FIRST-ORDER METHODS FOR NONSMOOTH NONCONVEX FUNCTIONAL CONSTRAINED OPTIMIZATION WITH OR WITHOUT SLATER POINTS

by
Zhichao Jia

A thesis submitted to The Johns Hopkins University in conformity
with the requirements for the degree of Master of Science in Engineering

Baltimore, Maryland
December, 2022

Abstract

Constrained optimization problems where both the objective and constraints may be nonsmooth and nonconvex arise across many learning and data science settings. In this thesis, we show a simple first-order method finds a feasible, ϵ-stationary point at a convergence rate of $O(\epsilon^{-4})$ without relying on compactness or Constraint Qualification (CQ). When CQ holds, this convergence is measured by approximately satisfying the Karush-Kuhn-Tucker conditions. When CQ fails, we guarantee the attainment of weaker Fritz-John conditions. As an illustrative example, we show our method still stably converges on piecewise quadratic SCAD regularized problems despite frequent violations of constraint qualification. The considered algorithm is similar to those of [1, 2] (whose guarantees both assume compactness and CQ), iteratively taking inexact proximal steps, computed via an inner loop applying a switching subgradient method to a strongly convex constrained subproblem. Our non-Lipschitz analysis of the switching subgradient method analysis appears to be new and may be of independent interest.

Thesis Readers

Dr. Benjamin Grimmer (Primary Advisor)
Assistant Professor
Department of Applied Mathematics and Statistics
Johns Hopkins University

Dedicated to Dr. Grimmer, who leads me to optimization.

Dedicated to my parents, who support my decisions.

Dedicated to myself, who is brave towards his dream and future.

Acknowledgements

I spent one year and a half at Johns Hopkins University significant to my academic career. I gained knowledge, explored my research interests, and met excellent people here. Their kindness, enthusiasm and knowledgeability impressed me a lot.

I am the most grateful to my advisor Ben Grimmer. His kindness, energy, patience and carefulness nourished my academic soul. He exemplifies the kind of mathematician, researcher, teacher and advisor I dream to be. Thank you, Ben. Thank you for teaching me knowledge and inspiring my minds. Thank you for leading me out when I was facing difficulties. Thank you for sharing me interesting and useful stuff in math, academia and life. Thank you for encouraging and helping me all the way.

I also want to express my appreciation to my second advisor, Jim Spall, and my third advisor, Nicolas Loizou. Thank you, Jim, for introducing me to stochastic optimization and teaching me to be academically rigorous. Thank you, Nicolas, for bringing me to large-scale optimization and showing me broad research topics. Thank you both for your kindness and help during my studies.

While I was preparing for this thesis, Dr. Grimmer provided me great help on the technical contents and writing, and supervised my revisions in detail. I would like to especially thank for his help in this whole process.

Throughout my graduate life at Hopkins, there are always supports and help from my old and new friends around me. Thank you, Yuning Ding, for struggling together abroad. Thank you, Lang Lang, for beautiful songs from your guitar, fancy or reliable

ideas and hot pots. Thank you, Ziyi Wei, for enjoyable collaboration in SPSA and having fun. Thank you, Ruoling Fan, for badminton games. Thank you, Zhiyuan Ding, for braised noodles. Thank you Danlin Li, Yi'an Qian, Ducheng Peng, Yichuan Wang, Mengtong Xu accompanied in my research. Thank you Yi Wu, Weihang Gao, Qianyun Zhao accompanied in my courses. Thank you Ruolan Sun, Langze An, Yunpeng Zhao, Yanhan Liao, Haohui Zhu, Jie Chen, Xidan Xu accompanied in my daily life. Thank you all for your partnership, companion and help.

At last, I would like to thank my family, especially my parents. Thank you dad for your unwavering support. Thank you mom for your continuous encouragement. Thank you all for your love from my hometown.

Contents

List of Figures

Chapter 1

Introduction

In this paper, we considered the difficult family of constrained optimization problems where both the objective and constraints may be nonconvex and nonsmooth. Specifically, we study problems of the following form:

$$\begin{aligned} \min_{x \in X} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \qquad i = 1, ..., m. \end{aligned} \tag{1.1}$$

Here, d is dimension of the problem and convex set $X \subseteq \mathbb{R}^d$ is its domain. The objective $f : X \to \mathbb{R}$ and constraints $g_i : X \to \mathbb{R}, i = 1, ..., m$ are assumed to be continuous on X, but need not be convex nor differentiable.

Constrained optimization problems with nonsmooth and nonconvex objective loss functions and constraints are common in modern data science and machine learning fields. For instance, phase retrieval, blind deconvolution and covariance matrix estimation could all be constructed as nonconvex and nonsmooth minimization problems [3–8]. If we further expect sparsity for our solutions, it is effective to introduce a regularizing constraint (e.g., convex choices like ℓ_1-norms or ℓ_2-norms, nonconvex choices like SCAD functions [9, 10] or ℓ_q-norms for $q \in (0, 1)$). The SCAD functions will serve as a running example throughout this work as they are very simple piecewise quadratic functions exhibiting nonsmoothness and nonconvexity, with real usage in several modern sparse optimization problems [11–15]. Other problems like multi-class Neyman-Pearson classification [1, 16, 17], which tries to minimize the loss on one class while controlling the losses on other classes under some values, are also prevalent constrained optimization models inheriting any nonsmoothness and nonconvexities from the loss functions.

Our approach to solving nonsmooth, nonconvex, constrained problems relies on two main ingredients outlined below: (in)exact proximal point methods and Fritz-John/Karush-Kuhn-Tucker stationarity conditions.

(In)exact Proximal Point Methods Several recent works [7, 18–22] have concerned solving nonconvex problems via inexact evaluation of a proximal operator. For settings without functional constraints (i.e., $m = 0$), these methods seek a stationary point of $\min_{x \in X} f(x)$ by iterating

$$x_{k+1} \approx \operatorname{prox}_{\alpha, f}(x_k) := \operatorname*{argmin}_{x \in X} \left\{ f(x) + \frac{1}{2\alpha}\|x - x_k\|^2 \right\} \tag{1.2}$$

with stepsize $\alpha > 0$. By restricting to the family of weakly convex functions (defined in (2.4)), this proximal subproblem is guaranteed to be convex with a unique solution for small enough α. When the proximal map can be evaluated exactly, an $\epsilon > 0$-stationary point (defined in Definitions 2.1 and 2.2) is found within $O(1/\epsilon^2)$ iterations. The inexact methods of [7, 22] show that using cheaper subgradient oracle calls such a point is found within $O(1/\epsilon^4)$ iterations.

We consider the following extension of these ideas to nonconvex inequality constraints by [1, 2] (their ideas and comparisons with our contributions are discussed in depth in Section 1.3). Consider the following proximal subproblem, penalizing the constraints similarly to the objective

$$x_{k+1} \approx \operatorname*{argmin}_{x \in X} \left\{ f(x) + \frac{1}{2\alpha}\|x - x_k\|^2 \mid g_i(x) + \frac{1}{2\alpha}\|x - x_k\|^2 \leq \tau \right\} \tag{1.3}$$

with stepsize $\alpha > 0$ and feasibility tolerance $\tau \geq 0$. Importantly, any feasible solution to this proximal subproblem x_{k+1} has feasibility bounded by $g_i(x_{k+1}) \leq \tau - \frac{1}{2\alpha}\|x_k - x_{k+1}\|_2^2$. Hence a sequence of x_k generated by inexactly evaluating this mapping remains feasible for (1.1) until reaching approximate stationarity (that is, $\|x_k - x_{k+1}\|_2 \geq \sqrt{2\alpha\tau}$ implies $g_i(x_{k+1}) \leq 0$ for each constraint i).

Fritz-John/Karush-Kuhn-Tucker Stationarity Let $\partial f(x)$ denote a generalized subdifferential of a function f and $N_X(x)$ denote the normal cone of X at x, formally defined in Chapter 2. Here we consider two classic measurements of stationarity: Fritz-John (FJ) conditions giving a weaker optimality condition and Karush-Kuhn-Tucker (KKT) conditions giving a stronger condition.

We say that a feasible x^* is a FJ point of (1.1) if there exists nonnegative multipliers $\gamma_0^* \in \mathbb{R}$ and $\gamma^* = (\gamma_1^*, ..., \gamma_m^*)^T \in \mathbb{R}^m$, and subgradients $\zeta_f \in \partial f(x^*)$ and $\zeta_{gi} \in \partial g_i(x^*)$ such that $(\gamma_0^*, \gamma_1^*, ..., \gamma_m^*)$ is a non-zero vector with

$$\begin{aligned} \gamma_i^* g_i(x^*) = 0, \qquad \forall i = 1, ..., m, \\ \gamma_0^* \zeta_f + \sum_{i=1}^{m} \gamma_i^* \zeta_{gi} \in -N_X(x^*). \end{aligned} \tag{1.4}$$

Note requiring $(\gamma_0^*, \gamma_i^*, ..., \gamma_m^*)$ to be a nonzero vector is equivalent to requiring $\gamma_0^* + \sum_{i=1}^m \gamma_i^* = 1$. This condition is necessary for x^* to be a global (or locally) minimizer [23]. However, this condition is known to fail to give meaningful insights into the quality of x^* as a solution whenever $\gamma_0^* = 0$ as (1.4) becomes independent of f [24]. This weakness is remedied by the stronger notion of KKT points, which implicitly require $\gamma_0^* \neq 0$. We say a feasible x^* is a KKT point for the problem (1.1) if there exists nonnegative Lagrange multipliers $\lambda^* \in \mathbb{R}^m$, $\zeta_f \in \partial f(x^*)$ and $\zeta_{gi} \in \partial g_i(x^*)$ such that

$$\begin{aligned} \lambda_i^* g_i(x^*) = 0, \qquad \forall i = 1, ..., m, \\ \zeta_f + \sum_{i=1}^{m} \lambda_i^* \zeta_{gi} \in -N_X(x^*). \end{aligned} \tag{1.5}$$

Note the KKT conditions strengthen FJ, requiring $\gamma_0^* \neq 0$, in particular $\gamma_0^* = 1$. The requirement that $\gamma_0^* \neq 0$ is equivalent to having the Mangasarian-Fromovitz Constraint Qualification (MFCQ) condition hold: Let $A(x) = \{i \mid g_i(x) = 0, i = 1, ..., m\}$. We say MFCQ holds at x if

$$\exists v \in -N_X^*(x) \qquad s.t. \quad \zeta_{gi}^T v < 0 \quad \forall i \in A(x), \forall \zeta_{gi} \in \partial g_i(x). \tag{1.6}$$

Measurements of approximate FJ and KKT stationarity can be vastly different when constraint qualification does not hold. When a strengthened (σ-strict) MFCQ condition (see (2.11)) is not satisfied at the stationarity point our method converging to, the associated Lagrange multipliers may blow up and approximate KKT stationarity may never be attained despite the iterates x_k of (1.3) converging. In contrast, approximately satisfying the FJ conditions can be ensured whenever x_k converges.

1.1 Contribution

We show an inexact proximal method can solve a wide range of nonsmooth, nonconvex constrained optimization problems, producing an approximate stationary point within $O(1/\epsilon^4)$ subgradient evaluations, matching the unconstrained rate. In particular, our proposed method uses a switching subgradient method approximately solving (1.3) to produce each subsequent x_{k+1}, see Algorithm 1. We show this scheme has the following three generally desirable properties missing from prior works [1, 2]:

Always Feasible Iterates By appropriately selecting the algorithmic parameters α and τ, we can ensure feasibility $g_i(x_{k+1}) \leq \tau - \frac{1}{2\alpha}\|x_k - x_{k+1}\|_2^2 \leq 0$ for the original problem (1.1). Maintaining not just approximate but actually feasible iterates is critical, for example, in settings of planning or control where feasibility corresponds to physical limitations or safety concerns [25, 26].

Stationarity with or without Constraint Qualification Ensuring constraint qualification over nonconvex constraints is nontrivial. This is illustrated for a common sparse regularization in Section 1.2 and numerical explored in Chapter 5. In Theorems 3.2 and 3.3 respectively, we show with or without constraint qualification, an inexact proximal point method produces an approximate KKT or FJ point using at most $O(1/\epsilon^4)$ subgradient evaluations.

Convergence Rates without Compactness Our guarantees apply without needing to assume compactness of the domain X, which prior works relied on. This is done by extending the analysis of the switching subgradient method to handle non-Lipschitz objective and constraint functions like those occurring in (1.3). This analysis and resulting subproblem convergence guarantee appear to be new and may be of independent interest.

1.2 Vignette: Failure of MFCQ Assumptions for Sparse Regularized Problems

Nonconvex regularization has become common due to its statistical benefits [27–30]. One of the simplest regularizers is the smoothly clipped absolute deviation (SCAD) function [9, 10], sums up piecewise quadratic clipped absolute deviations in each coordinate

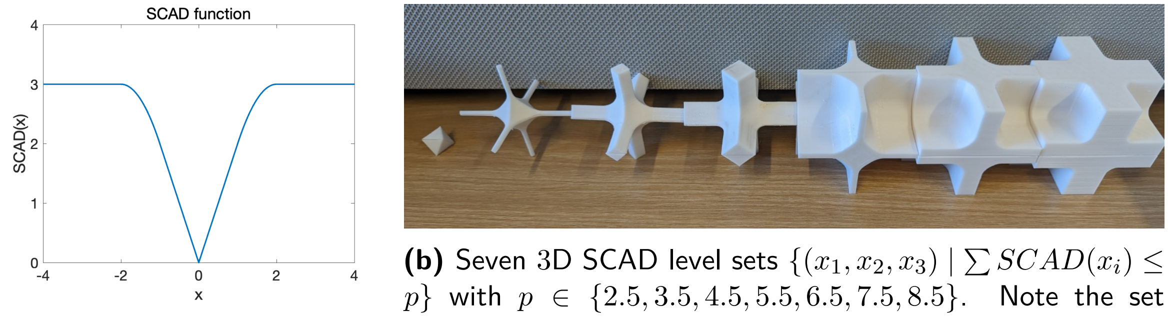
$$SCAD(x_i) = \begin{cases} 2|x_i| & 0 \le |x_i| \le 1, \\ -x_i^2 + 4|x_i| - 1 & 1 < |x_i| \le 2, \\ 3 & |x_i| > 2. \end{cases} \tag{1.7}$$

The constraint $g(x) := \sum_i SCAD(x_i) - p \le 0$ implies that at most $p/3$ entries of x have magnitude larger than two. Here we consider the problem of sparse phase retrieval problems (SPR), see (5.1), which minimizes a piecewise quadratic objective over this piecewise quadratic constraint set.

Note these piecewise quadratic constraints and objective form a simple family of nonsmooth nonconvex problems where we can approximately solve the convex subproblem (1.3). Despite this, two problems (one mild and one severe) prevent the convergence theory of prior works from being applied.

First, prior works do not apply as the set $\{x \mid g(x) \le 0\}$ is not compact for any

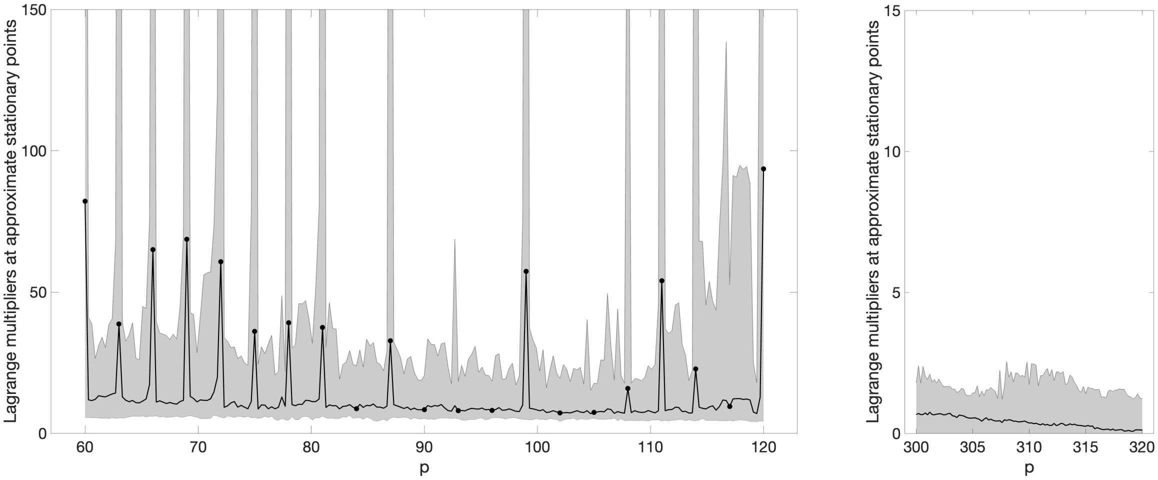
$p \geq 3$. If a bound on the size of a solution is known, then one could add a ball constraint $X = \{x \mid \|x\| \leq D\}$ to enure compactness. Our theory applies without such a modification.



(a) 1D SCAD function

(b) Seven 3D SCAD level sets $\{(x_1, x_2, x_3) \mid \sum SCAD(x_i) \leq p\}$ with $p \in \{2.5, 3.5, 4.5, 5.5, 6.5, 7.5, 8.5\}$. Note the set changes suddenly at $p \in \{3, 6, 9\}$.

Figure 1-1. 1-1a shows the one-dimensional SCAD function in $[-4, 4]$. 1-1b shows the feasible regions of the three-dimensional SCAD function in $[-5, 5]^3$.



(a) Small values of p

(b) Large values of p

Figure 1-2. Lagrange multipliers computed at approximate stationary points reached by iterating (1.3) on randomly generated SPR problems (see Chapter 5 for the exact construction). As p varies from 60 to 120, the black line shows the average multiplier computed after 30 iterations and the gray region shows the range between maximum and minimum values seen. Black dots are placed at each multiple of three, where the strengthened MFCQ condition fails to hold.

More subtly, prior works do not apply here as SPR often fails to have constraint qualification hold as p varies. As a result, none of the prior works' theories can be guaranteed to apply and yield KKT points. Figure 1-2 illustrates the failure of KKT conditions on randomly generated SPR instances. We see that when p is near a multiple of three, the limit point reached by iteratively applying (1.3) may have its associated Lagrange multiplier blow up. For large values of p, we see the multipliers tending to zero, corresponding to unconstrained stationarity.

We could infer from this graph that when p is a multiple of three, the strengthened MFCQ condition breaks. In this case, our theory still guarantees the iteration will find a stationary point that is an approximate FJ point, but which may fail to be an approximate KKT point (having extremely large values of the computed Lagrange multipliers). Despite this possibility, in several of our random problem instances, we see the Lagrange multiplier values stay reasonably despite MFCQ failing to hold everywhere, and so our method identifies the resulting point as an approximate KKT point (in addition to being FJ).

1.3 Related Work

Inexact Proximal Methods The idea of using inexact proximal-point methods to deal with nonsmooth problems is not new to this work. Double-loop algorithms that cost several inner steps to inexactly solve a convex proximal subproblem in each outer iteration have been designed and showed corresponding convergence results. For example, the algorithm proposed in [18] approximating nonconvex proximal points contributed to such an idea, and [19] presented a proximal variant of bundle methods solving nonconvex problems based on the work of [18]. In recent years, [22] developed this idea into unconstrained stochastic settings, where a general class of weakly convex and nonsmooth objective functions were analyzed.

Special Case of (Strongly) Convex Constraints A range of methods from the literature can be applied to inexactly solve the nonsmooth and strongly convex constrained subproblems constructed during the iterations of the inexact proximal method. [31] introduced a level-set method for convex constrained problems, improved by [32] to maintain feasibility. Another type of method transforms constrained minimization problems into minimax problems, and such saddle point problems could be solved by primal-dual type methods as done by [33] and [34]. Switching subgradient methods can also be applied, which have been analyzed in [35] and extended in [1], [36] and [37].

Comparison with Ma, Lin, and Yang [1] We consider a very similar inexact proximal point method with switching subgradient method being the oracle for the subproblems as Ma et al. [1], in which they also find nearly optimal and nearly feasible solutions for the subproblems. In their work, convergence of a stochastic subgradient algorithm was also analyzed. However, under deterministic setting, they can only achieve a nearly feasible and approximate stationary solution for the original optimization problem, while our method ensures feasibility. To attain KKT stationarity, they introduced a uniform Slater's condition as their stronger type of constraint qualification, which is stronger than the strengthened MFCQ condition proposed in this paper. What's more, their constant upper bound for the optimal dual variables and convergence rate for switching subgradient algorithm both rely on the boundedness of the domain X by some constant value, while we do not need such a requirement. Finally, we also show convergence results to FJ points instead of KKT points without constraint qualification.

Comparison with Boob, Deng, and Lan [2] As another comparison, Bood et al. [2] showed the framework of searching for nearly optimal and strictly feasible solutions for the subproblems, and keeping strict feasibility automatically during iterations to finally achieve a feasible approximate stationary solution for the main problem. They also proposed an algorithm which can be used for the subproblems, with corresponding convergence results in various problem settings. However, this algorithm cannot fit their framework of guaranteeing strict feasibility. Furthermore, they considered MFCQ, strong MFCQ and strong feasibility conditions as their stronger types of constraint qualifications, while the strong feasibility condition is stronger than the strengthened MFCQ condition proposed in this paper. For MFCQ and strong MFCQ conditions in their paper, although both of them are weaker than our strengthened MFCQ condition, they need an additional assumption to ensure the existence of an exact stationary solution that the iterated points converge to, which is necessary to show the boundedness of the optimal dual variables. They did not prove that this upper bound could be any constant value, while we attain a constant upper bound directly according to our strengthened MFCQ condition. What's more, they also require the domain to be compact, which is not necessary for us. We further show convergence results to FJ points instead of KKT points without constraint qualification.

Prior Nonconvex Fritz John and KKT-type Guarantees Birgin et al. [38] gave a general method that attains approximate stationarity using first, second, or higher order information. They adopted both scaled KKT points and unscaled KKT points to describe the stationarity, while the former one means the accuracy of KKT conditions satisfied at such points is proportional to the size of the Lagrange multipliers, and it has no influence on the accuracy for the latter one. The scaled KKT points with the linear combination of the gradients of only the constraints being near zero are similar as Fritz John points. Hinder and Ye [39] redefined Fritz John stationarity in a slightly stronger version compared to the natural definition which is more similar to ours, and used IPMs to find approximate Fritz John points under nonconvex constraints. They also introduced their new definitions of unscaled KKT points and termination criteria as comparisons with [38]. Besides, the ideas of adopting scaled KKT stationarity and the corresponding discussions on the size of Lagrange multipliers also occur in [40–43].

Alternative Approaches to Nonconvex Constraints Finally, we note three alternatives to the use of (inexact) proximal methods for nonconvex constrained problems considered here: Classic second-order approaches like sequential quadratic programming techniques [44] can be applied. Cubic regularization approaches [45] and penalized methods [46, 47] can also provide provably convergence guarantees. If the constraints are star convex with respect to a known point, then the radial methods of [48, 49] apply with convergence guarantees towards stationarity while maintaining fully feasible iterates.

Chapter 2

Preliminaries

Throughout the paper, we use the following notations. Let $\|\cdot\|$ denote the l_2-norm. For domain set X, we denote its normal cone at x as $N_X(x)$, and its dual cone as $N_X^*(x)$. The distance from a point x to a set S is denoted as $\mathrm{dist}(x, S) = \min_{s\in S} \|x - s\|$, and the convex hull of any set S is denoted as $\mathrm{co}\{S\}$. For any convex function $h : X \to \mathbb{R} \bigcup \{+\infty\}$, its set of subgradients at $x \in X$ is defined as:

$$\partial h(x) = \{\zeta \in \mathbb{R}^d | h(x') \geq h(x) + \zeta^T (x' - x), \quad \forall x' \in X\}. \tag{2.1}$$

More generally, for any potentially nonconvex function $h : X \to \mathbb{R} \bigcup \{+\infty\}$, its set of Clarke subgradients at x is defined as:

$$\partial h(x) = \mathrm{co}\{\lim_{i\to\infty} \nabla h(x_i) | x_i \to x \text{ and } h(x) \text{ is differentiable at any } x_i \in X\}. \tag{2.2}$$

A function $h(x)$ is μ-strongly convex on X if $h - \frac{\mu}{2}\|\cdot\|^2$ is convex. This is equivalent to having:

$$h(x') \geq h(x) + \zeta^T (x' - x) + \frac{\mu}{2}\|x' - x\|^2, \qquad \forall x, x' \in X, \forall \zeta \in \partial h(x). \tag{2.3}$$

A function $h(x)$ is ρ-weakly convex on X if $h + \frac{\rho}{2}\|\cdot\|^2$ is convex. This is equivalent to having:

$$h(x') \geq h(x) + \zeta^T(x' - x) - \frac{\rho}{2}\|x' - x\|^2, \qquad \forall x, x' \in X, \forall \zeta \in \partial h(x). \tag{2.4}$$

Without loss of generality, we simplify the m nonsmooth, nonconvex constraints of (1.1) into a single constraint as follows:

$$\begin{aligned} \min_{x \in X} \quad & f(x) \\ \text{s.t.} \quad & g(x) := \max_{i=1,\ldots,m} g_i(x) \leq 0. \end{aligned} \tag{2.5}$$

Not if each g_i is ρ-weakly convex, then g is ρ-weakly convex.

We follow the construction of (1.3) to build up our subproblems on the main problem (2.5). The subproblems are written as:

$$\begin{aligned} \min_{x \in X} \quad & F_k(x) := f(x) + \frac{\hat{\rho}}{2}\|x - x_k\|^2 \\ \text{s.t.} \quad & G_k(x) := g(x) + \frac{\hat{\rho}}{2}\|x - x_k\|^2 \leq 0. \end{aligned} \tag{2.6}$$

By properly selecting $\hat{\rho}$, both the objective function $F_k(x)$ and the constraint $G_k(x)$ are $(\hat{\rho} - \rho)$-strongly convex, and we set the feasibility tolerance. We find a nearly optimal and nearly feasible solution under the feasibility tolerance for the subproblem as our next iterate.

As the subproblems being approximately solved, our goal is to find approximate stationary solutions for our main problem (2.5). Without loss of generality, we describe the approximate stationarity for problem (1.1) according to Fritz John conditions and KKT conditions shown in (1.4) and (1.5), and give the following definitions.

Definition 2.1. *A point x is an ϵ-FJ point for problem (1.1) if $g_i(x) \leq 0$ $\forall i = 1, ..., m$, and there exists $\zeta_f \in \partial f(x)$, $\zeta_{gi} \in \partial g_i(x)$ and $\gamma_0 \geq 0$, $\gamma = (\gamma_1, ..., \gamma_m)^T \geq 0$, $\gamma_0 + \sum_{i=1}^m \gamma_i = 1$ such that:*

$$dist(\gamma_0\zeta_f + \sum_{i=1}^{m} \gamma_i\zeta_g, -N_X(x)) \leq \epsilon, \tag{2.7}$$

$$|\gamma_i g_i(x)| \leq \epsilon^2 \qquad \forall i = 1, ..., m. \tag{2.8}$$

Definition 2.2. *A point x is an ϵ-KKT point for problem (1.1) if $g_i(x) \leq 0$ $\forall i = 1, ..., m$, and there exists $\zeta_f \in \partial f(x)$, $\zeta_{gi} \in \partial g_i(x)$ and $\lambda = (\lambda_1, ..., \lambda_m)^T \geq 0$ such that:*

$$dist(\zeta_f + \sum_{i=1}^{m} \lambda_i\zeta_{gi}, -N_X(x)) \leq \epsilon, \tag{2.9}$$

$$|\lambda_i g_i(x)| \leq \epsilon^2 \qquad \forall i = 1, ..., m. \tag{2.10}$$

Let $\hat{x}_{k+1}$ denote the optimal solution for the subproblem (2.6), γ_{k0}, γ_k and λ_k correspond to the FJ and KKT stationarity at $\hat{x}_{k+1}$. We can show that for the main problem (2.5), with γ_{k0}, γ_k, λ_k and an appropriate $\hat{\rho}$, (2.8)/(2.10) is automatically satisfied when (2.7)/(2.9) is satisfied. This implies that when searching for approximate FJ or KKT stationary points for our main problem (2.5), we only need to consider whether (2.7) or (2.9) holds.

Inexact proximal point idea allows our method to find nearly optimal solutions for the subproblems (2.6) instead of exact solutions $\hat{x}_{k+1}$. Therefore, when $\hat{x}_{k+1}$ become an approximate stationary point for the main problem (2.5), we can only ensure x_k lying in the corresponding neighborhood of $\hat{x}_{k+1}$, but not exactly reach this approximate stationary solution. We give the following definitions to describe the points near the approximate stationary point.

Definition 2.3. *A point x is an (ϵ, η)-FJ point for problem (1.1) if there exists an ϵ-FJ point x' for problem (1.1) with $\|x - x'\| \leq \eta$.*

Definition 2.4. *A point x is an (ϵ, η)-KKT point for problem (1.1) if there exists an ϵ-KKT point x' for problem (1.1) with $\|x - x'\| \leq \eta$.*

Again, let $\hat{x}_{k+1}$ denote the optimal solution for the subproblem (2.6), γ_{k0}, γ_k and λ_k correspond to the FJ and KKT stationarity at $\hat{x}_{k+1}$. We can show that with γ_{k0}, γ_k, λ_k and an appropriate $\hat{\rho}$, if $\hat{x}_{k+1}$ is an approximate stationary point for the main problem (2.5), then the distance from x_k to $\hat{x}_{k+1}$ is bounded. This implies that as long as an approximate stationary point is found for our main problem (2.5), our iterate locates in its close neighborhood automatically.

The accuracy of KKT stationarity is proportional to the size of the optimal Lagrange multipliers. To guarantee KKT stationarity, it is necessary to give a constant upper bound for the optimal Lagrange multipliers in (1.5) for our subproblems (see problem (2.6)). Thus we define a stronger type of constraint qualification. Let $A(x) = \{i | g_i(x) = 0, i = 1, ..., m\}$. We say σ-strict MFCQ condition holds at x if there exists a constant $\sigma > 0$, such that:

$$\exists v \in -N_X^*(x) \quad and \quad \|v\| = 1 \qquad s.t. \quad \zeta_{gi}^T v \leq -\sigma \quad \forall i \in A(x), \forall \zeta_{gi} \in \partial g_i(x). \tag{2.11}$$

Specifically, when $N_X(x) = \{0\}$, we could equivalently state the condition as:

$$\|\zeta_{gi}\| \geq \sigma \quad \forall i \in A(x), \forall \zeta_{gi} \in \partial g_i(x). \tag{2.12}$$

Now we consider the σ-strong MFCQ condition for problem (1.1) based on the σ-strict MFCQ condition at x. The σ-strong MFCQ condition holds when σ-strict MFCQ condition is satisfied at any $x \in X$ with uniform σ. By assuming σ-strong MFCQ condition is satisfied for all the subproblems (see problem (2.6)), we could show boundedness for Lagrange multipliers in (1.5) for our subproblems, and attain convergence results to KKT points for the main problem (2.5).

Make the following four assumptions about (2.5) throughout the paper.

Assumption A. *$f(x)$ and $g(x)$ are continuous and ρ-weakly convex functions on X.*

Assumption B. *$f_{lb} = \inf_{x \in X} f(x) > -\infty$, $g_{lb} = \inf_{x \in X} g(x) > -\infty$.*

Assumption C. *For any $x \in X$, we can compute some $\zeta_f \in \partial f(x)$, $\zeta_g \in \partial g(x)$ with $\|\zeta_f\|, \|\zeta_g\| \leq M$.*

Assumption D. *We have access to an initial feasible point x_0 to problem* (2.5) *(i.e. $x_0 \in X$ and $g(x_0) \leq 0$).*

These assumptions suffice for our convergence theory to FJ points. Under the following additional assumption we show convergence results to KKT points.

Assumption E. *σ-strong MFCQ condition is satisfied for any subproblem* (2.6)*.*

For notation simplicity, let $D = \sqrt{\frac{-8g_{lb}}{\hat{\rho}-\rho}}$ indicate the diameter of the set $\{x|G_k(x) \leq 0\}$ due to the $(\hat{\rho} - \rho)$-strong convexity of $G_k(x)$, which in particular upper bounds the distance from the current iterate x_k to the optimal solution of the subproblem (2.6). Let $B = \frac{M+\hat{\rho}D}{\sigma}$ denote the uniform upper bound for the optimal dual variables of the subproblems (2.6) as Assumption E is satisfied.

Assume $\hat{x}_{k+1}$ is the optimal solution for the subproblem (2.6). Based on our previous illustrations, to ensure an (ϵ, ϵ)-FJ/KKT solution for the main problem (2.5), we only need to care about whether (2.7)/(2.9) is satisfied. We can show when $\|\hat{x}_{k+1} - x_k\|$ is less than or equal to some corresponding value, it is guaranteed that (2.7)/(2.9) holds. This implies that we attain approximate FJ or KKT stationary solutions for the main problem (2.5) as long as $\|\hat{x}_{k+1} - x_k\|$ is small enough. We provide the following Lemma to formalize this result.

Lemma 2.5. *Let $\hat{x}_{k+1}$ denote the optimal solution for the subproblem* (2.6) *with $\hat{\rho} > \max\{\rho, 1\}$. When Assumptions A-D hold and $\|\hat{x}_{k+1} - x_k\| \leq \frac{\epsilon}{\hat{\rho}}$, x_k is an (ϵ, ϵ)-FJ* *point; when Assumptions A-E hold and $\|\hat{x}_{k+1} - x_k\| \leq \frac{\epsilon}{\hat{\rho}(1+B)}$, x_k is an (ϵ, ϵ)-KKT point.*

Note that (2.7)/(2.9) is a sufficient and necessary condition for x_k to be an (ϵ, ϵ)-FJ/KKT point. However, Lemma 2.5 given above only indicates the sufficiency of the condition that $\|\hat{x}_{k+1} - x_k\|$ being small enough, without providing its necessity. In fact, when x_k is an (ϵ, ϵ)-KKT point, it is not necessary for $\|\hat{x}_{k+1} - x_k\| \leq \frac{\epsilon}{\hat{\rho}(1+B)}$ to be satisfied.

Chapter 3

Algorithms

In this section, we describe how the inexact proximal point method using the switching subgradient method as an oracle works for solving the main problem (2.5), and prove our main convergence guarantees. We first introduce the switching subgradient method for our constrained strongly convex subproblems, and then discuss the inexact proximal point method. All proofs are deferred to Section 4.

3.1 The Classic Switching Subgradient Method (without Lipschitz Continuity)

We introduce the classic switching subgradient method (see [35]) for solving the optimization problem below:

$$\begin{aligned} \min_{z \in Z} \quad & F(z) \\ \text{s.t.} \quad & G(z) \leq 0. \end{aligned} \tag{3.1}$$

Here we assume the domain Z is a convex set, and $F(z)$ and $G(z)$ are μ-strongly convex functions on Z. Let z^* be the optimal solution of this problem. Previous convergence analysis of algorithms for this problem usually assumed Lipschitz continuity for both $F(z)$ and $G(z)$, but we only need the following weaker condition previously considered for projected subgradient methods [50] as:

$$\begin{aligned}&\forall z \in Z,\ \forall \zeta_F \in \partial F(z),\ \zeta_G \in \partial G(z),\ \exists L_0 \geq 0,\ L_1 \geq 0,\\ &\text{s.t. } \|\zeta_F\|^2 \leq L_0^2 + L_1(F(z) - F(z^*)),\ \|\zeta_G\|^2 \leq L_0^2 + L_1(G(z) - G(z^*)).\end{aligned} \tag{3.2}$$

When $L_1 = 0$, $F(z)$ and $G(z)$ become L_0-Lipschitz continuous functions. Condition (3.2) allows $F(z)$ and $G(z)$ up to quadratic growth, which could be satisfied for strongly convex problems even in an unbounded domain Z. Note that it is impossible to have Lipschitz continuity for a strongly convex function in an unbounded domain. Specifically, for μ-strongly convex $h(x)$ defined on X' with x^* as its minimum point, strong convexity provides: $\forall x \in X', \forall \zeta_x \in \partial h(x), \|\zeta_x\| \geq \mu\|x - x^*\|$. As X' is unbounded, $\|x - x^*\|$ is unbounded, then $\|\zeta_x\|$ is also unbounded. This makes Lipschitz continuity of $h(x)$ fails to hold.

We define nearly optimal and nearly feasible solutions for problem (3.1).

Definition 3.1. *A point z is a (δ, τ)-optimal solution for problem (3.1) if $F(z) - F(z^*) \leq \delta$ and and $G(z) \leq \tau$, where z^* is the optimal solution.*

In fact, when we aim to find a (τ, τ)-optimal solution for problem (3.1) using the switching subgradient method (Algorithm 1), we only need to assume an even weaker condition based on (3.2), with the up to quadratic growth of $F(z)$ and $G(z)$ not necessarily satisfied in the whole domain Z. It can be written as:

$$\begin{aligned}&\forall z_1 \in \{z|G(z) \leq \tau\}, z_2 \in \{z|G(z) > \tau\},\ \forall \zeta_F \in \partial F(z_1),\ \zeta_G \in \partial G(z_2),\ \exists L_0 \geq 0,\ L_1 \geq 0,\\ &\text{s.t. } \|\zeta_F\|^2 \leq L_0^2 + L_1(F(z_1) - F(z^*)),\ \|\zeta_G\|^2 \leq L_0^2 + L_1(G(z_2) - G(z^*)).\end{aligned} \tag{3.3}$$

Here we analyze the switching subgradient method (Algorithm 1) to solve problem (3.1), finding a (τ, τ)-optimal solution for it. Basically, when the current iterate is not nearly feasible with tolerance τ, we compute the subgradient based on the constraint function and make an update seeking feasibility; otherwise we compute the subgradient of the objective function to make an update seeking optimality.

Algorithm 1 The Switching Subgradient Method

```
Require:
  $\mu$, $\tau > 0$, $z_0 \in \{z \in Z | G(z) \leq \tau\}$, $T > 0$, $\alpha_t$.
  set $I = \phi$, $J = \phi$
  for $t = 0, 1, ..., T-1$ do
    if $G(z_t) \leq \tau$ then
      $z_{t+1} = \text{proj}_Z(z_t - \alpha_t \zeta_{Ft})$, $\zeta_{Ft} \in \partial F(z_t)$, $I = I \bigcup \{t\}$
    else
      $z_{t+1} = \text{proj}_Z(z_t - \alpha_t \zeta_{Gt})$, $\zeta_{Gt} \in \partial G(z_t)$, $J = J \bigcup \{t\}$
    end if
  end for
Ensure:
  A $(\tau, \tau)$-optimal solution $\bar{z}_T = \frac{\sum_{t \in I}(t+1)F(z_t)}{\sum_{t \in I}(t+1)}$ for problem (3.1).
```

We give the convergence result for this method, generalizing [1, 36, 37] to non-Lipschitz settings.

Theorem 3.1. *With $\alpha_t = \frac{2}{\mu(t+2) + \frac{L_1^2}{\mu(t+1)}}$ and $\tau > 0$ in Algorithm 1, $\bar{z}_T$ is a (τ, τ)-optimal solution for problem (3.1) for all*

$$T \geq \max \left\{ \frac{8L_0^2}{\mu\tau}, \sqrt{\frac{2L_1^2 \|z_0 - z^*\|^2}{\mu\tau}} \right\}.$$

Note that the switching subgradient method can also attain a $(\tau, 0)$-optimal solution at the rate of $O(\tau^{-1})$ for problem (3.1), as long as we have access to a feasible initial point z_0 (i.e. $G(z_0) \leq 0$).

In our subproblem (2.6), F_k and G_k are both $(\hat{\rho} - \rho)$-strongly convex functions, which grow quadratically on potentially unbounded domain X. We can find upper bounds of quadratic growth (3.3) for F_k and G_k. The following Lemma is provided to show this result.

Lemma 3.2. *Condition (3.3) is satisfied for problem (2.6) with $L_0 = \sqrt{9M^2 - 6\hat{\rho} g_{lb}}$ and $L_1 = 6\hat{\rho}$.*

With Lemma 3.2 above, we apply the switching subgradient method to our subproblems (2.6) and give the following convergence result.

Corollary 3.3. *With $z_0 = x_k$, $\mu = \hat{\rho} - \rho$, $\alpha_t = \frac{2}{(\hat{\rho}-\rho)(t+2)+\frac{36\hat{\rho}^2}{(\hat{\rho}-\rho)(t+1)}}$ and $\tau > 0$ in Algorithm 1, $\bar{z}_T$ is a (τ, τ)-optimal solution for problem (2.6) for all*

$$T \geq \max\left\{\frac{24(3M^2 - 2\hat{\rho}g_{lb})}{\mu\tau}, \sqrt{\frac{72\hat{\rho}^2 D^2}{\mu\tau}}\right\}.$$

In previous convergence analysis of the switching subgradient method shown in other literature, the Lipschitz continuity assumption is necessary for both the objective function $F_k(x)$ and the constraint function $G_k(x)$. Due to their quadratic rates of growth, previous works require compactness of the domain X. In Corollary 3.3, we do not need Lipschitz continuity to guarantee the convergence of this method. As a result, compactness of X not needed to be assumed anymore.

Several stochastic variants of Algorithm 1 exist for solving stochastic versions of problem (3.1). An adaptive stochastic mirror descent method was introduced in [36] for the randomized version of problem (3.1), in which we could still get the exact functional values of the constraint, but can only evaluate the stochastic approximations of the subgradients of both the objective function and the constraint. With unbiased estimators of the subgradients available, Algorithm 1 can be applied to this kind of randomized problems with convergence results in expectation, without requiring the compactness of the domain or the stochastic subgradients to be bounded almost surely. A stochastic version of our quadratic growth upper bound (3.3) could be seen in [51] as a combination of the expected smoothness and finite gradient noise conditions around the optimal solution of problem (3.1), which is needed to show convergence of the stochastic version of Algorithm 1, without almost surely bounded subgradients being necessary anymore. In [37], they provided their alternating mirror descent stochastic approximation algorithm under stochastic estimations of the functional values of both the objective function and the constraint. Under this setting, they showed guarantees of finding nearly optimal solutions in expectation, still requiring the compactness of domain.

3.2 Proximally Guided Switching Subgradient Method

Using the switching subgradient method as our subproblem oracle, the inexact proximal point method searching for approximate stationary solutions for problem (2.5) proceeds according to Algorithm 2.

Algorithm 2 The Inexact Constrained Proximal Method

```
Require:
  A feasible point $x_0$ for problem (2.5), $\hat{\rho} > \max\{\rho, 1\}$, $\epsilon > 0$, parameters $\delta$ and $\tau$.
  for $k = 0, 1, ...$ do
    find $x_{k+1}$: a $(\delta, \tau)$-optimal solution for problem (2.6)
  end for
Ensure:
  An $(\epsilon, \epsilon)$-FJ or $(\epsilon, \epsilon)$-KKT point $x_{k-1}$ for problem (2.5).
```

To reach an (ϵ, ϵ)-FJ point, we pick:

$$\delta = \frac{(\hat{\rho} - \rho)\epsilon^2}{8\hat{\rho}^2} \quad \text{and} \quad \tau = \frac{(\hat{\rho} - \rho)\epsilon^2}{8\hat{\rho}^2}. \tag{3.4}$$

To reach an (ϵ, ϵ)-KKT point, we pick:

$$\delta = \frac{(\hat{\rho} - \rho)\epsilon^2}{8(1 + B)^2\hat{\rho}^2} \quad \text{and} \quad \tau = \frac{(\hat{\rho} - \rho)\epsilon^2}{8(1 + B)^2\hat{\rho}} \min\left\{\frac{1}{\hat{\rho} - \rho + \hat{\rho}B}, 1\right\}. \tag{3.5}$$

We guarantee the feasibility of our iterates x_k by the following two Lemmas.

Lemma 3.4. *Under Assumptions A–D with δ and τ as in (3.4), or Assumptions A–E with δ and τ as in (3.5), Algorithm 2 has $g(x_k) \le 0$ before x_k becomes an (ϵ, ϵ)-FJ point or an (ϵ, ϵ)-KKT point.*

Lemma 3.4 shows the automatic satisfaction of feasibility in Algorithm 2 until an (ϵ, ϵ)-FJ/KKT point is found and the algorithm stops.

In the framework of Algorithm 2, we adopt Algorithm 1 to serve as an oracle—a proximally guided switching subgradient method to attain nearly optimal and nearly feasible solutions for the subproblems (2.6), and finally reach an approximate stationary solution for the main problem (2.5). We provide the following convergence result of Algorithm 2 for reaching an (ϵ, ϵ)-FJ solution for problem (2.5).

Theorem 3.2. *Under Assumptions A–D with parameters as in (3.4), Algorithm 2 using Algorithm 1 as an oracle has x_K be an (ϵ, ϵ)-FJ point for problem (2.5) for some*

$$K < \frac{8\hat{\rho}^2(f(x_0) - f_{lb})}{3(\hat{\rho} - \rho)\epsilon^2} = \frac{\Delta_1}{\epsilon^2},$$

with $T = \max\left\{\frac{192\hat{\rho}^2(3M^2 - 2\hat{\rho}g_{lb})}{(\hat{\rho}-\rho)^2\epsilon^2}, \sqrt{\frac{576\hat{\rho}^4 D^2}{(\hat{\rho}-\rho)^2\epsilon^2}}\right\} = \max\left\{\frac{\Delta_2}{\epsilon^2}, \frac{\Delta_3}{\epsilon}\right\}$ steps of Algorithm 1 in each iteration of Algorithm 2, such an x_K is found using at most

$$\frac{\Delta_1 \max\{\Delta_2, \Delta_3\epsilon\}}{\epsilon^4}$$

total subgradient evaluations.

To guarantee an approximate KKT stationarity, it is necessary to give a uniform upper bound for the optimal dual variables (Lagrange multipliers) of the KKT conditions (1.5) for our subproblems (2.6). We show the Lemma below to achieve this.

Lemma 3.5. *Under Assumptions A–E, the optimal dual variables for problems (2.6) are uniformly upper bounded by $B = \frac{M + \hat{\rho}D}{\sigma}$.*

Then we provide the following convergence result of Algorithm 2 for reaching an (ϵ, ϵ)-KKT solution for problem (2.5).

Theorem 3.3. *Under Assumptions A–E with parameters as in (3.5), Algorithm 2 using Algorithm 1 as an oracle has x_K be an (ϵ, ϵ)-KKT point for problem (2.5) for* *some*

$$K < \frac{8(1+B)\hat{\rho}^2(f(x_0) - f_{lb})}{3(\hat{\rho} - \rho)\epsilon^2} = \frac{\Lambda_1}{\epsilon^2},$$

with $T = \max\left\{\frac{192(1+B)^2\hat{\rho}(3M^2 - 2\hat{\rho}g_{lb})\max\{\hat{\rho}-\rho+\hat{\rho}B,1\}}{(\hat{\rho}-\rho)^2\epsilon^2}, \sqrt{\frac{576(1+B)^2\hat{\rho}^3D^2\max\{\hat{\rho}-\rho+\hat{\rho}B,1\}}{(\hat{\rho}-\rho)^2\epsilon^2}}\right\} = \max\left\{\frac{\Lambda_2}{\epsilon^2}, \frac{\Lambda_3}{\epsilon}\right\}$ *steps of Algorithm 1 in each iteration of Algorithm 2, such an* x_K *is found using at most*

$$\frac{\Lambda_1 \max\{\Lambda_2, \Lambda_3\epsilon\}}{\epsilon^4}$$

total subgradient evaluations.

3.3 Stopping Criteria

While searching for approximate stationary points, we need to stop our algorithm as we reach our target. The stopping criteria for Algorithm 2 is shown as:

$$\|x_k - x_{k-1}\| \leq d_1 \quad \text{or} \quad g(x_k) > 0 \quad \text{or} \quad f(x_k) \geq f(x_{k-1}) - d_2. \tag{3.6}$$

In (3.6), to reach an (ϵ, ϵ)-FJ point, we pick:

$$d_1 = \frac{\epsilon}{2\hat{\rho}} \quad \text{and} \quad d_2 = \frac{3(\hat{\rho} - \rho)\epsilon^2}{8\hat{\rho}^2}. \tag{3.7}$$

To reach an (ϵ, ϵ)-KKT point, we pick:

$$d_1 = \frac{\sqrt{\hat{\rho} - \rho}\,\epsilon}{2(1+B)\sqrt{\hat{\rho} - \rho + \hat{\rho}B}\,\hat{\rho}} \quad \text{and} \quad d_2 = \frac{3(\hat{\rho} - \rho)\epsilon^2}{8(1+B)\hat{\rho}^2}. \tag{3.8}$$

The following two Lemmas about the stopping criteria help to guarantee Algorithm 2 to reach an (ϵ, ϵ)-FJ or (ϵ, ϵ)-KKT solution for problem (2.5).

Lemma 3.6. *Under Assumptions A–D with* d_1 *and* d_2 *as in* (3.7)*, the stopping criteria* (3.6) *does not hold when* x_{k-1} *is not an* (ϵ, ϵ)*-FJ point.*

Proof. When x_k is not an (ϵ, ϵ)-FJ point, we have $g(x_{k+1}) \leq 0$ due to Lemma 3.4, while (4.4) and (4.11) hold. Therefore, the stopping criteria (3.6) does not hold when x_{k-1} is not an (ϵ, ϵ)-FJ point. □

Lemma 3.7. *Under Assumptions A–E with d_1 and d_2 as in (3.8), the stopping criteria (3.6) does not hold when x_{k-1} is not an (ϵ, ϵ)-KKT point.*

Proof. When x_k is not an (ϵ, ϵ)-FJ point, we have $g(x_{k+1}) \leq 0$ due to Lemma 3.4, while (4.8) and (4.12) hold. Therefore, the stopping criteria (3.6) does not hold when x_{k-1} is not an (ϵ, ϵ)-FJ point. □

With Lemma 3.6 and Lemma 3.7 shown above, our stopping criteria is not satisfied when we have not reached an satisfactory approximate stationary point, and Algorithm 2 continues. When the stopping criteria holds, we reach our targeted approximate stationary point and stop our algorithm. Generally, it is possible that the stopping criteria fails to be satisfied when we have already achieved our targeted stationarity. In this case, it is reasonable to let Algorithm 2 continue working, since the break of (3.6) guarantees the feasibility and the least amount of descent of the next iterates, which will lead to convergence while the stopping criteria being satisfied finally.

3.4 Extension to Stochastic Gradient Oracles and High Probability Guarantees

In Section 3.1, we discussed the stochastic variants of the switching subgradient method. Under randomized problem settings with unbiased stochastic subgradient estimators and the measurements of the constraint function values being accurate, we can guarantee the targeted level of approximate feasibility for our subproblems, and preserve feasibility for our main problem. Since it is achievable for the expected objective value of our solution for the subproblem being sufficiently close to the optimal value after fixed number of iterations, it is possible that there exists a high probability guarantee for our objective value staying close to the expected value in a small size of neighborhood. Therefore, a high probability guarantee for the least amount of the objective value descent in each iteration of Algorithm 2 is attained, which indicates the high probability of the stochastic version of our method achieving approximate stationarity under a fixed computation complexity. Furthermore, when there are only noisy evaluations of the constraint functions available, we may also give a high probability guarantee for the approximate feasibility for our subproblems, which yields a high probability guarantee for the feasibility of our final solution for the main problem. This is similar as the way we guarantee the computation complexity of finding approximate stationarity in such stochastic settings.

Chapter 4

Convergence Analysis

4.1 Proof of Theorem 3.1 — Non-Lipschitz Strongly Convex Switching Subgradient Method Convergence

Our proof follows the styles of [1] and [50].

Let z^* be the optimal solution for problem (3.1), whose existence and uniqueness follow from strong convexity. Since the domain Z is convex, when $t \in I$, we have

$$\begin{aligned}\|z_{t+1} - z^*\|^2 &\leq \|z_t - \alpha_t \zeta_{Ft} - z^*\|^2 \\ &= \|z_t - z^*\|^2 - 2\alpha_t \zeta_{Ft}^T (z_t - z^*) + \alpha_t^2 \|\zeta_{Ft}\|^2 \\ &\leq \|z_t - z^*\|^2 - 2\alpha_t \zeta_{Ft}^T (z_t - z^*) + L_0^2 \alpha_t^2 + L_1 \alpha_t^2 (F(z_t) - F(z^*)).\end{aligned}$$

Since $F(z)$ is μ-strongly convex, (2.3) implies

$$\|z_{t+1} - z^*\|^2 \leq (1 - \mu\alpha_t)\|z_t - z^*\|^2 - (2\alpha_t - L_1\alpha_t^2)(F(z_t) - F(z^*)) + L_0^2\alpha_t^2$$

$$(2 - L_1\alpha_t)(F(z_t) - F(z^*)) \leq (\frac{1}{\alpha_t} - \mu)\|z_t - z^*\|^2 - \frac{1}{\alpha_t}\|z_{t+1} - z^*\|^2 + L_0^2\alpha_t.$$

Since $\alpha_t = \frac{2}{\mu(t+2) + \frac{L_1^2}{\mu(t+1)}}$, the above coefficient on $F(z_t) - F(z^*)$ is at least one

$$L_1\alpha_t = \frac{2L_1}{\mu(t+2)+\frac{L_1^2}{\mu(t+1)}} \le \frac{2L_1}{2\sqrt{\mu(t+2)\frac{L_1^2}{\mu(t+1)}}} \le 1.$$

Then the previous inequality becomes

$$F(z_t) - F(z^*) \le \frac{\mu t + \frac{L_1^2}{\mu(t+1)}}{2}\|z_t - z^*\|^2 - \frac{\mu(t+2)+\frac{L_1^2}{\mu(t+1)}}{2}\|z_{t+1} - z^*\|^2 + \frac{2L_0^2}{\mu(t+2)}.$$

Multiplying through by $(t+1)$ (a trick due to [1]) yields

$$(t+1)(F(z_t)-F(z^*)) \le \frac{\mu t(t+1)+\frac{L_1^2}{\mu}}{2}\|z_t - z^*\|^2 - \frac{\mu(t+1)(t+2)+\frac{L_1^2}{\mu}}{2}\|z_{t+1} - z^*\|^2 + \frac{2L_0^2}{\mu}.$$

Similarly, from the μ-strongly convex constraint $G(z)$, we have when $t \in J$ that

$$(t+1)(G(z_t)-G(z^*)) \le \frac{\mu t(t+1)+\frac{L_1^2}{\mu}}{2}\|z_t - z^*\|^2 - \frac{\mu(t+1)(t+2)+\frac{L_1^2}{\mu}}{2}\|z_{t+1} - z^*\|^2 + \frac{2L_0^2}{\mu}.$$

Summing the two inequalities above up for $t = 0, 1, 2, ..., T-1$ yields

$$\sum_{t\in I}(t+1)(F(z_t) - F(z^*)) + \sum_{t\in J}(t+1)(G(z_t) - G(z^*)) \le \frac{2L_0^2 T}{\mu} + \frac{L_1^2\|z_0 - z^*\|^2}{2\mu}.$$

For $t \in J$, we have $G(z_t) > \tau$. Since $G(z^*) \le 0$, we have $G(z_t) - G(z^*) > \tau$. Then the inequality becomes

$$\sum_{t\in I}(t+1)(F(z_t) - F(z^*)) + \sum_{t\in J}(t+1)\tau \le \frac{2L_0^2 T}{\mu} + \frac{L_1^2\|z_0 - z^*\|^2}{2\mu}.$$

Therefore, with $T \ge \max\{\frac{8L_0^2}{\mu\tau}, \sqrt{\frac{2L_1^2\|z_0-z^*\|^2}{\mu\tau}}\}$, we have

$$\begin{aligned}\sum_{t\in I}(t+1)(F(z_t)-F(z^*)) &\le \sum_{t\in I}(t+1)\tau - \sum_{t=0}^{T-1}(t+1)\tau + \frac{2L_0^2T}{\mu} + \frac{L_1^2\|z_0-z^*\|^2}{2\mu}\\ &= \sum_{t\in I}(t+1)\tau - \frac{T(T+1)}{2}\tau + \frac{2L_0^2T}{\mu} + \frac{L_1^2\|z_0-z^*\|^2}{2\mu}\\ &= \sum_{t\in I}(t+1)\tau - \frac{T\tau}{4}(T-\frac{8L_0^2}{\mu\tau}) - \frac{\tau}{4}(T^2-\frac{2L_1^2\|z_0-z^*\|^2}{\mu\tau})\\ &< \sum_{t\in I}(t+1)\tau.\end{aligned}$$

The convexity of $F(z)$ gives us

$$F(\bar{z}_T)-F(z^*) = F\left(\frac{\sum_{t\in I}(t+1)z_t}{\sum_{t\in I}(t+1)}\right) - F(z^*) \le \frac{\sum_{t\in I}(t+1)F(z_t)}{\sum_{t\in I}(t+1)} - F(z^*) < \tau.$$

The convexity of $G(z)$ gives us

$$G(\bar{z}_T) = G\left(\frac{\sum_{t\in I}(t+1)z_t}{\sum_{t\in I}(t+1)}\right) \le \frac{\sum_{t\in I}(t+1)G(z_t)}{\sum_{t\in I}(t+1)} < \tau.$$

Therefore, $\bar{z}_T$ is a (τ,τ)-optimal solution for problem (2.6).

4.2 Proof of Theorem 3.2

According to Lemma 3.4, our iterates x_k are always feasible, that is $g(x_k)\le 0$, for the main problem (2.5) before we reach an (ϵ,ϵ)-FJ point. For any x_k, with γ_{k0} and γ_k defined in (1.4) for problem (2.6), we construct the function $\mathcal{L}_k(x)$ for problem (2.6) as

$$\mathcal{L}_k(x) = \gamma_{k0}F_k(x)+\gamma_kG_k(x) = \gamma_{k0}(f(x)+\frac{\hat{\rho}}{2}\|x-x_k\|^2)+\gamma_k(g(x)+\frac{\hat{\rho}}{2}\|x-x_k\|^2). \quad (4.1)$$

Without loss of generality, suppose $\gamma_{k0}\ge 0$, $\gamma_k\ge 0$, and $\gamma_{k0}+\gamma_k=1$. Let $\hat{x}_{k+1}$ be the exact solution for problem (2.6). According to FJ conditions (1.4), there exists $\hat{\zeta}_{Fk}\in\partial F_k(\hat{x}_{k+1})$ and $\hat{\zeta}_{Gk}\in\partial G_k(\hat{x}_{k+1})$ which satisfies

$$\gamma_{k0}\hat{\zeta}_{Fk} + \gamma_k\hat{\zeta}_{Gk} \in -N_X(\hat{x}_{k+1}). \tag{4.2}$$

Since $\mathcal{L}_k(x)$ is $(\hat{\rho}-\rho)$-strongly convex, we have

$$\begin{aligned}\gamma_{k0}F_k(x_k) + \gamma_k G_k(x_k) \geq & \gamma_{k0}F_k(\hat{x}_{k+1}) + \gamma_k G_k(\hat{x}_{k+1}) + (\gamma_{k0}\hat{\zeta}_{Fk} + \gamma_k\hat{\zeta}_{Gk})^T(x_k - \hat{x}_{k+1}) \\ & + \frac{\hat{\rho}-\rho}{2}\|x_k - \hat{x}_{k+1}\|^2.\end{aligned}$$

According to FJ conditions, we also have $\gamma_k G_k(\hat{x}_{k+1}) = 0$. By (4.2) and since $x_k \in X$, we know $(\gamma_{k0}\hat{\zeta}_{Fk} + \gamma_k\hat{\zeta}_{Gk})^T(x_k - \hat{x}_{k+1}) \geq 0$. Since $g(x_k) \leq 0$ from Lemma 3.4, the previous inequality becomes

$$\gamma_{k0}f(x_k) \geq \gamma_{k0}F_k(\hat{x}_{k+1}) + \frac{\hat{\rho}-\rho}{2}\|\hat{x}_{k+1} - x_k\|^2.$$

Since x_{k+1} is a (δ, τ) solution for problem (2.6), $F_k(x_{k+1}) - F_k(\hat{x}_{k+1}) \leq \delta$. Then the previous inequality becomes

$$\begin{aligned}\gamma_{k0}f(x_k) &\geq \gamma_{k0}(f(x_{k+1}) + \frac{\hat{\rho}}{2}\|x_{k+1} - x_k\|^2 - \delta) + \frac{\hat{\rho}-\rho}{2}\|\hat{x}_{k+1} - x_k\|^2 \\ &\geq \gamma_{k0}(f(x_{k+1}) - \delta) + \frac{\hat{\rho}-\rho}{2}\|\hat{x}_{k+1} - x_k\|^2.\end{aligned}$$

Thus we attain a lower bound for descent of each step as

$$\gamma_{k0}(f(x_k) - f(x_{k+1})) \geq \frac{\hat{\rho}-\rho}{2}\|\hat{x}_{k+1} - x_k\|^2 - \gamma_{k0}\delta. \tag{4.3}$$

When $\gamma_{k0} = 0$, then $\|\hat{x}_{k+1} - x_k\| = 0$ and we have already reached an exact stationary point x_k for problem (2.5). Now we consider the case that $\gamma_{k0} > 0$ here. Let $\hat{\zeta}_{fk} = \hat{\zeta}_{Fk} - \hat{\rho}(\hat{x}_{k+1} - x_k) \in \partial f(\hat{x}_{k+1})$, $\hat{\zeta}_{gk} = \hat{\zeta}_{Gk} - \hat{\rho}(\hat{x}_{k+1} - x_k) \in \partial g(\hat{x}_{k+1})$. According to (4.2), before we reach the (ϵ, ϵ)-FJ point $\hat{x}_{k+1}$, there exists $\nu \in N_X(\hat{x}_{k+1})$ which satisfies:

$$\gamma_{k0}(\hat{\zeta}_{fk} + \hat{\rho}(\hat{x}_{k+1} - x_k)) + \gamma_k(\hat{\zeta}_{gk} + \hat{\rho}(\hat{x}_{k+1} - x_k)) + \nu = 0,$$

$$\|\gamma_{k0}\hat{\zeta}_{fk} + \gamma_k\hat{\zeta}_{gk} + \nu\| > \epsilon.$$

Then

$$\|\hat{x}_{k+1} - x_k\| > \frac{\epsilon}{\hat{\rho}}.$$

Thus, before we reach the (ϵ, ϵ)-FJ point $\hat{x}_{k+1}$, it can be derived from (4.3) that

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \frac{\hat{\rho} - \rho}{2\gamma_{k0}}\|\hat{x}_{k+1} - x_k\|^2 - \delta \\ &\geq \frac{\hat{\rho} - \rho}{2}\|\hat{x}_{k+1} - x_k\|^2 - \delta \\ &> \frac{\hat{\rho} - \rho}{2} \times \frac{\epsilon^2}{\hat{\rho}^2} - \frac{(\hat{\rho} - \rho)\epsilon^2}{8\hat{\rho}^2} \\ &= \frac{3(\hat{\rho} - \rho)\epsilon^2}{8\hat{\rho}^2}. \end{aligned} \tag{4.4}$$

By Assumption B, we could give an upper bound for the number of total iterations K of Algorithm 2 as

$$K < \frac{8\hat{\rho}^2(f(x_0) - f_{lb})}{3(\hat{\rho} - \rho)\epsilon^2}.$$

Using Algorithm 1 as an oracle for Algorithm 2, with $\tau = \frac{(\hat{\rho}-\rho)\epsilon^2}{8\hat{\rho}^2}$, L_0 and L_1 from Lemma 3.2. Taking the number of steps of Algorithm 1 as $T = \max\left\{\frac{8L_0^2}{(\hat{\rho}-\rho)\tau}, \sqrt{\frac{2L_1^2D^2}{(\hat{\rho}-\rho)\tau}}\right\}$ in each iterations of Algorithm 2, the number of total subgradient evaluations is upper bounded as

$$KT < \frac{8\hat{\rho}^2(f(x_0) - f_{lb})}{3(\hat{\rho} - \rho)\epsilon^2} \max\left\{\frac{192\hat{\rho}^2(3M^2 - 2\hat{\rho}g_{lb})}{(\hat{\rho} - \rho)^2\epsilon^2}, \sqrt{\frac{576\hat{\rho}^4D^2}{(\hat{\rho} - \rho)^2\epsilon^2}}\right\}.$$

4.3 Proof of Theorem 3.3

According to Lemma 3.4, our iterates are always feasible, that is $g(x_k) \leq 0$, for the main problem (2.5) before we reach an (ϵ, ϵ)-KKT point. For any x_k, with λ_k defined in (1.5) for problem (2.6), we construct the Lagrange function for problem (2.6) as

$$L_k(x) = F_k(x) + \lambda_k G_k(x) = f(x) + \frac{\hat{\rho}}{2}\|x - x_k\|^2 + \lambda_k(g(x) + \frac{\hat{\rho}}{2}\|x - x_k\|^2). \tag{4.5}$$

Without loss of generality, suppose $\lambda_k \geq 0$. Let $\hat{x}_{k+1}$ be the exact solution for problem (2.6). According to KKT conditions (1.5), there exists $\hat{\zeta}_{Fk} \in \partial F_k(\hat{x}_{k+1})$ and $\hat{\zeta}_{Gk} \in \partial G_k(\hat{x}_{k+1})$ which satisfies

$$\hat{\zeta}_{Fk} + \lambda_k \hat{\zeta}_{Gk} \in -N_X(\hat{x}_{k+1}). \tag{4.6}$$

Since $L_k(x)$ is $(1 + \lambda_k)(\hat{\rho} - \rho)$-strongly convex, we have

$$\begin{aligned} F_k(x_k) + \lambda_k G_k(x_k) \geq & F_k(\hat{x}_{k+1}) + \lambda_k G_k(\hat{x}_{k+1}) + (\hat{\zeta}_{Fk} + \lambda_k \hat{\zeta}_{Gk})^T (x_k - \hat{x}_{k+1}) \\ & + \frac{(1 + \lambda_k)(\hat{\rho} - \rho)}{2}\|\hat{x}_{k+1} - x_k\|^2. \end{aligned}$$

According to KKT conditions (1.5), we also have $\lambda_k G_k(\hat{x}_{k+1}) = 0$. By (4.6) and since $x_k \in X$, we know $(\hat{\zeta}_{Fk} + \lambda_k \hat{\zeta}_{Gk})^T (x_k - \hat{x}_{k+1}) \geq 0$. Since $g(x_k) \leq 0$ from Lemma 3.4, the previous inequality becomes

$$f(x_k) \geq F_k(\hat{x}_{k+1}) + \frac{\hat{\rho} - \rho}{2}\|\hat{x}_{k+1} - x_k\|^2.$$

Since x_{k+1} is a (δ, τ) solution for problem (2.6), $F_k(x_{k+1}) - F_k(\hat{x}_{k+1}) \leq \delta$, then

$$\begin{aligned} f(x_k) &\geq (f(x_{k+1}) + \frac{\hat{\rho}}{2}\|x_{k+1} - x_k\|^2 - \delta) + \frac{(1 + \lambda_k)(\hat{\rho} - \rho)}{2}\|\hat{x}_{k+1} - x_k\|^2 \\ &\geq f(x_{k+1}) - \delta + \frac{(1 + \lambda_k)(\hat{\rho} - \rho)}{2}\|\hat{x}_{k+1} - x_k\|^2. \end{aligned}$$

Thus we attain a lower bound for descent of each step as

$$f(x_k) - f(x_{k+1}) \geq \frac{(1+\lambda_k)(\hat{\rho}-\rho)}{2}\|\hat{x}_{k+1} - x_k\|^2 - \delta. \tag{4.7}$$

Let $\hat{\zeta}_{fk} = \hat{\zeta}_{Fk} - \hat{\rho}(\hat{x}_{k+1} - x_k) \in \partial f(\hat{x}_{k+1})$, $\hat{\zeta}_{gk} = \hat{\zeta}_{Gk} - \hat{\rho}(\hat{x}_{k+1} - x_k) \in \partial g(\hat{x}_{k+1})$. According to (4.6), before we reach the (ϵ, ϵ)-KKT point $\hat{x}_{k+1}$, there exists $\nu \in N_X(\hat{x}_{k+1})$ which satisfies:

$$(\hat{\zeta}_{fk} + \hat{\rho}(\hat{x}_{k+1} - x_k)) + \lambda_k(\hat{\zeta}_{gk} + \hat{\rho}(\hat{x}_{k+1} - x_k)) + \nu = 0,$$

$$\|\hat{\zeta}_{fk} + \lambda_k \hat{\zeta}_{gk} + \nu\| > \epsilon.$$

Then

$$\|\hat{x}_{k+1} - x_k\| > \frac{\epsilon}{(1+\lambda_k)\hat{\rho}}.$$

Thus, before we reach the (ϵ, ϵ)-KKT point $\hat{x}_{k+1}$, apply Lemma 3.5 here, it can be derived from (4.7) that

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \frac{(1+\lambda_k)(\hat{\rho}-\rho)}{2}\|\hat{x}_{k+1} - x_k\|^2 - \delta \\ &> \frac{(1+\lambda_k)(\hat{\rho}-\rho)}{2} \times \frac{\epsilon^2}{(1+\lambda_k)^2\hat{\rho}^2} - \frac{(\hat{\rho}-\rho)\epsilon^2}{8(1+B)^2\hat{\rho}^2} \\ &> \frac{(\hat{\rho}-\rho)\epsilon^2}{2(1+\lambda_k)\hat{\rho}^2} - \frac{(\hat{\rho}-\rho)\epsilon^2}{8(1+B)\hat{\rho}^2} \\ &\geq \frac{(\hat{\rho}-\rho)\epsilon^2}{2(1+B)\hat{\rho}^2} - \frac{(\hat{\rho}-\rho)\epsilon^2}{8(1+B)\hat{\rho}^2} \\ &= \frac{3(\hat{\rho}-\rho)\epsilon^2}{8(1+B)\hat{\rho}^2}. \end{aligned} \tag{4.8}$$

By Assumption B, we could give an upper bound for the number of total iterations K as

$$K < \frac{8(1+B)\hat{\rho}^2(f(x_0) - f_{lb})}{3(\hat{\rho} - \rho)\epsilon^2}.$$

Using Algorithm 1 as an oracle for Algorithm 2, with $\tau = \frac{(\hat{\rho}-\rho)\epsilon^2}{8(1+B)^2\hat{\rho}} \min\left\{\frac{1}{\hat{\rho}-\rho+\hat{\rho}B}, 1\right\}$, L_0 and L_1 from Lemma 3.2. Taking the number of steps of Algorithm 1 as $T = \max\left\{\frac{8L_0^2}{(\hat{\rho}-\rho)\tau}, \sqrt{\frac{2L_1^2D^2}{(\hat{\rho}-\rho)\tau}}\right\}$ in each iterations of Algorithm 2, the number of total subgradient evaluations is upper bounded as

$$KT < \frac{8(1+B)\hat{\rho}^2(f(x_0) - f_{lb})}{3(\hat{\rho} - \rho)\epsilon^2} \max\left\{\frac{192(1+B)^2\hat{\rho}(3M^2 - 2\hat{\rho}g_{lb})\max\{\hat{\rho} - \rho + \hat{\rho}B, 1\}}{(\hat{\rho} - \rho)^2\epsilon^2}, \sqrt{\frac{576(1+B)^2\hat{\rho}^3D^2\max\{\hat{\rho} - \rho + \hat{\rho}B, 1\}}{(\hat{\rho} - \rho)^2\epsilon^2}}\right\}.$$

4.4 Proof of Lemmas

4.4.1 Proof of Lemma 2.5

Given $\hat{\rho} > \max\{\rho, 1\}$, we discuss for FJ and KKT stationarity respectively.

For FJ stationarity, assume FJ conditions (1.4) are satisfied for problem (2.6) at $\hat{x}_{k+1}$ with $\gamma_{k0} \geq 0$, $\gamma_k \geq 0$, $\gamma_{k0} + \gamma_k = 1$, $\hat{\zeta}_{Fk} \in \partial F_k(\hat{x}_{k+1})$ and $\hat{\zeta}_{Gk} \in \partial G_k(\hat{x}_{k+1})$. Let $\hat{\zeta}_{fk} = \hat{\zeta}_{Fk} - \hat{\rho}(\hat{x}_{k+1} - x_k) \in \partial f(\hat{x}_{k+1})$ and $\hat{\zeta}_{gk} = \hat{\zeta}_{Gk} - \hat{\rho}(\hat{x}_{k+1} - x_k) \in \partial g(\hat{x}_{k+1})$, then there exists $\nu \in N_X(\hat{x}_{k+1})$ such that

$$\gamma_{k0}(\hat{\zeta}_{fk} + \hat{\rho}(\hat{x}_{k+1} - x_k)) + \gamma_k(\hat{\zeta}_{gk} + \hat{\rho}(\hat{x}_{k+1} - x_k)) = -\nu.$$

When $\|\hat{x}_{k+1} - x_k\| \leq \frac{\epsilon}{\hat{\rho}}$, we have

$$\|\gamma_{k0}\hat{\zeta}_{fk} + \gamma_k\hat{\zeta}_{gk} + \nu\| = \hat{\rho}\|\hat{x}_{k+1} - x_k\| \leq \epsilon.$$

Then (2.7) is satisfied. When $\gamma_k = 0$, $|\gamma_k g(\hat{x}_{k+1})| = 0$, thus we only consider the case that γ_k is positive. In this case, we have $G_k(\hat{x}_{k+1}) = 0$ according to FJ conditions, then

$$0 \geq g(\hat{x}_{k+1}) = -\frac{\hat{\rho}}{2}\|\hat{x}_{k+1} - x_k\|^2 \geq -\frac{\epsilon^2}{2\hat{\rho}}.$$

Therefore

$$|\gamma_k g(\hat{x}_{k+1})| \leq |g(\hat{x}_{k+1})| \leq \frac{\epsilon^2}{2\hat{\rho}} < \epsilon^2.$$

Then (2.8) is satisfied, and $\hat{x}_{k+1}$ is an ϵ-FJ point for problem (2.5). Due to $\|\hat{x}_{k+1} - x_k\| \leq \frac{\epsilon}{\hat{\rho}} < \epsilon$, x_k is an (ϵ, ϵ)-FJ point for problem (2.5).

Similarly, for KKT stationarity, assume KKT conditions (1.5) are satisfied for problem (2.5) with $\lambda_k \geq 0$, $\hat{\zeta}_{Fk} \in \partial F_k(\hat{x}_{k+1})$ and $\hat{\zeta}_{Gk} \in \partial G_k(\hat{x}_{k+1})$. Let $\hat{\zeta}_{fk} = \hat{\zeta}_{Fk} - \hat{\rho}(\hat{x}_{k+1} - x_k) \in \partial f(\hat{x}_{k+1})$ and $\hat{\zeta}_{gk} = \hat{\zeta}_{Gk} - \hat{\rho}(\hat{x}_{k+1} - x_k) \in \partial g(\hat{x}_{k+1})$, then there exists $\nu \in N_X(\hat{x}_{k+1})$ such that

$$\hat{\zeta}_{fk} + \hat{\rho}(\hat{x}_{k+1} - x_k) + \lambda_k(\hat{\zeta}_{gk} + \hat{\rho}(\hat{x}_{k+1} - x_k)) = -\nu.$$

When $\|\hat{x}_{k+1} - x_k\| \leq \frac{\epsilon}{\hat{\rho}(1+B)}$, apply Lemma (3.5) here, we have

$$\|\hat{\zeta}_{fk} + \lambda_k \hat{\zeta}_{gk} + \nu\| = \hat{\rho}(1 + \lambda_k)\|\hat{x}_{k+1} - x_k\| \leq \epsilon.$$

Then (2.9) is satisfied. When $\lambda_k = 0$, $|\lambda_k g(\hat{x}_{k+1})| = 0$, thus we only consider the case that λ_k is positive. In this case, we have $G_k(\hat{x}_{k+1}) = 0$ according to KKT conditions, then

$$0 \geq g(\hat{x}_{k+1}) = -\frac{\hat{\rho}}{2}\|\hat{x}_{k+1} - x_k\|^2 \geq -\frac{\epsilon^2}{2\hat{\rho}(1+B)^2}.$$

Therefore

$$|\lambda_k g(\hat{x}_{k+1})| \le \frac{\epsilon^2 \lambda_k}{2\hat{\rho}(1+B)^2} \le \frac{\epsilon^2}{2\hat{\rho}} < \epsilon^2.$$

Then (2.10) is satisfied, and $\hat{x}_{k+1}$ is an ϵ-KKT point for problem (2.5). Due to $\|\hat{x}_{k+1} - x_k\| \le \frac{\epsilon}{\hat{\rho}(1+B)} < \epsilon$, x_k is an (ϵ, ϵ)-KKT point for problem (2.5).

4.4.2 Proof of Lemma 3.2

Let $z^* = \hat{x}_{k+1}$ be the optimal solution for problem (2.6), and $\mu = \hat{\rho} - \rho$. Let $\zeta_{Fk} \in \partial F_k(z)$, $\zeta_{Gk} \in \partial G_k(z)$, $\zeta_f = \zeta_{Fk} - \hat{\rho}(z - z_0) \in \partial f(z)$, and $\zeta_g = \zeta_{Gk} - \hat{\rho}(z - z_0) \in \partial g(z)$. We aim to find L_0 and L_1 satisfying the quadratic growth condition below:

$$\forall z \in \{z | G_k(z) \le \tau\}, \|\zeta_{Fk}\|^2 \le L_0^2 + L_1(F_k(z) - F_k(z^*)), \tag{4.9}$$

$$\forall z \in \{z | G_k(z) > \tau\}, \|\zeta_{Gk}\|^2 \le L_0^2 + L_1(G_k(z) - G_k(z^*)). \tag{4.10}$$

For $F_k(z)$ we have

$$\begin{aligned}
& L_0^2 + L_1(F_k(z) - F_k(z^*)) \\
= & L_0^2 + L_1(F_k(z) - F_k(z^*)) - \|\zeta_{Fk}\|^2 + \|\zeta_{Fk}\|^2 \\
\ge & L_0^2 + L_1(f(z) + \frac{\hat{\rho}}{2}\|z - z_0\|^2 - F_k(z_0)) - \|\zeta_f + \hat{\rho}(z - z_0)\|^2 + \|\zeta_{Fk}\|^2 \\
= & L_0^2 + L_1(f(z) - f(z_0)) + \frac{L_1\hat{\rho}}{2}\|z - z_0\|^2 - \|\zeta_f\|^2 - 2\hat{\rho}\zeta_f^T(z - z_0) - \hat{\rho}^2\|z - z_0\|^2 + \|\zeta_{Fk}\|^2 \\
\ge & L_0^2 - L_1 M\|z - z_0\| + \frac{L_1\hat{\rho}}{2}\|z - z_0\|^2 - M^2 - 2\hat{\rho}\|\zeta_f\|\|z - z_0\| - \hat{\rho}^2\|z - z_0\|^2 + \|\zeta_{Fk}\|^2 \\
= & (L_0^2 - M^2) - (L_1 + 2\hat{\rho})M\|z - z_0\| + (\frac{L_1}{2} - \hat{\rho})\hat{\rho}\|z - z_0\|^2 + \|\zeta_{Fk}\|^2 \\
\overset{L_1 \ge 2\hat{\rho}}{=} & \underbrace{\left(\frac{L_1}{2} - \hat{\rho}\right)\hat{\rho}\left(\|z - z_0\| - \frac{(L_1 + 2\hat{\rho})M}{(L_1 - 2\hat{\rho})\hat{\rho}}\right)^2 + \left(L_0^2 - M^2 - \frac{(L_1 + 2\hat{\rho})^2 M^2}{2(L_1 - 2\hat{\rho})\hat{\rho}}\right)}_{A_1} + \|\zeta_{Fk}\|^2.
\end{aligned}$$

Pick $L_0 = 3M, L_1 = 6\hat{\rho}$, then $A_1 \ge 0$, and (4.9) is satisfied.

Similarly, for $G_k(z)$ we have

$$
\begin{aligned}
& L_0^2 + L_1(G_k(z) - G_k(z^*)) \\
&= L_0^2 + L_1(G_k(z) - G_k(z^*)) - \|\zeta_{Gk}\|^2 + \|\zeta_{Gk}\|^2 \\
&= L_0^2 + L_1(g(z) + \frac{\hat{\rho}}{2}\|z - z_0\|^2) - \|\zeta_g + \hat{\rho}(z - z_0)\|^2 + \|\zeta_{Gk}\|^2 \\
&= L_0^2 + L_1 g(z_0) + L_1(g(z) - g(z_0)) + \frac{L_1\hat{\rho}}{2}\|z - z_0\|^2 - \|\zeta_g\|^2 - 2\hat{\rho}\zeta_g^T(z - z_0) - \hat{\rho}^2\|z - z_0\|^2 + \|\zeta_{Gk}\|^2 \\
&\geq L_0^2 + L_1 g(z_0) - L_1 M\|z - z_0\| + \frac{L_1\hat{\rho}}{2}\|z - z_0\|^2 - M^2 - 2\hat{\rho}\|\zeta_g\|\|z - z_0\| - \hat{\rho}^2\|z - z_0\|^2 + \|\zeta_{Gk}\|^2 \\
&= (L_0^2 - M^2 + L_1 g(z_0)) - (L_1 + 2\hat{\rho})M\|z - z_0\| + (\frac{L_1}{2} - \hat{\rho})\hat{\rho}\|z - z_0\|^2 + \|\zeta_{Gk}\|^2 \\
&\overset{L_1 \geq 2\hat{\rho}}{=} \underbrace{\left(\frac{L_1}{2} - \hat{\rho}\right)\hat{\rho}\left(\|z - z_0\| - \frac{(L_1 + 2\hat{\rho})M}{(L_1 - 2\hat{\rho})\hat{\rho}}\right)^2 + \left(L_0^2 - M^2 + L_1 g(z_0) - \frac{(L_1 + 2\hat{\rho})^2 M^2}{2(L_1 - 2\hat{\rho})\hat{\rho}}\right)}_{A_2} + \|\zeta_{Gk}\|^2.
\end{aligned}
$$

Pick $L_0 = \sqrt{9M^2 - 6\hat{\rho}g(z_0)}, L_1 = 6\hat{\rho}$, then $A_2 \geq 0$, and (4.10) is satisfied.

Since we always have $g(z_0) \geq g_{lb}$, we pick $L_0 = \sqrt{9M^2 - 6\hat{\rho}g(z_0)}$ and $L_1 = 6\hat{\rho}$ as a uniform choice satisfied for all z_0, which makes (4.9) and (4.10) hold.

4.4.3 Proof of Lemma 3.4

First we show the feasibility of the iterates x_k before reaching an (ϵ, ϵ)-FJ point, with δ and τ in (3.4). Assume $G_k(x_k) = g(x_k) \leq 0$. For function $\mathcal{L}_k(x)$ (see (4.1)), $(\hat{\rho} - \rho)$-strong convexity gives us

$$
\begin{aligned}
\gamma_{k0}F_k(x_{k+1}) + \gamma_k G_k(x_{k+1}) \geq & \gamma_{k0}F_k(\hat{x}_{k+1}) + \gamma_k G_k(\hat{x}_{k+1}) + (\gamma_{k0}\hat{\zeta}_{Fk} + \gamma_k\hat{\zeta}_{Gk})^T(x_{k+1} \\
& - \hat{x}_{k+1}) + \frac{\hat{\rho} - \rho}{2}\|x_{k+1} - \hat{x}_{k+1}\|^2.
\end{aligned}
$$

Since $\lambda_{k0}\hat{\zeta}_{Fk} + \lambda_k\hat{\zeta}_{Gk} \in -N_X(\hat{x}_{k+1})$ by FJ conditions, and due to $x_{k+1} \in X$, we know

$$
(\gamma_{k0}\zeta_{Fk} + \gamma_k\zeta_{Gk})^T(x_{k+1} - \hat{x}_{k+1}) \geq 0.
$$

Also since $\gamma_k G_k(\hat{x}_{k+1}) = 0$ by FJ conditions, and x_{k+1} being an (δ, τ)-optimal solution for the subproblem (2.6) yields $F_k(x_{k+1}) - F_k(\hat{x}_{k+1}) \leq \delta$ and $G_k(x_{k+1}) \leq \tau$, then the first inequality becomes

$$\begin{aligned} \gamma_{k0}\delta + \gamma_k\tau &\geq \frac{\hat{\rho} - \rho}{2}\|\hat{x}_{k+1} - x_{k+1}\|^2 \\ \|\hat{x}_{k+1} - x_{k+1}\| &\leq \sqrt{\frac{2(\gamma_{k0}\delta + \gamma_k\tau)}{\hat{\rho} - \rho}}. \end{aligned}$$

Let $\hat{\zeta}_{fk} = \hat{\zeta}_{Fk} - \hat{\rho}(\hat{x}_{k+1} - x_k) \in \partial f(\hat{x}_{k+1})$, $\hat{\zeta}_{gk} = \hat{\zeta}_{Gk} - \hat{\rho}(\hat{x}_{k+1} - x_k) \in \partial g(\hat{x}_{k+1})$. Before we reach the (ϵ, ϵ)-FJ point $\hat{x}_{k+1}$ for the main problem (2.5), $\gamma_{k0} > 0$, according to the FJ conditions, there exists $\nu \in N_X(\hat{x}_{k+1})$ such that $\gamma_{k0}(\hat{\zeta}_{fk} + \hat{\rho}(\hat{x}_{k+1} - x_k)) + \gamma_k(\hat{\zeta}_{gk} + \hat{\rho}(\hat{x}_{k+1} - x_k)) + \nu = 0$ and $\|\gamma_{k0}\hat{\zeta}_{fk} + \gamma_k\hat{\zeta}_{gk} + \nu\| > \epsilon$, which yields $\|\hat{x}_{k+1} - x_k\| > \frac{\epsilon}{\hat{\rho}}$. Thus

$$\|x_{k+1} - x_k\|^2 \geq \frac{1}{2}\|\hat{x}_{k+1} - x_k\|^2 - \|\hat{x}_{k+1} - x_{k+1}\|^2 > \frac{\epsilon^2}{2\hat{\rho}^2} - \frac{2(\gamma_{k0}\delta + \gamma_k\tau)}{\hat{\rho} - \rho}.$$

Take $\delta = \frac{(\hat{\rho}-\rho)\epsilon^2}{8\hat{\rho}^2}$ and $\tau = \frac{(\hat{\rho}-\rho)\epsilon^2}{8\hat{\rho}^2}$, due to the fact that $\delta = \tau$, we have

$$\|x_{k+1} - x_k\|^2 > \frac{\epsilon^2}{2\hat{\rho}^2} - \frac{2\delta}{\hat{\rho} - \rho} = \frac{\epsilon^2}{2\hat{\rho}^2} - \frac{\epsilon^2}{4\hat{\rho}^2} = \frac{\epsilon^2}{4\hat{\rho}^2}. \tag{4.11}$$

Therefore

$$\begin{aligned} g(x_{k+1}) &= G(x_{k+1}) - \frac{\hat{\rho}}{2}\|x_{k+1} - x_k\|^2 \\ &< \tau - \frac{\hat{\rho}}{2} \times \frac{\epsilon^2}{4\hat{\rho}^2} \leq \frac{(\hat{\rho} - \rho)\epsilon^2}{8\hat{\rho}^2} - \frac{\epsilon^2}{8\hat{\rho}} = -\frac{\rho\epsilon^2}{8\hat{\rho}^2} < 0. \end{aligned}$$

This indicates that $g(x_{k+1}) \leq 0$ automatically if $g(x_k) \leq 0$, from which we can induce that all the iterates are feasible for the main problem (2.5) before reaching an (ϵ, ϵ)-FJ point.

Next we show the feasibility of the iterates x_k before reaching an (ϵ, ϵ)-KKT point, with δ and τ in (3.5). Assume $G_k(x_k) = g(x_k) \leq 0$. For the Lagrange function (see (4.5)), $(1 + \lambda_k)(\hat{\rho} - \rho)$-strong convexity of $L_k(x)$ gives us

$$\begin{aligned} F_k(x_{k+1}) + \lambda_k G_k(x_{k+1}) \geq & F_k(\hat{x}_{k+1}) + \lambda_k G_k(\hat{x}_{k+1}) + (\hat{\zeta}_{Fk} + \lambda_k \hat{\zeta}_{Gk})^T (x_{k+1} - \hat{x}_{k+1}) \\ & + \frac{(1 + \lambda_k)(\hat{\rho} - \rho)}{2} \|x_{k+1} - \hat{x}_{k+1}\|^2. \end{aligned}$$

Since $\hat{\zeta}_{Fk} + \lambda_k \hat{\zeta}_{Gk} \in -N_X(\hat{x}_{k+1})$ by KKT conditions, and due to $x_{k+1} \in X$, we know

$$(\zeta_{Fk} + \lambda_k \zeta_{Gk})^T (x_{k+1} - \hat{x}_{k+1}) \geq 0.$$

Also since $\lambda_k G_k(\hat{x}_{k+1}) = 0$ by KKT conditions, and x_{k+1} being an (δ, τ)-optimal solution for the subproblem (2.6) yields $F_k(x_{k+1}) - F_k(\hat{x}_{k+1}) \leq \delta$ and $G_k(x_{k+1}) \leq \tau$, the the first inequality becomes

$$\begin{aligned} \delta + \lambda_k \tau \geq & \frac{(1 + \lambda_k)(\hat{\rho} - \rho)}{2} \|\hat{x}_{k+1} - x_{k+1}\|^2 \\ \|\hat{x}_{k+1} - x_{k+1}\| \leq & \sqrt{\frac{2(\delta + \lambda_k \tau)}{(1 + \lambda_k)(\hat{\rho} - \rho)}} \leq \sqrt{\frac{2(\delta + \lambda_k \tau)}{\hat{\rho} - \rho}}. \end{aligned}$$

According to Lemma 3.5, $\lambda_k \leq \frac{M + \hat{\rho} D}{\sigma} = B$, then we have

$$\|\hat{x}_{k+1} - x_{k+1}\| \leq \sqrt{\frac{2(\delta + B\tau)}{\hat{\rho} - \rho}}.$$

Before we reach the (ϵ, ϵ)-KKT point $\hat{x}_{k+1}$ for the main problem (2.5), according to the KKT conditions, there exists $\nu \in N_X(\hat{x}_{k+1})$ such that $(\hat{\zeta}_{fk} + \hat{\rho}(\hat{x}_{k+1} - x_k)) + \lambda_k(\hat{\zeta}_{gk} + \hat{\rho}(\hat{x}_{k+1} - x_k)) + \nu = 0$ and $\|\hat{\zeta}_{fk} + \lambda_k \hat{\zeta}_{gk} + \nu\| > \epsilon$, which yields $\|\hat{x}_{k+1} - x_k\| > \frac{\epsilon}{(1+\lambda_k)\hat{\rho}}$. Apply Lemma 3.5 here, we have $\|\hat{x}_{k+1} - x_k\| > \frac{\epsilon}{(1+B)\hat{\rho}}$. Thus

$$\|x_{k+1}-x_k\|^2 \geq \frac{1}{2}\|\hat{x}_{k+1}-x_k\|^2 - \|\hat{x}_{k+1}-x_{k+1}\|^2 > \frac{\epsilon^2}{2(1+B)^2\hat{\rho}^2} - \frac{2(\delta+B\tau)}{\hat{\rho}-\rho}.$$

Take $\delta = \frac{(\hat{\rho}-\rho)\epsilon^2}{8(1+B)^2\hat{\rho}^2}$ and $\tau = \frac{(\hat{\rho}-\rho)\epsilon^2}{8(1+B)^2\hat{\rho}} \min\left\{\frac{1}{\hat{\rho}-\rho+\hat{\rho}B}, 1\right\}$, and since $\tau \leq \frac{(\hat{\rho}-\rho)\epsilon^2}{8(1+B)^2(\hat{\rho}-\rho+\hat{\rho}B)\hat{\rho}}$, we have

$$\|x_{k+1}-x_k\|^2 > \frac{(\hat{\rho}-\rho)\epsilon^2}{4(1+B)^2(\hat{\rho}-\rho+\hat{\rho}B)\hat{\rho}^2}. \tag{4.12}$$

Therefore

$$\begin{aligned} g(x_{k+1}) &= G(x_{k+1}) - \frac{\hat{\rho}}{2}\|x_{k+1}-x_k\|^2 \\ &< \tau - \frac{\hat{\rho}}{2} \times \frac{(\hat{\rho}-\rho)\epsilon^2}{4(1+B)^2(\hat{\rho}-\rho+\hat{\rho}B)\hat{\rho}^2} = 0. \end{aligned}$$

This indicates that $g(x_{k+1}) \leq 0$ automatically if $g(x_k) \leq 0$, from which we can induce that all the iterates are feasible for the main problem (2.5) before reaching an (ϵ, ϵ)-KKT point.

4.4.4 Proof of Lemma 3.5

Let $\hat{x}_{k+1}$ be the exact solution for problem (2.6). $(\hat{\rho}-\rho)$-strong convexity of $G_k(x)$ implies that the set $\{x | G_k(x) \leq 0\}$ has diameter $D = \sqrt{\frac{-8g_{lb}}{\hat{\rho}-\rho}}$. x_k and $\hat{x}_{k+1}$ both lying in this set yields $\|\hat{x}_{k+1} - x_k\| \leq D$.

Let λ_k be the dual variable for problem (2.6). According to FJ conditions (1.5), there exists $\hat{\zeta}_{Fk} \in \partial F_k(\hat{x}_{k+1})$ and $\hat{\zeta}_{Gk} \in \partial G_k(\hat{x}_{k+1})$ which satisfies $\hat{\zeta}_{Fk} + \lambda_k \hat{\zeta}_{Gk} \in -N_X(\hat{x}_{k+1})$. We then focus on the case that λ_k is positive. Under this condition, $G_k(\hat{x}_{k+1}) = 0$ by KKT conditions. Then there exists $\nu \in N_X(\hat{x}_{k+1})$ such that

$$\hat{\zeta}_{Fk} + \lambda_k \hat{\zeta}_{Gk} = -\nu$$

$$\lambda_k = \frac{\|\hat{\zeta}_{Fk}\|}{\|\hat{\zeta}_{Gk} + \frac{\nu}{\lambda_k}\|}. \tag{4.13}$$

According to Assumption E, $\exists v \in -N_X^*(\hat{x}_{k+1})$ and $\|v\| = 1$, $s.t.$ $\hat{\zeta}_{Gk}^T v \leq -\sigma$. Since $\nu \in N_X(\hat{x}_{k+1})$ and $v \in -N_X^*(\hat{x}_{k+1})$, we know $\nu^T v \leq 0$. Then

$$\|\hat{\zeta}_{Gk} + \frac{\nu}{\lambda_k}\| = \|\hat{\zeta}_{Gk} + \frac{\nu}{\lambda_k}\| \cdot \|v\| \geq -(\hat{\zeta}_{Gk} + \frac{\nu}{\lambda_k})^T v \geq \sigma.$$

Let $\hat{\zeta}_{fk} = \hat{\zeta}_{Fk} - \hat{\rho}(\hat{x}_{k+1} - x_k) \in \partial f_k(\hat{x}_{k+1})$. Since we have Assumption C and $\|\hat{x}_{k+1} - x_k\| \leq D$, (4.13) becomes

$$\lambda_k = \frac{\|\hat{\zeta}_{Fk}\|}{\|\hat{\zeta}_{Gk} + \frac{\nu}{\lambda_k}\|} \leq \frac{\|\hat{\zeta}_{fk}\| + \hat{\rho}\|\hat{x}_{k+1} - x_k\|}{\sigma} \leq \frac{M + \hat{\rho}D}{\sigma} = B.$$

Thus, constant value $B = \frac{M+\hat{\rho}D}{\sigma}$ could be an uniform upper bound for the optimal dual variables λ_k of problems (2.6).

Chapter 5

Numerical Experiments

In this section, we illustrate the diversity of different approximate stationary points reached by inexact proximal point methods. We consider a sparse phase retrieval (SPR) problem in Section 1.2. Although we always saw our iterates converge as $x_k \to x^*$, but with the different levels of sparsity controlled by the SCAD constraint, we saw three distinct behaviors. When we control our problem under a high sparsity, our method is more likely to converge to a sparse solution locates on the boundary of the feasible region (with the constraint being active), where we may or may not have the strengthened constraint qualification (σ-strict MFCQ condition) satisfied. This make our iterates converge to either an approximate KKT stationary solution with the Lagrange multipliers of reasonable magnitude, or only an approximate FJ stationary solution with the Lagrange multipliers blowing up to infinity. If we relax our sparsity level and allow a lower sparsity, our method will have higher probability to converge to a solution locates in the interior of the feasible region (with the constraint being inactive and hence the optimal Lagrange multiplier equals to zero). In this case, we can reach the same accuracy of approximate FJ and KKT stationarity.

Phase retrieval is a common problem in various applications, such as imaging, X-ray crystallography and transmission electron microscopy. The phase is recovered by solving linear equations up to a universal sign change. We construct our sparse phase retrieval problem as:

$$\begin{aligned} &\min_{x \in X} \ f(x) = \frac{1}{m}\|(Ax)^2 - b^2\|_1 = \frac{1}{m}\sum_{i=1}^{m} |(a_i^T x)^2 - b_i^2| \\ &\text{s.t. } g(x) = \sum_{i=1}^{n} SCAD(x_i) - p \le 0. \end{aligned} \tag{5.1}$$

To control sparsity, $SCAD : \mathbb{R} \to \mathbb{R}$ is a SCAD function given by:

$$SCAD(u) = \begin{cases} 2|u| & 0 \le |u| \le 1, \\ -u^2 + 4|u| - 1 & 1 < |u| \le 2, \\ 3 & |u| > 2. \end{cases} \tag{5.2}$$

In this problem, $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are weakly convex and nonsmooth continuous functions, $X = \{x | x_i \in [-10, 10] \ \forall i = 1, ..., n\}$, $m = 240$, $n = 120$, $A \in \mathbb{R}^{m \times n}$, a_i^T is the i-th row of A and $b \in \mathbb{R}^m$. The value of $p \in [0, 3n)$ varies to control the sparsity of our problem. We generate each element of A as $a_{ij} \sim N(0, 1)$. For the elements x_i^* of x^*, we generate 40 of them uniformly in $[-10, -5] \bigcup [5, 10]$, and set the other 80 entries as 0. We also generate $\eta \sim N(0, I_m)$ and $b^2 = (Ax^*)^2 + \eta$. Our algorithm starts from a feasible initial point $x_0 = [0.25, ..., 0.25]^T$ with the targeted stationarity $\epsilon = 0.01$.

According to Lemma B.1 in [22], $f(x)$ is expected to be 2-weakly convex. Let a_{max} denote the maximum absolute value of the elements of A, and we set $\rho = 2a_{max} > 2$ (this should be checked in practical, since $a_{max} > 1$ with high probability), $\hat{\rho} = 2\rho$. It is also easy to derive that any subgradient of $f(x)$ should not exceed $20n^{3/2}a_{max}^2$ in X, and we set M as this value. Besides, we have $f_{lb} = 0$ and $g_{lb} = -p$ as the functional lower bounds.

An interesting fact of $g(x)$ is that, when p can be divided by 3, then there always exists subproblems that do not have Slater points. Consider $x = [3, 3, \cdots, 3, 0, 0, \cdots, 0]$ which consists of $p/3$ entries of 3 and $(n - p/3)$ entries of 0. When our iterate locates at x, we will not have the σ-strict MFCQ condition hold for our subproblem at such a point. Thus by this way our method will not be able to reach an approximate KKT point for our main problem.

In practical, it is a difficult task to know the exact solutions $\hat{x}_{k+1}$ of the subproblems, thus the way we compute stationarity and the Lagrange multipliers should be designed carefully. Since our nearly optimal and feasible solutions x_{k+1} are sufficiently close to $\hat{x}_{k+1}$, we use x_{k+1} as the substitute of $\hat{x}_{k+1}$ in our computation. For the Lagrange multipliers, we accumulate the stepsizes for solving each subproblem respectively for $G_k(x)$ and $F_k(x)$ within a huge amount of steps (controlled by measuring the amount of movements of the subiterates) and compute their ratio as an approximate evaluation. That is, when $t \to \infty$, we have

$$z^* = \lim_{t\to\infty} z_t = z_0 + \lim_{t\to\infty}\left(\sum_{t\in I}\alpha_t\zeta_{Fkt} + \sum_{t\in J}\alpha_t\zeta_{Gkt}\right).$$

Notice that z_t converges to z^*, and $\sum_{t=0}^{\infty}\alpha_t \to \infty$. Therefore

$$\begin{aligned}
&\frac{\sum_{t\in I}\alpha_t}{\sum_{t=0}^{\infty}\alpha_t}\lim_{t\to\infty}\zeta_{Fkt} + \frac{\sum_{t\in J}\alpha_t}{\sum_{t=0}^{\infty}\alpha_t}\lim_{t\to\infty}\zeta_{Gkt}\\
&=\lim_{t\to\infty}\left(\frac{\sum_{t\in I}\alpha_t\zeta_{Fkt}}{\sum_{t=0}^{\infty}\alpha_t} + \frac{\sum_{t\in J}\alpha_t\zeta_{Gkt}}{\sum_{t=0}^{\infty}\alpha_t}\right)\\
&=\lim_{t\to\infty}\frac{z^*-z_0}{\sum_{t=0}^{\infty}\alpha_t} = 0.
\end{aligned}$$

This shows that we can use $\frac{\sum_{t\in I}\alpha_t}{\sum_{t=0}^{\infty}\alpha_t}$ as an approximation of γ_{k0}, and $\frac{\sum_{t\in J}\alpha_t}{\sum_{t=0}^{\infty}\alpha_t}$ as an approximation of γ_k in the FJ conditions at the optimal point for the subproblems. We can accordingly compute the optimal Lagrange multipliers as $\lambda_k = \frac{\sum_{t\in J}\alpha_t}{\sum_{t\in I}\alpha_t}$ in the KKT conditions. Furthermore, due to the large time consumption of the potentially extremely large value of the inner step number guarantees, we stop our inner steps as long as a sufficiently tiny movement of the averaged nearly feasible subiterates is detected (i.e. we stop when $\|\bar{z}_{t+1} - \bar{z}_t\| \le \eta$ as $t \in I$, with $\eta = 10^{-8}$, and note that $\bar{z}_t = \frac{\sum_{t\in I}(t+1)F(z_t)}{\sum_{t\in I}(t+1)}$ where $I \bigcup J = \{0, 1, ..., t-1\}$). This allows us to finish our numerical experiments with reasonable time consumption.

5.1 Sparse Phase Retrieval with FJ Stationarity

In our first numerical experiment, we randomly generate our data and set $p = 120$. In this example below, we could only find an approximate FJ stationary point for the main problem, but not converge to an approximate KKT stationary point. This is because with $p = 120$ that could be divided by 3, it tends to be no Slater points in the subproblems when we are reaching the final iteration, and the σ-strong MFCQ condition fails to hold as we end our method. We could correspondingly see that the Lagrange multipliers blow up to infinity, and get the objective function value decrease without approaching a local minimum. The numerical results are shown in Figure 5-1.

Figure 5-1a shows that our iterates do not end at a local minimum point, and Figure 5-1b provides the feasibility of our iterates showing we approach the boundary of the feasible region. From Figure 5-1c we can see that our targeted FJ stationarity is attained finally, but Figure 5-1d indicates that KKT stationarity is not satisfactorily yielded. This is due to the blow up of the optimal Lagrange multipliers for our subproblems, which can be seen in Figure 5-1e and 5-1f.

5.2 Sparse Phase Retrieval with KKT Stationarity

In the second numeric, we use another set of randomly generated data and set $p = 121$. Under this setting, we would be sure to have Slater points for every subproblem no matter where our current iterate locates. This is because the subgradient set of $g(x)$ at any $x \in \{x|g(x) = 0\}$ contains the zero vector only when all the entries $x_i, i = 1, ..., n$ of x satisfy $x_i \in (-\infty, 2] \bigcup \{0\} \bigcup [2, \infty)$, $g(x)$ can be divided by 3 in this situation. Otherwise, for any $\zeta_g \in g(x)$, we can derive an lower bound for $\|\zeta_g\|$ to make the σ-strong MFCQ condition hold, which means Slater points exist. Therefore, it is guaranteed that our algorithm can converge to an approximate FJ point, which is also an approximate KKT point with a worse level of approximate stationarity if the final Lagrange multiplier is positive. We can decide the level of approximate FJ or KKT stationarity as our target, and set our parameters accordingly. Considering the case that the absolute values of 40 entries of x larger than or equal to 3 and 80 of them being 0, it is derived that $\sigma \leq \sqrt{2\hat{\rho}}$, and thus we set $\sigma = 2\sqrt{2}$. Specially, to show the difference processes of converging to the same level of approximate FJ and KKT stationary points, we set our targeted stationarity $\epsilon = 0.02$. As we have expected, when we aim to find an (ϵ, ϵ)-FJ point, with the corresponding parameters set, our method will stop earlier without reaching the same level of approximate KKT stationarity, due to the positive values of the Lagrange multipliers; When we aim to find an (ϵ, ϵ)-KKT point and set the corresponding parameters, our method will finally attain an (ϵ, ϵ)-KKT point, which consumes much more iterations than reaching the same level of approximate FJ stationarity. We could see that the Lagrange multipliers stay in a reasonable positive range without blowing up, and the objective values converge to local minimums. Let x_{lo} denote the stationary point we are converging to. The numerical results are shown in Figure 5-2 and Figure 5-3.

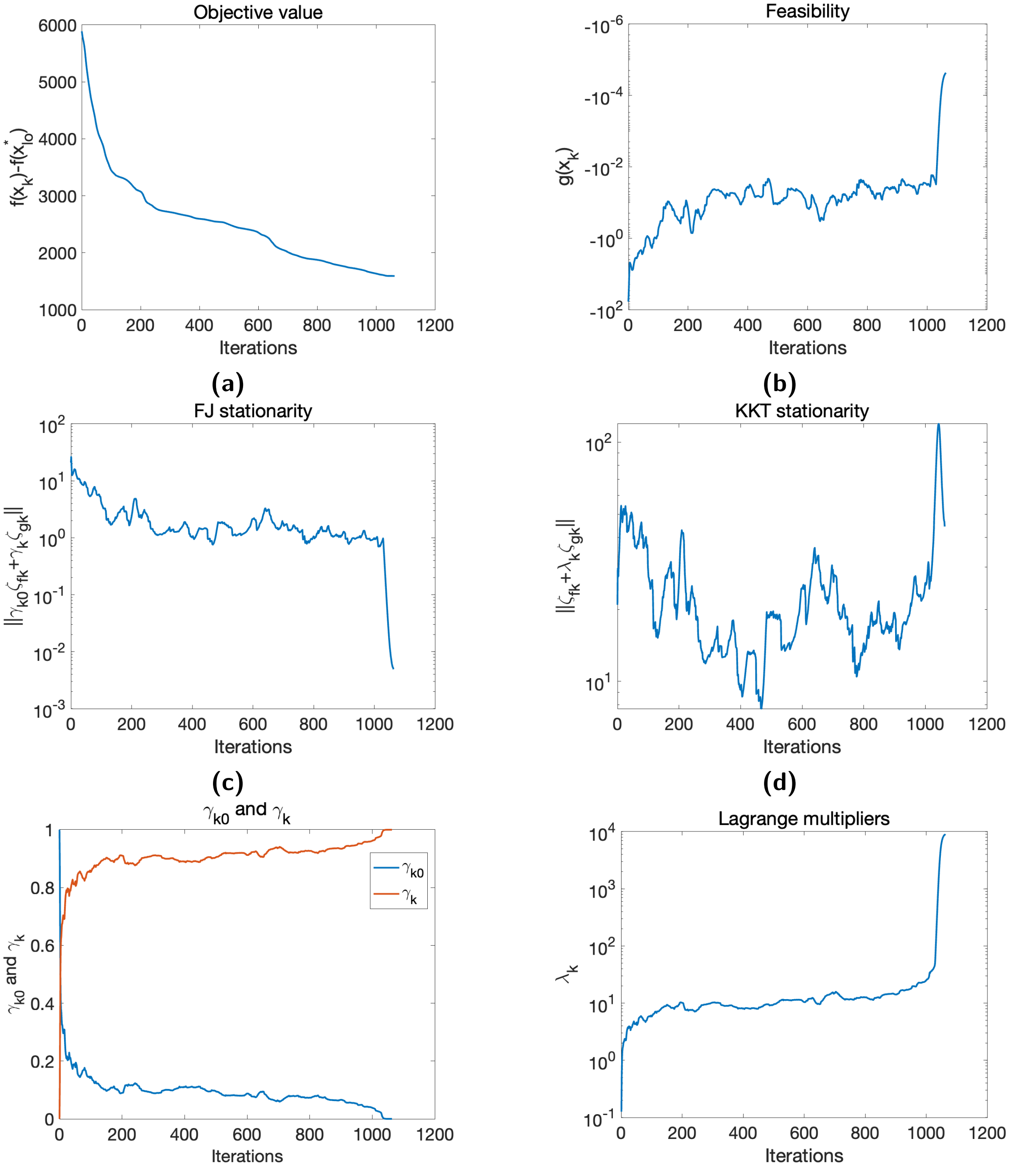
Figure 5-2 is generated as we are seeking for the targeted FJ stationarity. Figure 5-2a shows that our iterates converge to a local minimum point, and Figure 5-2b provides the feasibility of our iterates showing we approach the boundary of the feasible region. From Figure 5-2c we can see that our targeted FJ stationarity is attained finally, and Figure 5-2d shows that we also reach the corresponding level of KKT stationarity. Figure 5-2e and 5-2f reveal the varying of the Lagrange multipliers.

Figure 5-3 is generated as we are seeking for the targeted KKT stationarity. Figure 5-3a shows that our iterates converge to a local minimum point, and Figure 5-3b provides the feasibility of our iterates showing we approach the boundary of the feasible region. From Figure 5-3c we can see that our targeted KKT stationarity is attained finally. Figure 5-3d indicate that with the Lagrange multipliers larger than 1, although we have already reached our targeted FJ stationarity and the Lagrange multipliers do not blow up, the KKT stationarity we attained do not meet our goal, and our algorithm continues running until we achieve our targeted KKT stationarity.

5.3 Phase Retrieval with Inactive Stationarity

In the third numeric, we use another set of randomly generated data and set $p = 320$. Under such a relatively large value of p, we do not expect for sparse solutions anymore, and the loose restriction of the constraint results in higher possibility of attaining inactive stationarity. In this example, we would finally reach an approximate stationary solution with the constraint inactive at that point, which means its stationarity measured by FJ and KKT conditions are the same. The numerical results are shown in Figure 5-4.

Figure 5-4a shows that our iterates converge to a local minimum point, and Figure 5-4b provides the feasibility of our iterates showing we approach the boundary of the feasible region. From Figure 5-4c we can see that our targeted FJ stationarity is attained finally, and Figure 5-4d shows that our targeted KKT stationarity is also attained at the same time and the same level. Accordingly, the Lagrange multipliers converge to zero in Figure 5-4e and 5-4f.

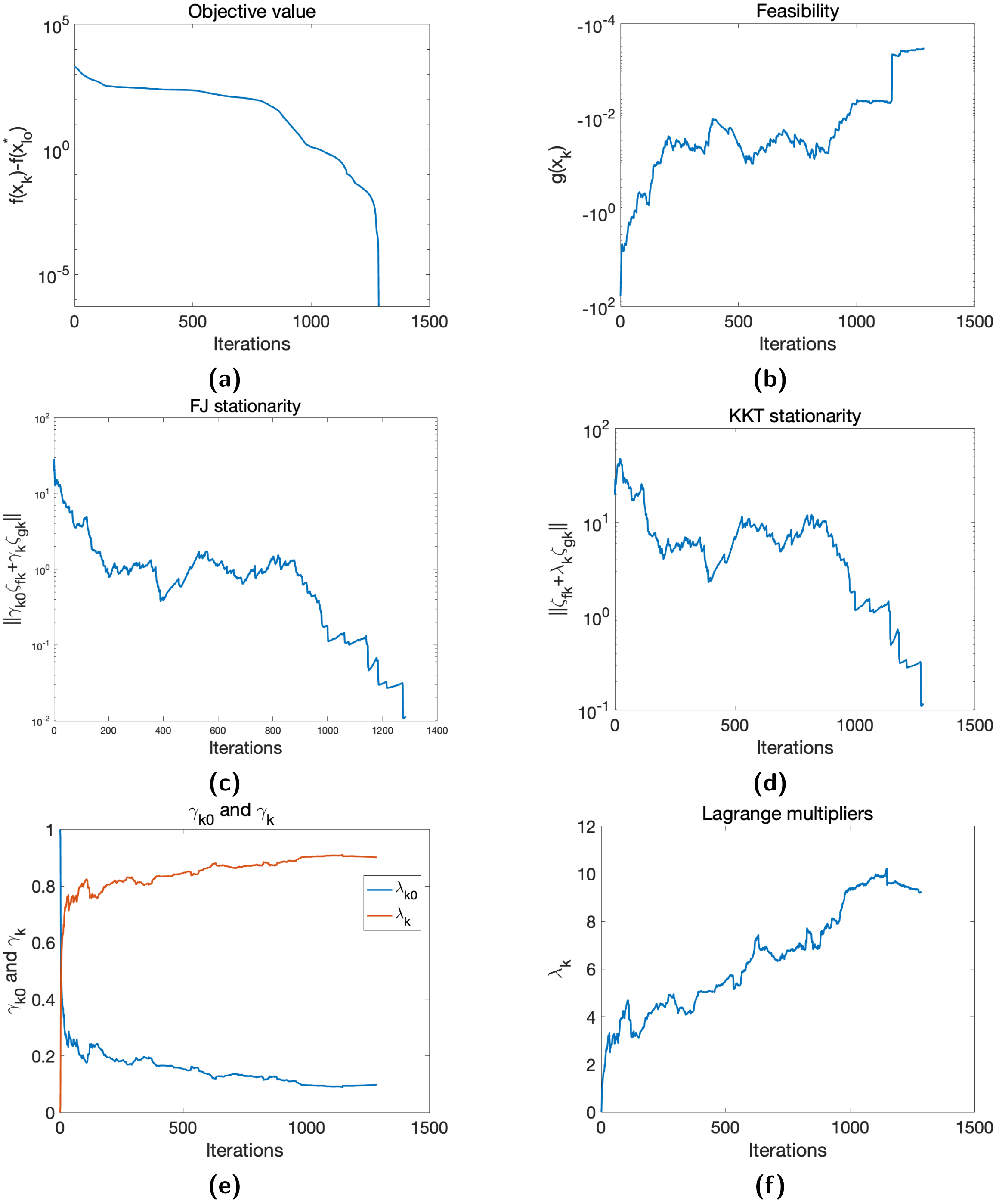


(a) **(b)**

(c) **(d)**

(e) **(f)**

Figure 5-1. Finding an (ϵ, ϵ)-FJ point for Example 1



(a)

(b)

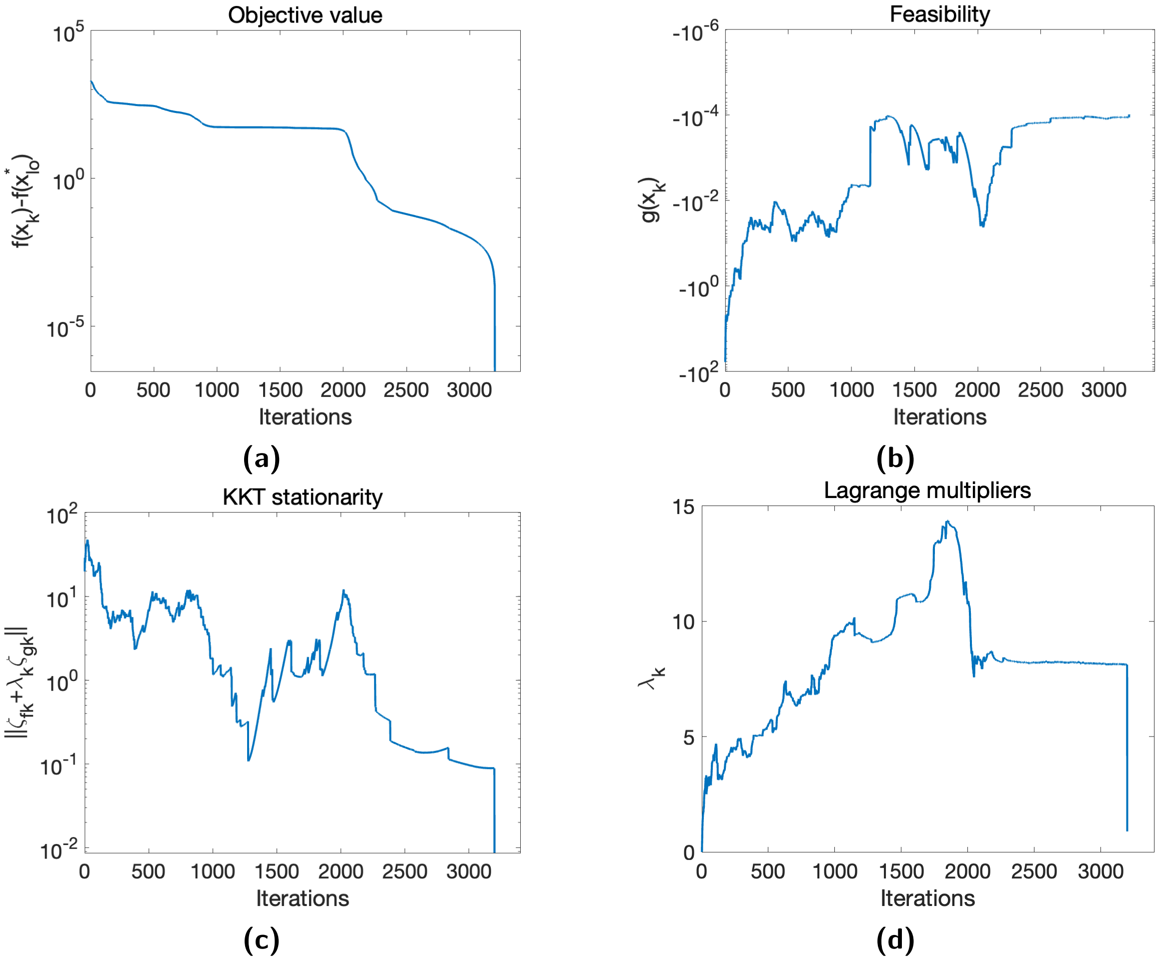
(c)

(d)

(e)

(f)

Figure 5-2. Finding an (ϵ, ϵ)-FJ point for Example 2



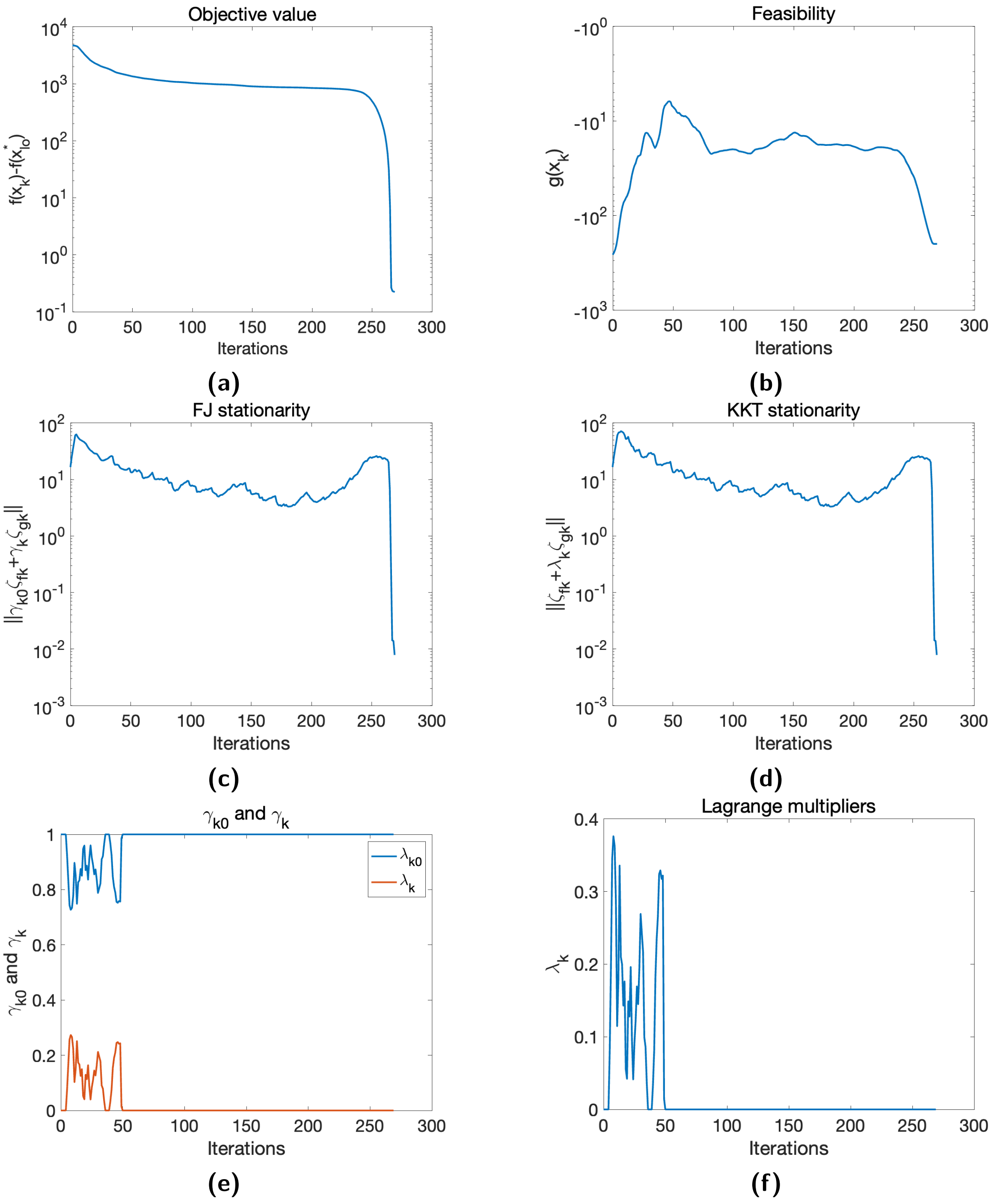
(a)

(b)

(c)

(d)

Figure 5-3. Finding an (ϵ, ϵ)-KKT point for Example 2



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5-4. Finding an (ϵ, ϵ)-FJ/KKT point for Example 3

Chapter 6

Conclusion

In this paper, we analyzed an inexact proximal point method using switching subgradient method as an oracle for nonconvex nonsmooth constrained optimization. We derived new convergence rates to attain FJ and KKT stationarity, while guaranteeing feasibility for our solutions and removing the restrictions on the compactness of domain. The performance of our method for solving sparse phase retrieval problems turns out to be consistent with our theoretical expectations.

For the future directions, it is worthy to explore more stable methods to make the Fritz John condition certificate when constraint qualification fails to be satisfied. Furthermore, stochastic versions of our method could be designed and analyzed, like [7, 22] in unconstrained setting. Finally, we may seek for guarantees of speedup for our method when we have sharpness (see [52]) at the stationary points.

References

1. Ma, R., Lin, Q. & Yang, T. *Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints* in *International Conference on Machine Learning* (2020), 6554–6564.
2. Boob, D., Deng, Q. & Lan, G. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming,* 1–65 (2022).
3. Davis, D., Drusvyatskiy, D. & Paquette, C. The nonsmooth landscape of phase retrieval. *IMA Journal of Numerical Analysis* **40,** 2652–2695 (2020).
4. Duchi, J. C. & Ruan, F. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA* **8,** 471–529 (2019).
5. Charisopoulos, V., Davis, D., Díaz, M. & Drusvyatskiy, D. Composite optimization for robust blind deconvolution. *arXiv preprint arXiv:1901.01624* (2019).
6. Chen, Y., Chi, Y. & Goldsmith, A. J. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory* **61,** 4034–4059 (2015).
7. Davis, D. & Drusvyatskiy, D. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988* (2018).
8. Davis, D. & Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization* **29,** 207–239 (2019).
9. Antoniadis, A. Wavelets in statistics: a review. *Journal of the Italian Statistical Society* **6,** 97–130 (1997).
10. Fan, J. & Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96,** 1348–1360 (2001).
11. Kim, Y., Choi, H. & Oh, H.-S. Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association* **103,** 1665–1673 (2008).
12. Fan, J., Feng, Y. & Wu, Y. Network exploration via the adaptive LASSO and SCAD penalties. *The annals of applied statistics* **3,** 521 (2009).
13. Xie, H. & Huang, J. SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics* **37,** 673–696 (2009).
14. Gasso, G., Rakotomamonjy, A. & Canu, S. Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE Transactions on Signal Processing* **57,** 4686–4698 (2009).

15. Zeng, J., Yin, W. & Zhou, D.-X. Moreau envelope augmented Lagrangian method for nonconvex optimization with linear constraints. *Journal of Scientific Computing* **91,** 1–36 (2022).

16. Weston, J. & Watkins, C. *Multi-class support vector machines* tech. rep. (Citeseer, 1998).

17. Tian, Y. & Feng, Y. Neyman-Pearson Multi-class Classification via Cost-sensitive Learning. *arXiv preprint arXiv:2111.04597* (2021).

18. Hare, W. & Sagastizábal, C. Computing proximal points of nonconvex functions. *Mathematical Programming* **116,** 221–258 (2009).

19. Hare, W. & Sagastizábal, C. A redistributed proximal bundle method for nonconvex optimization. *SIAM Journal on Optimization* **20,** 2442–2473 (2010).

20. Salzo, S. & Villa, S. Inexact and accelerated proximal point algorithms. *Journal of Convex analysis* **19,** 1167–1192 (2012).

21. Paquette, C., Lin, H., Drusvyatskiy, D., Mairal, J. & Harchaoui, Z. *Catalyst for gradient-based nonconvex optimization* in *International Conference on Artificial Intelligence and Statistics* (2018), 613–622.

22. Davis, D. & Grimmer, B. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization* **29,** 1908–1930 (2019).

23. John, F. in *Traces and emergence of nonlinear programming* 197–215 (Springer, 2014).

24. Mangasarian, O. L. & Fromovitz, S. The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. *Journal of Mathematical Analysis and applications* **17,** 37–47 (1967).

25. Allgöwer, F. & Zheng, A. *Nonlinear model predictive control* (Birkhäuser, 2012).

26. Yu, Z., Cui, P. & Crassidis, J. L. Design and optimization of navigation and guidance techniques for Mars pinpoint landing: Review and prospect. *Progress in Aerospace Sciences* **94,** 82–94 (2017).

27. Wen, F., Pei, L., Yang, Y., Yu, W. & Liu, P. Efficient and robust recovery of sparse signal and image using generalized nonconvex regularization. *IEEE Transactions on Computational Imaging* **3,** 566–579 (2017).

28. Wen, F., Chu, L., Liu, P. & Qiu, R. C. A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning. *IEEE Access* **6,** 69883–69906 (2018).

29. Zhang, H. *et al.* Meta-analysis based on nonconvex regularization. *Scientific reports* **10,** 1–16 (2020).

30. Pieper, K. & Petrosyan, A. Nonconvex regularization for sparse neural networks. *Applied and Computational Harmonic Analysis* (2022).

31. Aravkin, A. Y., Burke, J. V., Drusvyatskiy, D., Friedlander, M. P. & Roy, S. Level-set methods for convex optimization. *Mathematical Programming* **174,** 359–390 (2019).

32. Lin, Q., Nadarajah, S. & Soheili, N. A level-set method for convex optimization with a feasible solution path. *SIAM Journal on Optimization* **28,** 3290–3311 (2018).

33. Nemirovski, A. Prox-method with rate of convergence O (1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization* **15,** 229–251 (2004).

34. Hamedani, E. Y. & Aybat, N. S. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv preprint arXiv:1803.01401* **2** (2018).

35. Polyak, B. A General Method for Solving Extremum Problems. *Soviet Mathematics. Doklady* **8** (Jan. 1967).

36. Bayandina, A., Dvurechensky, P., Gasnikov, A., Stonyakin, F. & Titov, A. in *Large-scale and distributed optimization* 181–213 (Springer, 2018).

37. Lan, G. & Zhou, Z. Algorithms for stochastic optimization with expectation constraints. *arXiv preprint arXiv:1604.03887* (2016).

38. Birgin, E. G., Gardenghi, J., Martínez, J. M., Santos, S. A. & Toint, P. L. Evaluation complexity for nonlinear constrained optimization using unscaled KKT conditions and high-order models. *SIAM Journal on Optimization* **26,** 951–967 (2016).

39. Hinder, O. & Ye, Y. Worst-case iteration bounds for log barrier methods for problems with nonconvex constraints. *arXiv preprint arXiv:1807.00404* (2018).

40. Cartis, C., Gould, N. I. & Toint, P. L. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM Journal on Optimization* **21,** 1721–1739 (2011).

41. Cartis, C., Gould, N. I. & Toint, P. L. On the evaluation complexity of cubic regularization methods for potentially rank-deficient nonlinear least-squares problems and its relevance to constrained nonlinear optimization. *SIAM Journal on Optimization* **23,** 1553–1574 (2013).

42. Cartis, C., Gould, N. I. & Toint, P. L. On the complexity of finding first-order critical points in constrained nonlinear optimization. *Mathematical Programming* **144,** 93–106 (2014).

43. Cartis, C., Gould, N. & Toint, P. L. Strong evaluation complexity bounds for arbitrary-order optimization of nonconvex nonsmooth composite functions. *arXiv preprint arXiv:2001.10802* (2020).

44. Nocedal, J. & Wright, S. J. in *Numerical Optimization* 529–562 (Springer New York, New York, NY, 2006).

45. Cartis, C., Gould, N. I. M. & Toint, P. L. On the Evaluation Complexity of Constrained Nonlinear Least-Squares and General Constrained Nonlinear Optimization Using Second-Order Methods. *SIAM J. Numer. Anal.* **53,** 836–851 (Jan. 2015).

46. Facchinei, F., Kungurtsev, V., Lampariello, L. & Scutari, G. Ghost Penalties in Nonconvex Constrained Optimization: Diminishing Stepsizes and Iteration Complexity. *Mathematics of Operations Research* **46,** 595–627. eprint: https://doi.org/10.1287/moor.2020.1079 (2021).

47. Wang, X., Ma, S. & Yuan, Y.-X. Penalty methods with stochastic approximation for stochastic nonlinear programming. *Mathematics of Computation* **86,** 1793–1820 (2017).

48. Grimmer, B. *Radial Duality Part I: Foundations* 2021. arXiv: 2104.11179 [math.OC].

49. Grimmer, B. *Radial Duality Part II: Applications and Algorithms* 2021. arXiv: 2104.11185 [math.OC].

50. Grimmer, B. Convergence rates for deterministic and stochastic subgradient methods without Lipschitz continuity. *SIAM Journal on Optimization* **29,** 1350–1365 (2019).

51. Gower, R. M. *et al. SGD: General analysis and improved rates* in *International Conference on Machine Learning* (2019), 5200–5209.

52. Davis, D., Drusvyatskiy, D., MacPhee, K. J. & Paquette, C. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications* **179,** 962–982 (2018).