, while the sparse mean-variance models vary in a quite narrow range, say 0–10^{-8}. The difference is attributed to the different orders of magnitude of the



(a). Number of active stocks

(b). Number of short selling stocks

Figure 4.3: Performance with $\alpha = -0.2$



(a). Number of active stocks

(b). Number of short selling stocks

Figure 4.4: Performance with $\alpha = -0.5$

optimal objective functions. For the 37-stock case, the orders of the minimax and mean-variance models are both -1. However, the respective orders are 1 and -11 for the 1200-stock problem.

Experiment 3. Performances under Fixed Level of Sparsity

In the last experiment, the rolling window process mentioned in Experiment 1 is repeated over five observation periods to compare different sparse models with 1200 stocks. The desired rate of return G is retaken to be the maximal rate of return of all stocks. The level of short selling ($\alpha = -0.2$) and the number of periods ($T = 11$) remain unchanged, while the out-of-sample observation period is reset as 11 weeks. For example, data from period 688–698 are used to determine the optimal weights, and then we use them to compute the out-of-sample performance (e.g., the rate of return or Sharpe ratio) of period 699–709. As the same regularization parameter τ in different sparse models generally corresponds to different levels of sparsity, there is little comparability between different sparse models with the same τ. Therefore, a more practical method is to compare them at the same level of sparsity. The comparison in what follows is completed under this consideration.

The example tests the out-of-sample performance of the l_1-sparse and $l_{1/2}$-sparse minimax models and l_1-sparse minimax Sharpe ratio model under five levels of sparsity (see the last five columns in Table 4.2) from periods 699–709 to 703–713. The l_1-sparse and $l_{1/2}$-sparse mean-variance models are considered as benchmarks. The level of sparsity, say 11–20, means the number of active stocks is between 11 and 20, which can be achieved by adjusting the value of τ (see Experiment 2). However, in general, more than one portfolio falls into the target level of sparsity. For this situation, the smallest risk of these portfolios and its corresponding rate of return, Sharpe ratio, and number of short selling stocks are considered. If the portfolio with the minimum risk is still not unique, we select one with the maximal rate of return.

Table 4.2: Performance of different sparse models

(a). l_1-sparse minimax model

l_1-MM	Equal-weighted				11–20				21–30				31–40				41–50				51–60			
	R	Risk$_{MM}$	SR$_{MM}$	S	R	Risk$_{MM}$	SR$_{MM}$	S	R	Risk$_{MM}$	SR$_{MM}$	S	R	Risk$_{MM}$	SR$_{MM}$	S	R	Risk$_{MM}$	SR$_{MM}$	S	R	Risk$_{MM}$	SR$_{MM}$	S
Period 699–709	0.007	0.032	0.218	0	0.012	0.112	0.109	10	0.033	0.248	0.133	21	0.034	0.443	0.076	31	0.046	0.511	0.090	40	0.072	0.616	0.118	50
Period 700–710	0.009	0.032	0.270	0	0.019	0.271	0.070	10	0.042	0.383	0.111	20	0.033	0.453	0.073	29	0.058	0.502	0.116	39	0.077	0.668	0.116	50
Period 701–711	0.008	0.032	0.237	0	0.097	0.325	0.298	12	0.150	0.424	0.353	21	0.169	0.612	0.277	31	0.238	0.713	0.334	40	0.286	0.949	0.301	51
Period 702–712	0.006	0.032	0.199	0	-0.004	0.307	-0.014	11	0.003	0.633	0.005	22	-0.002	0.974	-0.002	31	0.010	1.086	0.010	42	0.024	1.237	0.019	51
Period 703–713	0.004	0.029	0.150	0	0.022	0.179	0.123	9	0.099	0.252	0.393	21	0.127	0.387	0.329	31	0.155	0.525	0.296	40	0.196	0.677	0.289	51

(b). $l_{1/2}$-sparse minimax model

$l_{1/2}$-MM	Equal-weighted				11–20				21–30				31–40				41–50				51–60			
	R	Risk$_{MM}$	SR$_{MM}$	S	R	Risk$_{MM}$	SR$_{MM}$	S	R	Risk$_{MM}$	SR$_{MM}$	S	R	Risk$_{MM}$	SR$_{MM}$	S	R	Risk$_{MM}$	SR$_{MM}$	S	R	Risk$_{MM}$	SR$_{MM}$	S
Period 699–709	0.007	0.032	0.218	0	0.030	0.235	0.128	13	0.049	0.332	0.148	23	0.057	0.424	0.134	36	0.065	0.598	0.108	42	0.072	0.687	0.105	48
Period 700–710	0.009	0.032	0.270	0	0.050	0.476	0.105	14	0.008	0.625	0.013	24	0.033	0.762	0.043	37	0.033	0.835	0.039	42	0.044	0.967	0.046	53
Period 701–711	0.008	0.032	0.237	0	0.067	0.337	0.198	11	0.142	0.506	0.280	24	0.201	0.736	0.273	34	0.278	0.942	0.295	44	0.299	1.131	0.264	51
Period 702–712	0.006	0.032	0.199	0	0.051	3.182	0.016	11	0.758	6.499	0.117	16	-0.009	0.418	-0.022	31	-0.010	1.259	-0.008	44	0.016	1.817	0.009	50
Period 703–713	0.004	0.029	0.150	0	0.037	0.098	0.384	5	0.104	0.381	0.273	23	0.155	0.411	0.377	32	0.194	0.603	0.321	43	0.235	0.710	0.331	53

(c). l_1-sparse mean-variance model

l_1-MV	Equal-weighted				11–20				21–30				31–40				41–50				51–60			
	R	Risk$_{MV}$	SR$_{MV}$	S	R	Risk$_{MV}$	SR$_{MV}$	S	R	Risk$_{MV}$	SR$_{MV}$	S	R	Risk$_{MV}$	SR$_{MV}$	S	R	Risk$_{MV}$	SR$_{MV}$	S	R	Risk$_{MV}$	SR$_{MV}$	S
Period 699–709	0.007	0.000	31.384	0	0.014	0.004	3.177	7	0.013	0.004	3.156	11	0.014	0.004	3.259	14	0.014	0.004	3.393	18	0.015	0.004	3.653	27
Period 700–710	0.009	0.000	32.612	0	0.012	0.009	1.442	6	0.011	0.008	1.362	7	0.010	0.008	1.288	16	0.009	0.007	1.193	24	0.007	0.007	1.040	35
Period 701–711	0.008	0.000	25.308	0	0.021	0.027	0.809	5	0.025	0.026	0.963	7	0.032	0.027	1.215	15	0.029	0.023	1.249	23	0.027	0.022	1.226	27
Period 702–712	0.006	0.000	21.137	0	0.011	0.008	1.337	7	0.012	0.008	1.404	9	0.013	0.008	1.531	12	0.010	0.008	1.288	14	0.019	0.010	1.909	23
Period 703–713	0.004	0.000	18.594	0	0.006	0.006	0.995	6	0.005	0.005	0.916	9	0.004	0.006	0.737	10	0.005	0.006	0.843	14	0.005	0.006	0.843	21

(d). $l_{1/2}$-sparse mean-variance model

$l_{1/2}$-MV	Equal-weighted				11–20				21–30				31–40				41–50				51–60			
	R	Risk$_{MV}$	SR$_{MV}$	S	R	Risk$_{MV}$	SR$_{MV}$	S	R	Risk$_{MV}$	SR$_{MV}$	S	R	Risk$_{MV}$	SR$_{MV}$	S	R	Risk$_{MV}$	SR$_{MV}$	S	R	Risk$_{MV}$	SR$_{MV}$	S
Period 699–709	0.007	0.000	31.384	0	0.033	0.016	2.060	8	0.015	0.006	2.592	12	0.009	0.007	1.341	19	0.030	0.015	1.940	24	0.008	0.007	1.231	36
Period 700–710	0.009	0.000	32.612	0	0.006	0.009	0.687	6	0.040	0.034	1.197	17	0.006	0.009	0.692	20	0.006	0.009	0.718	25	0.007	0.009	0.769	32
Period 701–711	0.008	0.000	25.308	0	0.055	0.043	1.263	7	0.054	0.043	1.247	13	0.053	0.045	1.162	18	0.052	0.045	1.145	30	0.052	0.045	1.155	32
Period 702–712	0.006	0.000	21.137	0	0.024	0.012	2.078	5	0.021	0.011	1.851	12	0.021	0.011	1.867	14	0.021	0.011	1.865	19	0.024	0.012	2.081	23
Period 703–713	0.004	0.000	18.594	0	0.018	0.006	2.911	4	0.015	0.006	2.484	10	0.015	0.006	2.441	15	0.008	0.006	1.534	17	0.013	0.006	2.424	26

(e). l_1-sparse minimax Sharpe ratio model (parametric method)

l_1-SR(p)	Equal-weighted				11–20				21–30				31–40				41–50				51–60			
	R	Risk$_{MM}$	SR$_{MM}$	S	R	Risk$_{MM}$	SR$_{MM}$	S	R	Risk$_{MM}$	SR$_{MM}$	S	R	Risk$_{MM}$	SR$_{MM}$	S	R	Risk$_{MM}$	SR$_{MM}$	S	R	Risk$_{MM}$	SR$_{MM}$	S
Period 699–709	0.007	0.032	0.218	0	0.026	0.274	0.095	11	0.069	0.533	0.130	20	0.090	0.804	0.112	31	0.113	1.042	0.108	42	0.130	1.199	0.108	50
Period 700–710	0.009	0.032	0.270	0	0.023	0.431	0.054	11	0.028	0.563	0.050	20	0.063	0.893	0.071	32	-0.084	1.148	-0.073	42	0.097	1.312	0.074	50
Period 701–711	0.008	0.032	0.237	0	0.028	0.244	0.113	11	0.032	0.517	0.062	20	0.064	0.771	0.084	30	0.107	1.184	0.091	40	0.151	1.462	0.103	52
Period 702–712	0.006	0.032	0.199	0	0.025	0.295	0.083	10	0.046	0.556	0.082	21	0.065	0.791	0.082	30	0.082	1.056	0.077	40	0.082	1.316	0.063	50
Period 703–713	0.004	0.029	0.150	0	-0.001	0.306	-0.005	10	0.006	0.552	0.011	21	0.000	0.820	0.000	30	0.000	1.168	0.000	40	0.008	1.563	0.005	52

(f). l_1-sparse minimax Sharpe ratio model (ADMM)

l_1-SR(a)	Equal-weighted				11–20				21–30				31–40				41–50				51–60			
	R	Risk_{MM}	SR_{MM}	S	R	Risk_{MM}	SR_{MM}	S	R	Risk_{MM}	SR_{MM}	S	R	Risk_{MM}	SR_{MM}	S	R	Risk_{MM}	SR_{MM}	S	R	Risk_{MM}	SR_{MM}	S
Period 699–709	0.007	0.032	0.218	0	0.019	0.171	0.111	9	0.037	0.278	0.133	16	0.052	0.495	0.105	25	-0.004	0.019	-0.210	32	0.094	0.777	0.121	39
Period 700–710	0.009	0.032	0.270	0	-0.011	0.407	-0.027	7	0.028	0.903	0.031	15	0.041	0.872	0.047	22	0.078	1.472	0.053	29	0.106	3.655	0.029	36
Period 701–711	0.008	0.032	0.237	0	0.038	0.826	0.046	5	0.062	2.067	0.030	11	0.081	1.841	0.044	24	0.065	1.275	0.051	29	0.137	3.341	0.041	41
Period 702–712	0.006	0.032	0.199	0	0.009	0.087	0.103	6	0.021	0.107	0.197	13	0.046	0.164	0.281	29	0.097	0.348	0.279	31	0.091	0.316	0.288	35
Period 703–713	0.004	0.029	0.150	0	0.025	0.087	0.288	7	-0.014	0.438	-0.032	15	0.065	0.259	0.251	20	-0.007	0.104	-0.067	27	0.013	0.047	0.274	31

In Tables 4.2(a) to 4.2(f), R, $\text{Risk}_{\texttt{MV}}/\text{Risk}_{\texttt{MV}}$, $\text{SR}_{\texttt{MM}}/\text{SR}_{\texttt{MV}}$, and S represent the out-of-sample rate of return, out-of-sample risk, out-of-sample Sharpe ratio, and the number of short selling stocks. The result of equal-weighted rule is also presented for reference. The equal-weighted rule performs best in terms of the Sharpe ratio due to its extremely low level of risk. On the contrary, the rates of return of five sparse models are more favorable than those of the equal-weighted strategy. From Tables 4.2(a), 4.2(b), 4.2(e), and 4.2(f), we observe that, for all the sparse minimax models, a more sparse optimal portfolio tends to have a lower rate of return and a lower level of risk. However, changes in the rate of return and risk are not so significant for the l_1-sparse and $l_{1/2}$-sparse mean-variance models.

Next, we observe the l_1-sparse and $l_{1/2}$-sparse minimax models. When the level of sparsity is extremely high, particularly with $11 - 20$ active stocks, the $l_{1/2}$-sparse minimax model outperforms the l_1-sparse minimax rule, both in the aspect of the rate of return and Sharpe ratio. For the less sparse optimal portfolios, particularly with 41–50 or 51–60 active stocks, they perform closely. Namely, the $l_{1/2}$-sparse minimax model would be a desirable choice for investors who seek extremely sparse portfolios, while the l_1-sparse minimax model is more beneficial to those who prefer relatively less sparse portfolios due to its computational simplicity (see Table 4.1). For the l_1-sparse and $l_{1/2}$-sparse mean-variance models, we do not observe any superiority of the $l_{1/2}$-sparse formulation. As a whole, their out-of-sample performances appear to be commensurate for all levels of sparsity. Then, we compare two methods for

Table 4.3: Performance with different p

Period 701–711	11–20				21–30				31–40				41–50				51–60			
	R	Risk_{MM}	SR_{MM}	S	R	Risk_{MM}	SR_{MM}	S	R	Risk_{MM}	SR_{MM}	S	R	Risk_{MM}	SR_{MM}	S	R	Risk_{MM}	SR_{MM}	S
$l_{1/3}$-MM	0.065	0.422	0.154	13	0.163	0.453	0.361	22	0.150	0.683	0.220	31	0.233	0.837	0.279	42	0.277	0.984	0.281	51
$l_{1/2}$-MM	0.084	0.322	0.261	12	0.113	0.555	0.204	22	0.165	0.704	0.235	33	0.215	0.799	0.269	41	0.277	0.983	0.282	51
$l_{2/3}$-MM	0.077	0.455	0.170	14	0.117	0.528	0.222	22	0.183	0.682	0.269	33	0.213	0.789	0.270	41	0.269	1.031	0.261	52

Table 4.4: Performance with different α

Period 702–712	$\alpha = -0.02$				$\alpha = -0.05$				$\alpha = -0.2$				$\alpha = -0.5$			
(level 11–20)	R	Risk	SR	S	R	Risk	SR	S	R	Risk	SR	S	R	Risk	SR	S
l_1-MM	0.003	0.130	0.020	14	0.010	0.130	0.074	11	-0.004	0.307	-0.014	11	0.103	0.505	0.205	13
$l_{1/3}$-MM	-0.006	0.157	-0.040	16	0.004	0.155	0.029	0	0.060	0.292	0.206	13	0.104	0.495	0.209	12
$l_{1/2}$-MM	-0.000	0.142	-0.003	11	0.000	0.181	0.002	13	0.051	3.182	0.016	11	0.104	0.502	0.208	12
$l_{2/3}$-MM	0.001	0.130	0.006	13	-0.000	0.204	-0.002	13	0.031	0.408	0.075	13	0.102	0.498	0.204	12
l_1-MV	0.013	0.009	1.401	14	0.020	0.012	1.623	11	0.011	0.008	1.337	7	0.019	0.010	1.971	5
$l_{1/2}$-MV	-0.007	0.013	-0.561	16	0.019	0.011	1.727	13	0.024	0.012	2.078	5	0.035	0.015	2.305	5
l_1-SR(p)	0.004	0.141	0.028	14	0.010	0.204	0.049	17	0.025	0.295	0.083	10	0.079	0.795	0.100	12
l_1-SR(a)	0.006	0.094	0.064	12	0.013	0.169	0.077	11	0.009	0.087	0.103	6	0.035	0.337	0.104	7

solving the l_1-sparse minimax Sharpe ratio model (4.3). The result in Table 4.2(e) seems better and more stable than that in Table 4.2(f), but the advantages are not evident. And we also need to notice that the latter is more efficient in terms of the computational time (see Table 4.1).

Remarkably, the Sharpe ratios of the minimax model and mean-variance model are not comparable in that they are calculated by their respective risk, which are not comparable. Therefore, the only performance measure for comparing the sparse minimax models and sparse mean-variance models is the out-of-sample rate of return. Tables 4.2(a) and 4.2(c) (resp. Tables 4.2(b) and 4.2(d)) illustrate that the optimal portfolios of the l_1-sparse (resp. $l_{1/2}$-sparse) minimax model tend to achieve higher rates of return than those of the l_1-sparse (resp. $l_{1/2}$-sparse) mean-variance model. From Tables 4.2(a) and 4.2(e) (or 4.2(f)), the l_1-sparse minimax model and l_1-sparse minimax Sharpe ratio model perform similarly. Although the computation of the l_1-sparse minimax model is easier, the l_1-sparse minimax Sharpe ratio model still would be a good choice for investors who do not have the desired return in advance.

We also conduct the above experiment with different p (i.e., $p = 1/3$ and $2/3$) and α (i.e., $\alpha = -0.02$, -0.05, and -0.5), respectively. The results of periods 701–711 and 702–712 are listed in Tables 4.3 and 4.4 as the representatives. For the three l_p-sparse minimax models, it seems that they have similar results when the number of active stocks is between $41 - 50$ or $51 - 60$. For other levels of sparsity, the performances are distinct, and it is hard to say which is better. Concerning the effect of α, with a larger level of sparsity, we observe that higher rates of return and Sharpe ratios are obtained for all the models and that the risks of three sparse minimax models increase. However, the risks of two sparse mean-variance models are stable with different α.

In conclusion, all sparse minimax models are efficient for promoting the sparsity of the optimal portfolios. Moreover, with the level of sparsity fixed, the numerical performance of all the sparse minimax models is satisfactory compared to that of the sparse mean-variance models. The $l_{1/2}$-sparse minimax model is advantageous when the investor requires an extremely sparse portfolio, while the l_1-sparse minimax model is favorable for investment with a less strict requirement for sparsity. For the l_1-sparse minimax Sharpe ratio model, it is preferred when the desired return is not given in advance. Furthermore, optimal portfolios including fewer stocks of the l_p-sparse ($0 < p \leq 1$) minimax models tend to have lower rates of return and lower levels of risk. However, for the l_p-sparse mean-variance models, the corresponding changes are not so significant.

Chapter 5

Conclusions

In this thesis, we proposed an extrapolated inexact quasisubgradient method and a penalty extrapolated ADMM method, provided their convergence results, and illustrated their efficiency by numerical testing. Furthermore, we also constructed sparse minimax portfolio selection models by using the l_p $(0 < p \leq 1)$ norm and then explored their properties theoretically and numerically.

We investigated the quasisubgradient method with extrapolation in respect of the convergence in objective values, iteration complexity and rate of convergence. When the diminishing stepsize is decaying as a power function and the extrapolation rule is decreasing not less than a power function, the method provides a sublinear rate $\mathcal{O}\left(\tau^{k^s}\right)$ (for some $0 < s < 1$ and $0 < \tau < 1$) of convergence to the optimal solution set or to a ball of the optimal solution set, which is faster than $\mathcal{O}\left(1/k^h\right)$ for each $h > 0$. This is new in the literature. In a similar way, we proposed a primal-dual quasisubgradient method with extrapolation. Convergence results of both methods with extrapolation are consistent with those without extrapolation when the extrapolation step is appropriately selected. The numerical results indicate that the number of iteration required for obtaining an approximate optimal solution when using the quasisubgradient method with extrapolation is much less than that when using the corresponding quasisubgradient methods without extrapolation.

For the penalty extrapolated ADMM algorithm, we established the subsequential convergence, iteration complexity, and global convergence of the inner iterations and obtained the convergence to the stationary point of the outer iterations. The iteration complexity of the inner algorithm is $\mathcal{O}(1/k)$, which is the same as some nonextrapolated convex and nonconvex ADMM methods in He & Yuan (2012, 2015), and Hong et al. (2016) and better than $\mathcal{O}(1/\sqrt{k})$ for extrapolated convex ADMM methods in Chen et al. (2015). From the numerical experiment of a nonconvex QCQP problem, we observed that our proposed method has an advantage of running time compared to the SDR method, which is a popular method for solving QCQP problems.

We also considered the l_p-sparse $(0 < p \leq 1)$ minimax portfolio model and the l_1-sparse minimax Sharpe ratio model and developed a parametric algorithm for solving the second model. A descent property of the l_p norm of the optimal portfolio with respect to the regularization parameter was obtained for the proposed models. In numerical experiments, we found that all the sparse minimax models are efficient for promoting the sparsity of the optimal portfolios. The $l_{1/2}$-sparse minimax model is advantageous when the investor requires an extremely sparse portfolio. However, the l_1-sparse minimax model is favorable for the investment with a less strict requirement for sparsity. The l_1-sparse minimax Sharpe ratio model is preferred when the desired return is not given in advance.

References

Alecsa, C. D., László, S. C., & Viorel, A. (2019). A gradient-type algorithm with backward inertial steps associated to a nonconvex minimization problem. *Numerical Algorithms*, 1–28.

Al-Khayyal, F. A. (1992). Generalized bilinear programming: Part I. models, applications and linear programming relaxation. *European Journal of Operational Research*, *60*(3), 306–314.

Almutairi, H., & Elhedhli, S. (2009). A new lagrangean approach to the pooling problem. *Journal of Global Optimization*, *45*(2), 237.

Alvarez, F. (2004). Weak convergence of a relaxed and inertial hybrid projection-proximal point algorithm for maximal monotone operators in hilbert space. *SIAM Journal on Optimization*, *14*(3), 773–782.

Alves, M. M., Eckstein, J., Geremia, M., & Melo, J. G. (2020). Relative-error inertial-relaxed inexact versions of douglas-rachford and ADMM splitting algorithms. *Computational Optimization and Applications*, *75*(2), 389–422.

Anstreicher, K. M. (2012). On convex relaxations for quadratically constrained quadratic programming. *Mathematical Programming*, *136*(2), 233–251.

Attouch, H., Bolte, J., Redont, P., & Soubeyran, A. (2010). Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, *35*(2), 438–457.

Attouch, H., & Cabot, A. (2020). Convergence of a relaxed inertial proximal algorithm for maximally monotone operators. *Mathematical Programming*, *184*(1), 243–287.

Attouch, H., & Teboulle, M. (2004). Regularized lotka-volterra dynamical system as continuous proximal-like method in optimization. *Journal of Optimization Theory and Applications*, *121*(3), 541–570.

Audet, C., Brimberg, J., Hansen, P., Digabel, S. L., & Mladenović, N. (2004). Pooling problem: Alternate formulations and solution methods. *Management Science*, *50*(6), 761–776.

Auslender, A., & Teboulle, M. (2004). Interior gradient and epsilon-subgradient descent methods for constrained convex minimization. *Mathematics of Operations Research*, *29*(1), 1–26.

Auslender, A., & Teboulle, M. (2006). *Asymptotic cones and functions in optimization and variational inequalities*. Springer Science & Business Media.

Aussel, D., Corvellec, J.-N., & Lassonde, M. (1995). Mean value property and subdifferential criteria for lower semicontinuous functions. *Transactions of the American Mathematical Society*, *347*(10), 4147–4161.

Avriel, M., Diewert, W. E., Schaible, S., & Zang, I. (1988). *Generalized concavity.* New York: Plenum Press.

Bacon, C. R. (2008). *Practical portfolio performance measurement and attribution* (2nd ed., Vol. 546). New York: John Wiley & Sons.

Bazaraa, M. S., Goode, J., & Nashed, M. Z. (1974). On the cones of tangents with applications to mathematical programming. *Journal of Optimization Theory and Applications*, *13*(4), 389–426.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, *2*(1), 183–202.

Benninga, S., & Czaczkes, B. (2014). *Financial modeling* (4th ed.). London: MIT press.

Benson, H. P. (2007). A simplicial branch and bound duality-bounds algorithm for the linear sum-of-ratios problem. *European Journal of Operational Research*, *182*(2), 597–611.

Ben-Tal, A., Eiger, G., & Gershovitz, V. (1994). Global minimization by reducing the duality gap. *Mathematical Programming*, *63*(1), 193–212.

Berkelaar, A. B., Roos, K., & Terlaky, T. (1997). The optimal set and optimal partition approach to linear and quadratic programming. In *Advances in sensitivity analysis and parametic programming* (pp. 159–202). Springer.

Bertsekas, D. P., Nedìc, A., & Ozdaglar, A. (2003). *Convex analysis and optimization.* Athena Scientific.

Best, M. J., & Grauer, R. R. (1991). On the sensitivity of mean-variance-efficient portfolios to changes in asset means: Some analytical and computational results. *The Review of Financial Studies*, *4*(2), 315–342.

Bloemhof-Ruwaard, J. M., & Hendrix, E. M. (1996). Generalized bilinear programming: An application in farm management. *European Journal of Operational Research*, *90*(1), 102–114.

Bolte, J., Sabach, S., & Teboulle, M. (2014). Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, *146*(1), 459–494.

Boţ, R. I., Csetnek, E. R., & Hendrich, C. (2015). Inertial douglas-rachford splitting for monotone inclusion problems. *Applied Mathematics and Computation*, *256*, 472–487.

Bradley, S. P., & Frey Jr, S. C. (1974). Fractional programming with homogeneous functions. *Operations Research*, *22*(2), 350–357.

Brodie, J., Daubechies, I., De Mol, C., Giannone, D., & Loris, I. (2009). Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences*, *106*(30), 12267–12272.

Burke, J. V., & Ferris, M. C. (1993). Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, *31*(5), 1340–1359.

Cai, X., Teo, K.-L., Yang, X., & Zhou, X. (2000). Portfolio optimization under a minimax rule. *Management Science*, *46*(7), 957–972.

Chambolle, A., & Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, *40*(1), 120–145.

Chambolle, A., & Pock, T. (2016). On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, *159*(1), 253–287.

Chang, T.-H., Hong, M., & Wang, X. (2014). Multi-agent distributed optimization via inexact consensus ADMM. *IEEE Transactions on Signal Processing*, *63*(2), 482–497.

Chao, M., Zhang, Y., & Jian, J. (2020). An inertial proximal alternating direction method of multipliers for nonconvex optimization. *International Journal of Computer Mathematics*, 1–19.

Chartrand, R. (2007). Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, *14*(10), 707–710.

Chen, C., Chan, R. H., Ma, S., & Yang, J. (2015). Inertial proximal ADMM for linearly constrained separable convex optimization. *SIAM Journal on Imaging Sciences*, *8*(4), 2239–2267.

Chen, C., He, B., Ye, Y., & Yuan, X. (2016). The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, *155*(1-2), 57–79.

Chen, C., Li, X., Tolman, C., Wang, S., & Ye, Y. (2013). Sparse portfolio selection via quasi-norm regularization. *arXiv: 1312.6350*.

Crouzeix, J.-P., Martínez-Legaz, J.-E., & Volle, M. (1998). *Generalized convexity, generalized monotonicity: Recent results*. Kluwer Academic Publishers.

Cruz, J. B., & Pérez, L. L. (2010). Convergence of a projected gradient method variant for quasiconvex objectives. *Nonlinear Analysis: Theory, Methods & Applications*, *73*(9), 2917–2922.

Cruz, J. B., Pérez, L. L., & Melo, J. (2011). Convergence of the projected gradient method for quasiconvex multiobjective optimization. *Nonlinear Analysis: Theory, Methods & Applications*, *74*(16), 5268–5273.

Dai, Z., & Wen, F. (2018). A generalized approach to sparse and stable portfolio optimization problem. *Journal of Industrial & Management Optimization*, *14*(4), 1651.

Daniilidis, A., Hadjisavvas, N., & Martínez-Legaz, J.-E. (2001). An appropriate subdifferential for quasiconvex functions. *SIAM Journal on Optimization*, *12*(2), 407–420.

Das, S. R., Markowitz, H. M., Scheid, J., & Statman, M. (2011). Portfolios for investors who want to reach their goals while staying on the mean-variance efficient frontier. *The Journal of Wealth Management*, *14*(2), 25–31.

Daubechies, I., Defrise, M., & De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, *57*(11), 1413–1457.

DeMiguel, V., Garlappi, L., Nogales, F. J., & Uppal, R. (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, *55*(5), 798–812.

DeMiguel, V., Garlappi, L., & Uppal, R. (2007). Optimal versus naive diversification: How inefficient is the $1/n$ portfolio strategy? *Review of Financial Studies*, *22*(5), 1915–1953.

Deng, W., Lai, M.-J., Peng, Z., & Yin, W. (2017). Parallel multi-block ADMM with $o(1/k)$ convergence. *Journal of Scientific Computing*, *71*(2), 712–736.

dos Santos Gromicho, J. A. (1998). *Quasiconvex optimization and location theory* (Vol. 9). Dordrecht: Kluwer Academic Publishers.

Eckstein, J., & Bertsekas, D. P. (1992). On the douglas—rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, *55*(1), 293–318.

Elton, E. J., Gruber, M. J., Brown, S. J., & Goetzmann, W. N. (2014). *Modern portfolio theory and investment analysis* (9th ed.). New York.

Erbeyoğlu, G., & Bilge, Ü. (2016). Pso-based and sa-based metaheuristics for bilinear programming problems: An application to the pooling problem. *Journal of Heuristics*, *22*(2), 147–179.

Ermol'ev, Y. M. (1966). Methods of solution of nonlinear extremal problems. *Cybernetics*, *2*(4), 1–14.

Fastrich, B., Paterlini, S., & Winker, P. (2015). Constructing optimal sparse portfolios using regularization methods. *Computational Management Science*, *12*(3), 417–434.

Floudas, C. A., & Aggarwal, A. (1990). A decomposition strategy for global optimum search in the pooling problem. *ORSA Journal on Computing*, *2*(3), 225–235.

Fomin, S., Fulton, W., Li, C.-K., & Poon, Y.-T. (2005). Eigenvalues, singular values, and littlewood-richardson coefficients. *American Journal of Mathematics*, *127*(1), 101–127.

Foulds, L. R., Haugland, D., & Jørnsten, K. (1992). A bilinear approach to the pooling problem. *Optimization*, *24*(1-2), 165–180.

Fukuda, M., & Kojima, M. (2001). Branch-and-cut algorithms for the bilinear matrix inequality eigenvalue problem. *Computational Optimization and Applications*, *19*(1), 79–105.

Fukushima, M. (1992). Application of the alternating direction method of multipliers to separable convex programming problems. *Computational Optimization and Applications*, *1*(1), 93–111.

Gabay, D., & Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, *2*(1), 17–40.

Goudou, X., & Munier, J. (2009). The gradient and heavy ball with friction dynamical systems: The quasiconvex case. *Mathematical Programming*, *116*(1), 173–191.

Grant, M., Boyd, S., & Ye, Y. (2020). *CVX: Matlab software for disciplined convex programming (version 2.2).* http://cvxr.com/cvx/.

Greenberg, H. J., & Pierskalla, W. P. (1973). Quasi-conjugate functions and surrogate duality. *Cahiers du Centre d'étude de Recherche Operationelle*, *15*, 437–448.

Hadjisavvas, N., Komlósi, S., & Schaible, S. S. (2005). *Handbook of generalized convexity and generalized monotonicity* (Vol. 76). New York: Springer-Verlag.

Hajinezhad, D., Chang, T.-H., Wang, X., Shi, Q., & Hong, M. (2016). Nonnegative matrix factorization using ADMM: Algorithm and convergence analysis. In *2016 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4742–4746).

Hajinezhad, D., & Shi, Q. (2018). Alternating direction method of multipliers for a class of nonconvex bilinear optimization: convergence analysis and applications. *Journal of Global Optimization*, *70*(1), 261–288.

Harjunkoski, I., Pörn, R., Westerlund, T., & Skrifvars, H. (1997). Different strategies for solving bilinear integer non-linear programming problems with convex transformations. *Computers & Chemical Engineering*, *21*, S487–S492.

Harjunkoski, I., Westerlund, T., Pörn, R., & Skrifvars, H. (1998). Different transformations for solving non-convex trim-loss problems by minlp. *European Journal of Operational Research*, *105*(3), 594–603.

He, B., Liao, L., Han, D., & Yang, H. (2002). A new inexact alternating directions method for monotone variational inequalities. *Mathematical Programming*, *92*(1), 103–118.

He, B., Liao, L., & Qian, M. (2006). Alternating projection based prediction-correction methods for structured variational inequalities. *Journal of Computational Mathematics*, 693–710.

He, B., Tao, M., & Yuan, X. (2012). Alternating direction method with gaussian back substitution for separable convex programming. *SIAM Journal on Optimization*, *22*(2), 313–340.

He, B., & Yuan, X. (2012). On the O($1/n$) convergence rate of the douglas–rachford alternating direction method. *SIAM Journal on Numerical Analysis*, *50*(2), 700–709.

He, B., & Yuan, X. (2015). On non-ergodic convergence rate of douglas–rachford alternating direction method of multipliers. *Numerische Mathematik*, *130*(3), 567–577.

Hishinuma, K., & Iiduka, H. (2020). Fixed point quasiconvex subgradient method. *European Journal of Operational Research*, *282*(2), 428–437.

Hong, M., Luo, Z., & Razaviyayn, M. (2016). Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, *26*(1), 337–364.

Hong, M., & Luo, Z.-Q. (2017). On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming*, *162*(1-2), 165–199.

Hu, Y., Li, C., Meng, K., Qin, J., & Yang, X. (2017). Group sparse optimization via $l_{p,q}$ regularization. *The Journal of Machine Learning Research*, *18*(1), 960–1011.

Hu, Y., Li, J., & Yu, C. K.-W. (2020). Convergence rates of subgradient methods for quasi-convex optimization problems. *Computational Optimization and Applications*, *77*, 183-–212.

Hu, Y., Yang, X., & Sim, C.-K. (2015). Inexact subgradient methods for quasi-convex optimization problems. *European Journal of Operational Research*, *240*(2), 315–327.

Hu, Y., Yang, X., & Yu, C. K.-W. (2016). Subgradient methods for saddle point problems of quasiconvex optimization. *Pure and Applied Functional Analysis*, *2*(1), 83–97.

Huang, X., & Yang, X. (2003). A unified augmented lagrangian approach to duality and exact penalization. *Mathematics of Operations Research*, *28*(3), 533–552.

Jagannathan, R., & Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, *58*(4), 1651–1683.

Jensen, M. C. (1968). The performance of mutual funds in the period 1945-1964. *The Journal of Finance*, *23*(2), 389–416.

Jia, Z., Wu, Z., & Dong, X. (2019). An inexact proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth optimization problems. *Journal of Inequalities and Applications*, *125*, 1–16.

Jiang, B., Lin, T., Ma, S., & Zhang, S. (2019). Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. *Computational Optimization and Applications*, *72*(1), 115–157.

Jiao, H., & Liu, S. (2015). A practicable branch and bound algorithm for sum of linear ratios problem. *European Journal of Operational Research*, *243*(3), 723–730.

Johnstone, P. R., & Moulin, P. (2017). Local and global convergence of a general inertial proximal splitting scheme for minimizing composite functions. *Computational Optimization and Applications*, *67*(2), 259–292.

Kiwiel, K. C. (2001). Convergence and efficiency of subgradient methods for quasiconvex minimization. *Mathematical Programming*, *90*(1), 1–25.

Kiwiel, K. C., & Murty, K. (1996). Convergence of the steepest descent method for minimizing quasiconvex functions. *Journal of Optimization Theory and Applications*, *89*(1), 221–226.

Konno, H. (1971). *Bilinear programming: Part II. application of bilinear programming.* (Tech. Rep.). STANFORD UNIV CALIF DEPT OF OPERATIONS RESEARCH.

Konno, H., & Kuno, T. (1990). Generalized linear multiplicative and fractional programming. *Annals of Operations Research*, *25*(1), 147–161.

Konno, H., & Kuno, T. (1992). Linear multiplicative programming. *Mathematical Programming*, *56*(1), 51–64.

Konno, H., & Yamazaki, H. (1991). Mean-absolute deviation portfolio optimization model and its applications to tokyo stock market. *Management Science*, *37*(5), 519–531.

Konnov, I. V. (2003). On convergence properties of a subgradient method. *Optimization Methods and Software*, *18*(1), 53–62.

Kruger, A. Y. (2003). On Fréchet subdifferentials. *Journal of Mathematical Sciences*, *116*(3), 3325–3358.

Kuno, T., Konno, H., & Yamamoto, Y. (1992). A parametric successive underestimation method for convex programming problems with an additional convex multiplicative constraint. *Journal of the Operations Research Society of Japan*, *35*(3), 290–299.

Langenberg, N., & Tichatschke, R. (2012). Interior proximal methods for quasiconvex optimization. *Journal of Global Optimization*, *52*(3), 641–661.

Li, G., & Pong, T. K. (2015). Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, *25*(4), 2434–2460.

Liberti, L., & Pantelides, C. C. (2006). An exact reformulation algorithm for large nonconvex nlps involving bilinear terms. *Journal of Global Optimization*, *36*(2), 161–189.

Linderoth, J. (2005). A simplicial branch-and-bound algorithm for solving quadratically constrained quadratic programs. *Mathematical Programming*, *103*(2), 251–282.

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

Lu, Z. (2014b). Iterative reweighted minimization methods for l_p regularized unconstrained nonlinear programming. *Mathematical Programming*, *147*(1), 277–307.

Luenberger, D. G. (2013). *Investment science* (2nd ed.). New York: Oxford University Press.

Luo, Z., Ma, W.-K., So, A. M.-C., Ye, Y., & Zhang, S. (2010). Semidefinite relaxation of quadratic optimization problems. *IEEE Signal Processing Magazine*, *27*(3), 20–34.

Maingé, P.-E. (2009). Asymptotic convergence of an inertial proximal method for unconstrained quasiconvex minimization. *Journal of Global Optimization*, *45*(4), 631–644.

Maingé, P.-E., & Merabet, N. (2010). A new inertial-type hybrid projection-proximal algorithm for monotone inclusions. *Applied Mathematics and Computation*, *215*(9), 3149–3162.

Markowitz, H. M. (1952). Portfolio selection. *The Journal of Finance*(2), 77–91.

Markowitz, H. M., & Van Dijk, E. L. (2003). Single-period mean-variance analysis in a changing world. *Financial Analysts Journal*, *59*(2), 30–44.

Martin, P. G., & McCann, B. B. (1989). *The investor's guide to fidelity funds.* Wiley.

Martínez-Legaz, J., & Sach, P. (1999). A new subdifferential in quasiconvex analysis. *Journal of Convex Analysis*, *6*(1), 1–11.

McCafferty, T. A. (2002). *The market is always right.* New York: McGraw Hill Professional.

Mordukhovich, B. S. (1976). Maximum principle in the problem of time optimal response with nonsmooth constraints. *Journal of Applied Mathematics and Mechanics*, *40*(6), 960–969.

Mordukhovich, B. S. (2006). *Variational analysis and generalized differentiation I: Basic theory* (Vol. 330). New York: Springer Science & Business Media.

Nedić, A., & Bertsekas, D. P. (2010). The effect of deterministic noise in subgradient methods. *Mathematical Programming*, *125*(1), 75–99.

Nesterov, Y. (2013). Gradient methods for minimizing composite functions. *Mathematical Programming*, *140*(1), 125–161.

Nocedal, J., & Wright, S. (2006). *Numerical optimization.* New York: Springer Science & Business Media.

Ochs, P., Chen, Y., Brox, T., & Pock, T. (2014). ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, *7*(2), 1388–1419.

Osman, H., & Demirli, K. (2010). A bilinear goal programming model and a modified Benders decomposition algorithm for supply chain reconfiguration and supplier selection. *International Journal of Production Economics*, *124*(1), 97–105.

Pan, S., & Chen, J.-S. (2007). Entropy-like proximal algorithms based on a second-order homogeneous distance function for quasi-convex programming. *Journal of Global Optimization*, *39*(4), 555–575.

Papalambros, P. Y., & Wilde, D. J. (2000). *Principles of optimal design: Modeling and computation.* Cambridge university press.

Pauca, V. P., Piper, J., & Plemmons, R. J. (2006). Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and its Applications*, *416*(1), 29–47.

Plastria, F. (1985). Lower subdifferentiable functions and their minimization by cutting planes. *Journal of Optimization Theory and Applications*, *46*(1), 37–53.

Pock, T., & Sabach, S. (2016). Inertial proximal alternating linearized minimization (ipalm) for nonconvex and nonsmooth problems. *SIAM Journal on Imaging Sciences*, *9*(4), 1756–1787.

Polyak, B. T. (1967). A general method for solving extremal problems. *(Russian) Doklady Akademii Nauk SSSR*, *174*(1), 33–36.

Polyak, B. T. (1978). Nonlinear programming methods in the presence of noise. *Mathematical Programming*, *14*(1), 87–97.

Polyak, B. T. (1987). *Introduction to optimization.* New York: Optimization Software.

Qi, Y., Zhang, Y., & Ma, S. (2019). Parametrically computing efficient frontiers and reanalyzing efficiency-diversification discrepancies and naive diversification. *INFOR: Information Systems and Operational Research*, *57*(3), 430–453.

Quiroz, E. P., Quispe, E., & Oliveira, P. R. (2008). Steepest descent method with a generalized armijo search for quasiconvex functions on riemannian manifolds. *Journal of Mathematical Analysis and Applications*, *341*(1), 467–477.

Quiroz, E. P., Ramirez, L. M., & Oliveira, P. R. (2015). An inexact proximal method for quasiconvex minimization. *European Journal of Operational Research*, *246*(3), 721–729.

Ramík, J., & Vlach, M. (2012). *Generalized concavity in fuzzy optimization and decision analysis* (Vol. 41). Boston: Kluwer Academic Publishers.

Reklaitis, G. V., Ravindran, A., & Ragsdell, K. M. (1983). *Engineering optimization: Methods and applications*. New York: Wiley.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton University Press.

Rockafellar, R. T. (1985). Extensions of subgradient calculus with applications to optimization. *Nonlinear Analysis: Theory, Methods & Applications*, *9*(7), 665–698.

Rockafellar, R. T., & Wets, R. J.-B. (2009). *Variational analysis* (Vol. 317). Springer Science & Business Media.

Rubinov, A. M., Huang, X., & Yang, X. (2002). The zero duality gap property and lower semicontinuity of the perturbation function. *Mathematics of Operations Research*, *27*(4), 775–791.

Saab, R., Chartrand, R., & Yilmaz, O. (2008). Stable sparse approximations via nonconvex optimization. In *2008 ieee international conference on acoustics, speech and signal processing* (pp. 3885–3888).

Sharpe, W. F. (1966). Mutual fund performance. *The Journal of Business*, *39*(1), 119–138.

Sharpe, W. F. (1967). A linear programming algorithm for mutual fund portfolio selection. *Management Science*, *13*(7), 499–510.

Shen, L., & Pan, S. (2015). A corrected semi-proximal admm for multi-block convex optimization and its application to dnn-sdps. *arXiv:1502.03194*.

Sherali, H. D., & Adams, W. P. (2013). *A reformulation-linearization technique for solving discrete and continuous nonconvex problems* (Vol. 31). Springer Science & Business Media.

Sherali, H. D., & Alameddine, A. (1992). A new reformulation-linearization technique for bilinear programming problems. *Journal of Global Optimization*, *2*(4), 379–410.

Shor, N. Z. (1985). *Minimization methods for non-differentiable functions*. New York: Springer-Verlag.

Sortino, F. A., & Van Der Meer, R. (1991). Downside risk. *Journal of Portfolio Management*, *17*(4), 27.

Souza, S. d. S., Oliveira, P. R., da Cruz Neto, J. X., & Soubeyran, A. (2010). A proximal method with separable bregman distances for quasiconvex minimization over the nonnegative orthant. *European Journal of Operational Research*, *201*(2), 365–376.

Stancu-Minasian, I. M. (2012). *Fractional programming: Theory, methods and applications* (Vol. 409). Springer Science & Business Media.

Studniarski, M., & Ward, D. E. (1999). Weak sharp minima: characterizations and sufficient conditions. *SIAM Journal on Control and Optimization*, *38*(1), 219–236.

Tawarmalani, M., Richard, J.-P. P., & Chung, K. (2010). Strong valid inequalities for orthogonal disjunctions and bilinear covering sets. *Mathematical Programming*, *124*(1), 481–512.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Toh, K.-C., Todd, M. J., & Tütüncü, R. H. (1999). Sdpt3—a matlab software package for semidefinite programming, version 1.3. *Optimization Methods and Software*, *11*(1-4), 545–581.

Treynor, J. (1965). How to rate management of investment funds. *Harvard Business Review*, *43*, 63–75.

Tseng, P. (1993). Dual coordinate ascent methods for non-strictly convex minimization. *Mathematical Programming*, *59*(1), 231–247.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, *109*(3), 475–494.

Van Ngai, H., Luc, D. T., & Théra, M. (2002). Extensions of Fréchet ϵ-subdifferential calculus and applications. *Journal of Mathematical Analysis and Applications*, *268*(1), 266–290.

Vidal, C. J., & Goetschalckx, M. (2001). A global supply chain model with transfer pricing and transportation cost allocation. *European Journal of Operational Research*, *129*(1), 134–158.

Wang, Y., Yin, W., & Zeng, J. (2019). Global convergence of ADMM in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, *78*(1), 29–63.

Wen, B., Chen, X., & Pong, T. K. (2017). Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. *SIAM Journal on Optimization*, *27*(1), 124–145.

Weyl, H. (1912). Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, *71*(4), 441–479.

Woodside-Oriakhi, M., Lucas, C., & Beasley, J. E. (2011). Heuristic algorithms for the cardinality constrained efficient frontier. *European Journal of Operational Research*, *213*(3), 538–550.

Wu, Z., & Li, M. (2019). General inertial proximal gradient method for a class of nonconvex nonsmooth optimization problems. *Computational Optimization and Applications*, *73*(1), 129–158.

Xu, Y., & Yin, W. (2013). A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, *6*(3), 1758–1789.

Xu, Y., Yin, W., Wen, Z., & Zhang, Y. (2012). An alternating direction algorithm for matrix completion with nonnegative factors. *Frontiers of Mathematics in China*, *7*(2), 365–384.

Yang, L., Pong, T. K., & Chen, X. (2017). 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.

Young, M. R. (1998). A minimax portfolio selection rule with linear programming solution. *Management Science*, *44*(5), 673–683.

Yu, C. K.-W., Hu, Y., Yang, X., & Choy, S. K. (2019). Abstract convergence theorem for quasi-convex optimization problems with applications. *Optimization*, *68*(7), 1289–1304.

Zhang, X., Barrio, R., Martínez, M. A., Jiang, H., & Cheng, L. (2019). Bregman proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. *IEEE Access*, *7*, 126515–126529.