The Hong Kong Polytechnic University
Department of Applied Mathematics

Subgradient Methods for Convex Programming and Quasi-Convex Programming

Yaohua HU

A thesis submitted in partial fulfilment of
the requirements for the degree of Doctor of Philosophy

September 2012

CERTIFICATE OF ORIGINALITY

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

YAOHUA HU

To my family

Abstract

The purpose of this thesis is to propose some new types of subgradient methods, investigate convergence properties of the proposed algorithms, and illustrate the high efficiency and wide applicability by numerical experiments for both convex and quasi-convex optimization problems.

In the part of convex programming, we propose a primal subgradient method and a dual subgradient method, based on the gradient sampling technique, to solve a non-differentiable convex (constrained) optimization problem. The motivation comes from the fact that the gradient is cheap to compute comparing with the subgradient in many applications. The proposed algorithms consist of perturbing the projection vector to the (relative) interior of the effective domain of the objective function or the constrain set, approaching the subgradient via the convex combination of (relative) gradients at random nearby points, and proceeding the projected subgradient iteration. Using the constant/vanishing sampling radius and the constant/divergent stepsize rules, we demonstrate convergence to the (approximate) optimal value with probability 1. Numerical results demonstrate that the gradient sampling technique improves the convergence behavior of subgradient methods, and that our proposed algorithms are comparable with some existing subgradient algorithms.

In the part of quasi-convex programming, motivated by practical and theoretical reasons, we consider a generic inexact subgradient method (we call it the approximate quasi-subgradient method) to solve a nondifferentiable quasi-convex constrained optimization problem. The inexact terms stem from computation errors and noise, which come from practical considerations and applications. Assuming that the computational errors and noise are deterministic and bounded, we study the effect of the inexact terms on subgradient methods when the constraint set is compact or when the objective

function satisfies a generalized weak sharp minima condition. In both cases, using the constant/diminishing stepsize rule, we describe convergence results in both objective values and iterates, where the tolerances are given explicitly in terms of errors and noise. We also consider the finite convergence to the approximate optimal value and efficiency estimates of iterates. Several numerical experiments illustrate that the approximate quasi-subgradient method is comparable with some existing algorithm and suitable for large-scale problems. Furthermore, motivated by distributed optimization problems in networks, where both the data at each node and transmitted data are required to reach some quantization level, we propose and investigate a quantized approximate quasi-subgradient method, by using a quantization operator after proceeding the subgradient iteration along the approximate quasi-subgradient.

Acknowledgments

I would like to express my gratitude to all the people who contributed in making my three years stay at The Hong Kong Polytechnic University a fun time, full of invaluable experiences.

First of all, I wish to express my deep gratitude to my supervisors: Professor Xiaoqi Yang and Dr. Chee-Khian Sim. Thanks to them for providing me a challenging research topic and helping me to find the right research direction to follow, for teaching me how to analyze everything with patience, and for careful reading of my manuscripts and giving useful comments and suggestions on them over the last three years. I am greatly indebted to them, not only professionally, but also personally, for their generous help on both my research work and my life in Hong Kong.

A number of people have given me great help and advice. Thanks to Dr. Ka-Fai Cedric Yiu (The Hong Kong Polytechnic University) for always paying attention to my research work and for posing many challenging questions, which always improve my research work. Thanks to Professor Chong Li (Zhejiang University) for providing me many suggestions and exciting me many new ideas on my research work. Thanks to Professor Donghui Fang (Jishou University) for the stimulating discussion on how to apply the subgradient method for DC (difference of convex functions) programming. Thanks to Professor Stephen Wright (University of Wisconsin) for introducing me the stochastic subgradient method and illustrating how this method works powerfully in sparse optimization at SIAM Conference on Optimization held in Germany in 2011. Thanks to Professor Defeng Sun (National University of Singapore) for sharing me his experience in research work and for pointing some interesting trends to future research on numerical optimization when we met at two conferences held in Germany in 2011 and China in 2012 respectively.

I also wish to thank many good friends and colleagues: Bian Chuanxin, Chen Zhangyou, Deng Shirong, Gao Mingjie, Hong Yuming, Hu Sheng-long, Jin Haiyang, Li Gao, Li Zhibao, Ma Cheng, Meng Kaiwen, Pan Jie, Song Yisheng, Tian Boshi, Wang Hao, Wang Jue, Wei Yan, Xie Cong, Xia Bing, Xu Yazhong, Xu Yi, Xu Ziwei, Yang Jin, Zhang Ling, Zhang Xingfa, Zhang Yanfang, Zhou Yang and Zhu Cai for their friendship and help. They played an important role in the seminar courses and made my time at The Hong Kong Polytechnic University more enjoyable. Thanks to Chen Xi, Chi Peilian, Deng Xin, Du Hongfei, Han Ping, Huang Yani, Leng Jing, Liu Liyu, Song Tao, Xiao Bin, Yang Bin, Yang Xiao, Zhang Shengqiu and Zhang Yi, who are members of the badminton team of Chinese Students and Scholar Association of The University of Hong Kong. Their badminton training keeps me the excellent physical quality, which supports my hard work on the research field.

I sincerely thank supporting staff at Department of Applied Mathematics: Ms. Au Wah-yin (Jojo), Miss. Au Wing-chi (Anita), Mr. Cheng Tak-shing (Joseph), Miss. Chuang Fei Fei (Sylvia), Mrs. Fanny Cheung, Miss. Hung Hau-lam (Eunice), Ms. Hau Ling-ling (Cynthia), Mr. Lam Kwok-lung (Eric), Miss Law Siu-lin (Margaret), Ms. Lee Ka-man (Mansfield), Mr. Lo Charn-tong (Peter), Ms. Man Wai-lin (Ida) and Miss. Kwong Pui Kwan (Alice) for their kindness and support. I shall always treasure the wonderful research environment and the excellent research facilities of the Department of Applied Mathematics at The Hong Kong Polytechnic University.

Last but by no means least, I would like to thank my family. Mom and Dad, you gave me the whole life and have always encouraged me to keep on studying. I must offer my heartfelt thanks to my perfect wife, Qin Jing, for her love, encouragement, support and specially for her best gift to me: our new daughter Hu Yiyi Eta.

Contents

List of Notation

$\mathbb{R}$	the set of real numbers
$\mathbb{R}_+$	the set of positive real numbers
$\bar{\mathbb{R}}$	$\mathbb{R} \cup \{+\infty\}$
$\mathbb{N}$	the set of positive integers
$[a, b]$	the interval between a and b
$\lceil a \rceil$	the largest integer not greater than a
inf	infimum
min	minimization
sup	supremum
max	maximization
$\arg\min$	the set of arguments of the minimization
α^+	$\max\{0, \alpha\}$
$\exp(a)$	the exponential function e^a
$\underline{\lim}\, a_k$	limit inferior of a sequence $\{a_k\}$
$\overline{\lim}\, a_k$	limit superior of a sequence $\{a_k\}$
$\langle x, y \rangle$	inner product of two vectors $x, y \in \mathbb{R}^n$
$\|x\|$	Euclidean norm of $x \in \mathbb{R}^n$
cl$Z$	closure of Z
int$Z$	interior of Z
conv$Z$	convex hull of Z
cone$Z$	convex cone hull of Z
aff$Z$	affine space of Z
ri$Z$	interior of Z
rbd$Z$	relative boundary of Z

$B(x, \delta)$	closed ball of radius δ centered at x
B	the unit ball at the origin
$\text{dist}(x, Z)$	Euclidean distance of a vector x from a set Z
$P_Z(x)$	projection of x onto Z
$N_Z(x)$	normal cone to Z at x
$\nabla f(x)$	gradient of function f at x
$\nabla_X f(x)$	relative gradient of function f at x with respect to X
$\partial f(x)$	subdifferential of function f at x
$\partial_\epsilon f(x)$	ϵ-subdifferential of function f at x
$\partial^* f(x)$	Greenberg-Pierskalla subdifferential of function f at x
$\bar{\partial}^* f(x)$	quasi-subdifferential of function f at x
$\bar{\partial}^*_\epsilon f(x)$	ϵ-quasi-subdifferential of function f at x
$S_{f,\alpha}$	strict sublevel set of function f
$\bar{S}_{f,\alpha}$	sublevel set of function f
$U_{f,\alpha}$	strict superlevel set of function f
$\bar{U}_{f,\alpha}$	superlevel set of function f
$\text{rank}(A)$	rank of a matrix A
$\|A\|_*$	nuclear norm of A

List of Algorithms

GM	Gradient method
SGM	Classical subgradient method
IncSGM	Incremental subgradient method
DSGM	Dual subgradient method
PDSGM	Primal-dual subgradient method
IntSGM	Interior subgradient method
MDA	Mirror descent algorithm
MFA-SGM	Subgradient method based on a merit function approach
GS-SGM	Subgradient method based on gradient sampling technique
GS-SGM$^{\mathrm{S}}$	A simple version of GS-SGM
GS-DSGM	Dual subgradient method based on the gradient sampling technique
QSGM	Quasi-subgradient method
AQSGM	Approximate quasi-subgradient method
QAQSGM	Quantized approximate quasi-subgradient method

List of Tables

List of Figures

Chapter 1

Preview and Introduction

Because of the wide and growing use of optimization in science, engineering, economics, and industry, much attention has been given to the development of optimization algorithms. Knowledge of the properties and efficiency of these algorithms leads to a better understanding of their performance on various applications, and points the trend to future research on improving and extending optimization algorithms. Methods for minimizing functions with discontinuous gradients are gaining importance, and experts in computational methods of mathematical programming tend to agree that significant progress in the development of algorithms for minimizing nonsmooth functions is the key to the construction of efficient and powerful techniques for solving large-scale problems.

In nondifferentiable convex optimization, the two well-known and extensively studied methods are subgradient methods (see e.g. [52, 57, 58, 80, 82, 94, 106]) and bundle methods (see e.g. [46, 59, 55, 56, 69, 75, 101, 102]). The former relies on the knowledge of the objective function's subgradient and is not a descent method. The later uses both function values and subgradients, and usually enforces a descent property. Since the subgradient method has an extremely simple formula and requires very low storage, it appears a popular and powerful method on large-scale optimization problems. Hence, in this thesis, we will study the subgradient methods and investigate their properties and applications.

1.1 Review on Subgradient Methods for Convex Programming

Consider the following convex optimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in X, \end{aligned} \tag{1.1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, and X is a closed and convex set. Problems of type (1.1.1) are encountered in many application areas: for instance, in economics (see [90]), mechanics (see [79]), network design (see [15]), image process (see [96]), control theory (see [32]), optimal shape design (see [45]), data mining (see [22]), and machine learning (see [53]).

Subgradient methods are popular and practical techniques used to solve problem (1.1.1). Subgradient methods originated with the works of Polyak [92] and Ermoliev [38] and they were later developed by Shor [106] in the 1970s. In the last 40 years, many properties of subgradient methods have been discovered, generalizations and extensions have been proposed, and many applications have been found (see e.g. [16, 17, 46, 57, 58, 80, 82, 94, 106]).

When the objective function in (1.1.1) is continuously differentiable, the well-known gradient method was originally proposed by Goldstein [41] and Levitin and Polyak [70] in 1960s, and was then deeply developed and widely applied by many researchers (see e.g. [14, 16, 18, 21, 74, 89, 108]), which is described as follows.

Gradient method (GM)
Select a stepsize sequence $\{v_k\}$, start with an initial point $x_0 \in X$, and generate a sequence $\{x_k\} \in X$ via the iteration

$$x_{k+1} = P_X(x_k - v_k \nabla f(x_k)),$$

where P_X denotes the projection operator onto X.

The main idea of the subgradient method is to generalize the gradient method by replacing the gradient with an arbitrary subgradient. Therefore, the classical subgradient method for problem (1.1.1) is described as follows.

Classical subgradient method (SGM)

Select a stepsize sequence $\{v_k\}$, start with an initial point $x_0 \in X$, and generate a sequence $\{x_k\} \in X$ via the iteration

$$x_{k+1} = P_X(x_k - v_k g_k), \tag{1.1.2}$$

where P_X denotes the projection operator onto X and $g_k \in \partial f(x_k)$ is a subgradient of f at x_k. The subdifferential of a convex function f at x_k is defined by (see e.g. [17, 97, 98])

$$\partial f(x_k) := \{g : f(x) \geq f(x_k) + \langle g, x - x_k \rangle, \forall x \in \mathbb{R}^n\}. \tag{1.1.3}$$

An essential property of subgradient methods, which plays a key role in the convergence analysis, is the following basic inequality

$$\|x_{k+1} - x\|^2 \leq \|x_k - x\|^2 - 2v_k(f(x_k) - f(x)) + v_k^2\|g_k\|^2, \forall x \in X, \tag{1.1.4}$$

which follows from the subgradient iteration (1.1.2) and the definition of convex subdifferential (1.1.3). Using the basic inequality (1.1.4), convergence properties of subgradient methods, in both objective values and iterates, were widely studied in [16, 17, 46, 57, 58, 94, 106].

When f is differentiable at x_k, $\partial f(x_k) = \{\nabla f(x_k)\}$ and the only choice for g_k is $\nabla f(x_k)$, and thus the subgradient method reduces to the gradient method. Furthermore, in practice, it is usually considered the case when only an ϵ-subgradient can be obtained ($\epsilon > 0$), that is, g_k is allowed to only satisfy the relaxed subgradient inequality, i.e.,

$$f(x) \geq f(x_k) + \langle g_k, x - x_k \rangle - \epsilon, \forall x \in \mathbb{R}^n,$$

and then the subgradient method turns into an approximate subgradient method (also called the ϵ-subgradient method). Benefitted from practical reasons, approximate subgradient methods were widely studied in [2, 36, 46, 61, 68, 106]. Kiwiel [61] proposed a unified convergence framework for approximate subgradient methods; he presented convergence in both objective values and iterates, and gave efficiency estimates, using both the diminishing and nonvanishing stepsize rules. Larsson et al. [68] proposed and analyzed conditional ϵ-subgradient methods for solving convex constrained optimization problems and convex-concave saddle-point problems. In order to improve

conditional subgradient methods, D'Antonios and Frangioni [36] combined the deflection and the conditional subgradient technique into one iteration, and investigated the unified convergence analysis for the deflected conditional ϵ-subgradient method, using both the dynamic and diminishing stepsize rules. Applying the dual approach, Mijangos [78] studied the approximate dual subgradient method to solve constrained network flow problems. Furthermore, Auslender and Teboulle [2] proposed and developed an interior ϵ-subgradient method for convex constrained optimization problems over polyhedral sets, in particular over $\mathbb{R}^n_+$, via replacing the Euclidean distance function by a logarithmic-quadratic distance-like function.

Besides errors in ϵ-subgradient, the issue of noise on subgradient methods has been studied for convex constrained optimization problems. Regardless of the random noise, Polyak [93, 94] first studied the effect of noise, which is deterministic and bounded, on subgradient methods for convex programming. Polyak presented the convergence property of the subgradient method with noise, using both the diminishing and Shor-type (i.e., $v_k = \alpha v^k$, where $\alpha > 0$ and $0 < v < 1$) stepsize rules. A surprising conclusion is that the sequence, generated by the subgradient method with noise, exactly converges to the optimal solution when the objective function has a unique sharp minimum and satisfies a linear growth property, even if the noise is nonvanishing.

It is well-known that, in subgradient methods, the stepsize is a critical parameter that influences the convergence property and efficiency. The following types of stepsize rules are usually used and studied in the subgradient method literature (see e.g. [11, 61, 94, 106]).

(a) *Constant stepsize rule.* The stepsize v_k is fixed to a positive scalar v.

(b) *Diminishing stepsize rule.* The stepsize v_k satisfies

$$v_k > 0, \quad \lim_{k\to\infty} v_k = 0, \quad \sum_{k=0}^{\infty} v_k = +\infty. \tag{1.1.5}$$

(c) *Divergent stepsize rule.* The stepsize v_k satisfies

$$v_k > 0, \quad \sum_{i=1}^{\infty} v_k^2 < +\infty, \quad \sum_{i=1}^{\infty} v_k = +\infty. \tag{1.1.6}$$

(d) *Dynamic stepsize rule.* The stepsize v_k is given by

$$v_k = \gamma \frac{f(x_k) - f_*}{\|g_k\|^2}, \quad 0 < \gamma < 2, \tag{1.1.7}$$

where f_* denotes the optimal value of problem (1.1.1).

The constant stepsize rule is usually utilized if we are interested in quantifying the progress and efficiency estimates of the algorithm per iteration. The diminishing and divergent stepsize rules are used if the interest is in establishing the convergence property (in both objective values and iterates) of the algorithm as the number of iterations k tends to infinity. The dynamic stepsize rule, introduced by Polyak [92], is only considered when the optimal value is estimated and usually shows a better performance than other types of stepsize rules.

Nowadays, the subgradient method still remains an important tool for large-scale nonsmooth and stochastic optimization problems, due to its simple formulation and low storage requirement. In particular, due to special structures of applications, several types of subgradient methods have been proposed to efficiently solve convex optimization problems.

The incremental subgradient method

Consider the convex constrained optimization problem, where the objective function is a summation of a number of component convex functions, i.e.,

$$\begin{aligned} \min \quad & f(x) = \sum_{i=1}^{m} f_i(x) \\ \text{s.t.} \quad & x \in X. \end{aligned} \tag{1.1.8}$$

This type of optimization problems arises in the Lagrangian dual of the coupling constraints of large-scale separable optimization problems (see [14, 66]) or distributed optimization problems in large-scale networks (see [54, 95]). The iteration of the classical subgradient method for solving problem (1.1.8) is

$$x_{k+1} = P_X\big(x_k - v_k \sum_{i=1}^{m} g_{i,k}\big),$$

where $g_{i,k}$ is a subgradient of f_i at x_k.

An incremental subgradient method was proposed to solve this type of optimization problems (see [61, 80, 88, 105]). The motivation of the incremental approach originates from the incremental gradient method (the backpropagation method) for differentiable unconstrained optimization problems (see [14, 74, 76, 108]). The main idea of the incremental subgradient method is to perform each iteration as a cycle of m subiterations, which are the subgradient iterations for each component function, acting on previous subiterates. Therefore, the main improvement of the incremental subgradient method over the classical subgradient method is that at each iteration, x is changed incrementally through a sequence of m subgradient subiterations, and the incremental subgradient method for problem (1.1.8) is described as follows.

Incremental subgradient method (IncSGM)
Select a stepsize sequence $\{v_k\}$, start with an initial point $x_0 \in X$, and generate a sequence $\{x_k\} \in X$ via the following iteration. Each iteration starts with

$$\psi_{0,k} = x_k,$$

through m steps

$$\psi_{i,k} = P_X\big(\psi_{i-1,k} - v_k g_{i,k}\big), \quad g_{i,k} \in \partial f_i(\psi_{i-1,k}), \quad i = 1, \ldots, m,$$

and finally arrives at

$$x_{k+1} = \psi_{m,k}.$$

The convergence analysis of the incremental subgradient method is to view it as an approximate subgradient method and inherits that of the approximate subgradient method. Nedić and Bertsekas [80] established convergence in objective values of the incremental subgradient method, using the constant, diminishing and dynamic stepsize rules. Furthermore, Kiwiel [61] exhibited a unified convergence framework for the incremental subgradient method, including convergence in both objective values and iterates, using various stepsize rules.

The incremental subgradient method inherits the convergence behavior and powerful numerical performance from the incremental gradient method (see [14, 18]). It was illustrated in [14, 76, 108] that the incremental gradient method is highly efficient in

solving large-scale differentiable least squares problems arising in the training of neural networks, while numerical results in [80] indicated that the incremental subgradient method rapidly reaches the area near an optimal solution.

The dual subgradient method

Consider the primal problem, which is the following convex constrained optimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq 0 \\ & x \in X, \end{aligned} \tag{1.1.9}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, $g = (g_1, \cdots, g_m)^T$ with each $g_i : \mathbb{R}^n \to \mathbb{R}$ being convex, and X is a closed and convex set. By relaxing the inequality constraints $g(x) \leq 0$, the dual problem of (1.1.9) arising from the Lagrangian relaxation is given by

$$\begin{aligned} \max \quad & q(u) \\ \text{s.t.} \quad & u \in \mathbb{R}^m_+, \end{aligned} \tag{1.1.10}$$

where q is the Lagrangian dual function defined by

$$q(u) = \inf_{x \in X} \{f(x) + \langle u, g(x) \rangle\}.$$

The dual subgradient method was proposed in Shor [106] and widely studied in [67, 78, 82, 104], that is to utilize the classical subgradient method to solve the dual problem (1.1.10), and is described as follows.

Dual subgradient method (DSGM)
Select a stepsize sequence $\{v_k\}$, start with an initial point $u_0 \in \mathbb{R}^m_+$, and generate sequences $\{x_k\} \in X$ and $\{u_k\} \in \mathbb{R}^m_+$ via the iteration

$$u_{k+1} = P_{\mathbb{R}^m_+}(u_k + v_k g_k),$$

where g_k is a subgradient of the dual function q at u_k given by

$$g_k = g(x_k), \quad x_k \in \arg\min_{x \in X} \{f(x) + \langle u_k, g(x) \rangle\}. \tag{1.1.11}$$

The dual subgradient method attracts much attention due to its highly efficient performance for special structured optimization problems, which has a favorable dual structure (for instance, the dual function q is simple to evaluate), such as convex resource allocation problems in large-scale networks (see [67, 78, 82]). This is the motivation of the dual subgradient method. The convergence to a (approximate) dual optimal value of the dual subgradient method were studied in [67, 78, 104], while the progress was quantified per iteration by using the constant stepsize rule in [82].

One drawback of the dual subgradient method is that the subgradient scheme does not directly provide an optimal solution of the primal problem, that is, every cluster point of $\{x_k\}$ generated in (1.1.11) is not the primal optimal solution, even not a feasible point. To conquer this obstacle, Nemirovskii and Yudin [84] proposed using an averaging scheme to recover the primal optimal solution, i.e., the averaged vector $\hat{x}_k$ was defined by

$$\hat{x}_k = \frac{1}{k}\sum_{i=0}^{k-1} x_i, \quad \text{for all } k \geq 1.$$

The averaging scheme was widely developed in many works (see e.g. [67, 104, 82]). Sherali and Choi [104] applied an averaging scheme to recover the primal solution of linear optimization problems and extended the results to more general averaging schemes, i.e.,

$$\tilde{x}_k = \sum_{i=0}^{k} w_{i,k} x_i, \quad \text{for all } k \geq 1,$$

where $\sum_{i=0}^{k} w_{i,k} = 1$ and $w_{i,k} \geq 0$ for $i = 1, \cdots, k$. Nedić and Ozdaglar [82] utilized the averaging scheme to provide estimates on the primal feasibility and primal optimality of the averaged vector per iteration. Furthermore, Larsson et al. [67] showed that the limit of the averaged subgradient sequence satisfied the first-order optimality condition, while the original generated sequence did not satisfy the optimality condition.

The primal-dual subgradient method

Consider the following saddle point problem

$$\begin{array}{ll} \min\limits_{x}\max\limits_{\mu} & \mathcal{L}(x,\mu) \\ \text{s.t.} & x \in X, \mu \in M, \end{array} \tag{1.1.12}$$

where X is a closed and convex set in $\mathbb{R}^n$, M is a closed and convex set in $\mathbb{R}^m$, and $\mathcal{L} : X \times M \to \mathbb{R}$ is a convex-concave function. A solution of the saddle point problem (1.1.12) is a vector pair $(x^*, \mu^*) \in X \times M$ such that

$$\mathcal{L}(x^*, \mu) \leq \mathcal{L}(x^*, \mu^*) \leq \mathcal{L}(x, \mu^*), \forall x \in X, \forall \mu \in M.$$

Such a vector pair (x^*, μ^*) is also referred to as the saddle point of the function $\mathcal{L}$ over the set $X \times M$.

Saddle point problems arise in many application areas, for instance, networking applications, constrained optimization duality, zero-sum games, and general equilibrium theory. Combining the primal and dual processes together, the primal-dual subgradient method generates a sequence of primal and dual iterates converging to a saddle point of problem (1.1.12) (see [68, 83, 87, 103]). Therefore, similar to the classical subgradient method, the primal-dual subgradient method for (1.1.12) is presented as follows.

Primal-dual subgradient method (PDSGM)
Select a stepsize sequence $\{v_k\}$, start with initial points $x_0 \in X$, $\mu_0 \in M$, and generate sequences $\{x_k\} \in X$ and $\{\mu_k\} \in M$ via the iteration

$$x_{k+1} = P_X(x_k - v_k \mathcal{L}_x(x_k, \mu_k)),$$

$$\mu_{k+1} = P_M(x_k + v_k \mathcal{L}_\mu(x_k, \mu_k)),$$

where $\mathcal{L}_x(x_k, \mu_k)$ and $\mathcal{L}_\mu(x_k, \mu_k)$ denote arbitrary subgradients of $\mathcal{L}$ at (x^k, μ^k) with respect to x and μ.

Using both the diminishing and divergence stepsize rules, Larsson et al. [68], and Nesterov [87] studied convergence properties of primal-dual subgradient methods along with the averaging scheme. It is worth mentioning that, by using the primal-dual gap, the primal-dual subgradient method possesses a natural stopping criterion, which is unavailable in classical subgradient methods or dual subgradient methods.

Sen and Sherali [103] proposed a class of primal-dual subgradient methods that employed Lagrangian dual functions along with suitable penalty functions. The dual iterates were generated by the Lagrangian dual function, while the primal iterates were produced via the penalty function. Using several classical types of penalty functions, the sequence of primal and dual iterates converges to a saddle point of problem (1.1.12).

Moreover, in [103], a geometric convergence rate was established under some additional assumption.

In contrast to previous works focusing on convergence of the iterates to a saddle point, Nedić and Ozdaglar [83] utilized the constant stepsize rule and estimated the convergence rate of the generated sequence per iteration. In particular, under the bounded subgradient assumption, it was illustrated in [83, Proposition 3.1] that the function value $\mathcal{L}(\hat{x}_k, \hat{\mu}_k)$ at averaged vector $(\hat{x}_k, \hat{\mu}_k)$ converges to the optimal value at rate $1/k$ within error level vM^2, explicitly given in terms of the stepsize v and the bound on subgradients M. Furthermore, it was also illustrated in [83] that the primal-dual subgradient method can be applied in a wide range of problems, where the dual subgradient methods cannot be used (lack any special structure), because the primal-dual subgradient method can avoid difficulties associated with computing subgradients of the dual function.

Interior subgradient method

It is well-known that the classical subgradient method (1.1.2) can be rewritten as (see [2, 5, 12])

$$x_{k+1} = \arg\min_{x \in X} \Big\{ v_k \langle x, g_k \rangle + \frac{1}{2} \|x - x_k\|^2 \Big\}. \tag{1.1.13}$$

To solve the nondifferentiable convex constrained optimization problems, all these methods mentioned above are projected subgradient methods based on the Euclidean projection operator, which produces iterates that hit the boundary of the constraint set. Unfortunately, there are two main disadvantages of the Euclidean projection operator. One is that the Euclidean projection operator destroys the nice descent property and often leads to a zig-zagging phenomenon resulting in slow convergence rate. Moreover, the Euclidean projection operator itself may be computationally expensive, if the constraint set is not simple.

In order to deal with these difficulties, Auslender and Teboulle [2, 5] designed the interior subgradient method for convex optimization problems. The main idea of the interior subgradient method is to replace the Euclidean distance in (1.1.13) with a non-Euclidean distance-like function that can automatically eliminate the constraints and

also produce interior iterates. Therefore, the interior subgradient method is described as follows.

Interior subgradient method (IntSGM)
Select a stepsize sequence $\{v_k\}$, start with an initial point $x_0 \in X$, and generate a sequence $\{x_k\} \in X$ via the iteration

$$x_{k+1} = \arg\min_{x \in X} \Big\{ v_k \langle x, g_k \rangle + d(x, x_k) \Big\}, \tag{1.1.14}$$

where $d : X \times X \to \mathbb{R}_+$ is a proximal distance satisfying some mild properties (e.g. lower semi-continuity and strong convexity).

Using both the diminishing and dynamic stepsize rules, Auslender and Teboulle [2, 5] established convergence properties and provided efficiency estimates of interior subgradient methods under mild assumptions. Moreover, indicated in [5], when the proximal distance d is chosen as the logarithmic-quadratic distance-like function (see [1, 3, 4, 6]) or the Kullback-Liebler relative entropy distance (see [12, 13]), the subproblem (1.1.14), presented below, can be solved by an analytical formula and the resulting algorithms are particularly attractive.

(i) The logarithmic-quadratic distance-like function, introduced by Auslender and Teboulle in [6], is given by

$$d(x, y) = \sum_{i=1}^{n} y_i^2 \omega(x_i / y_i)$$

with

$$\omega(t) = \frac{\sigma}{2}(t-1)^2 + \mu(t - \log t - 1),$$

where $\sigma \geq \mu > 0$. It is easy to verify that $d(\cdot, y)$ is a proper, nonnegative, lower semi-continuous and convex function, and that $d(x, y) = 0$ if and only if $x = y$. Solving (1.1.14), one can obtain the following explicit formulae:

$$(x_{k+1})_i = (x_k)_i (\omega^*)'(-v_k \frac{(g_k)_i}{(x_k)_i}), i = 1, \cdots, n,$$

where $(x)_i$ denotes the i-th element of vector x, ω^* is the conjugate dual function of ω and thus

$$(\omega^*)'(t) = \frac{1}{2\sigma}\Big\{(\sigma - \mu) + t + \sqrt{((\sigma - \mu) + t)^2 + 4\mu\sigma}\Big\}.$$

(ii) The Kullback-Liebler relative entropy distance is given by

$$d(x,y)=\sum_{i=1}^{n}x_i\log\frac{x_i}{y_i}+y_i-x_i,\forall x\in\mathbb{R}^n_+,\forall y\in\mathbb{R}^n_{++}.$$

Hence, (1.1.14) can be easily solved analytically to yields the following explicit formulae:

$$(x_{k+1})_i=\frac{(x_k)_i\exp(-v_k(g_k)_i)}{\sum\limits_{j=1}^{n}(x_k)_j\exp(-v_k(g_k)_j)},i=1,\cdots,n.$$

When the Kullback-Liebler relative entropy distance is utilized in (1.1.14), the resulting interior subgradient method is actually a mirror descent algorithm. The mirror descent algorithm was introduced by Nemirovsky and Yudin [85] and could efficiently solve convex optimization problems over the unit simplex, with millions of variables (see [13]). The mirror descent algorithm is described as follows.

Mirror descent algorithm (MDA)

Select a stepsize sequence $\{v_k\}$, start with initial points $y_0\in\mathrm{dom}\nabla\psi^*$, $x_0=\nabla\psi^*(y_0)$, and generate a sequence $\{x_k\}\in X$ via the iteration

$$\begin{aligned}y_{k+1}&=\nabla\psi(x_k)-v_kg_k,\\x_{k+1}&=\nabla\psi^*(y_{k+1})\\&=\nabla\psi^*(\nabla\psi(x_k)-v_k).\end{aligned}\tag{1.1.15}$$

where $\psi:X\to\mathbb{R}$ is a given continuously differentiable and strongly convex function on int$X$ and ψ^* is its conjugate dual function.

The mirror descent algorithm (1.1.15) looks hard to understand and thus not even to mention how to improve. Beck and Teboulle [12] viewed the mirror descent algorithm as an interior subgradient method, where a Bregman-like distance is used in (1.1.14). This relationship linked these two methods and gave another aspect to understand each method.

The Bregman-like distance generated by a differentiable function ψ is defined by $B_\psi:X\times X\to\mathbb{R}$, given by

$$B_\psi(x,y)=\psi(x)-\psi(y)-\langle x-y,\nabla\psi(y)\rangle.$$

The interior subgradient method with the Bregman-like distance is given by

$$x_{k+1}=\arg\min_{x\in X}\Big\{v_k\langle x,g_k\rangle+B_\psi(x,x_k)\Big\}.\tag{1.1.16}$$

It was illustrated in [12, Proposition 3.2] that the mirror descent algorithm (1.1.15) corresponds exactly to the interior subgradient method (1.1.16). In particular, when $\psi(x) = \frac{1}{2}\|x\|^2$, $B_\psi(x, y) = \frac{1}{2}\|x - y\|^2$ and thus this method reduced to the classical subgradient method.

Subgradient method based on a merit function approach

Ruszczyński [100] developed a subgradient method with averaging for convex optimization problem (1.1.1) based on a merit function approach. In this algorithm, the author used the averaging scheme on both the subgradient approach and the successive directions' generation, and then combined these two averages into one iterative process. Therefore, the subgradient algorithm based on a merit function approach designed in [100] is presented as follows.

Subgradient method based on a merit function approach (MFA-SGM)
Select a scalar $a > 0$ and a stepsize sequence $\{v_k\} \in (0, \min(1, 1/a)]$, start with initial points $x_0 \in X$, $z_0 \in \partial f(x_0)$, and generate sequences $\{x_k\} \in X$ and $\{z_k\} \in \mathbb{R}^n$ via the iteration

$$\begin{aligned} y_k &= P_X(x_k - z_k), \\ x_{k+1} &= x_k + v_k(y_k - x_k), \\ z_{k+1} &= z_k + av_k(g_{k+1} - z_k), \end{aligned} \tag{1.1.17}$$

with $g_{k+1} \in \partial f(x_{k+1})$.

Using the technique based on a merit function approach in the space of (x, z), it was demonstrated in [100, Theorem 2.1] that the sequence of the iterates converges to an optimal solution and the sequence of the corresponding subgradients converges to a subgradient at that solution.

1.2 Review on Nonconvex Programming

Convex optimization plays a central role in many branches of applied mathematics and application areas. However, for many problems encountered in economics and engineering, the notion of convexity is too restrictive. By far, many real-life problems

cannot be described by convex mathematical models, while nonconvex functions provide a much more accurate representation of realities. During the past decades, a significant increase of research activities in nonconvex programming has been witnessed (see e.g. [9, 44, 49, 51, 72, 99]).

This limitation of convex mathematical models leads to various extensions of convex functions, which still maintain some nice properties of convex functions for characterizing optimality conditions and designing algorithms. This leads to the introduction of several generalizations of the convex function. These generalized convex functions are frequently used in various fields such as economics, engineering, management science, probability theory and various applied sciences (see e.g. [9, 35, 44]).

Quasi-convexity

A well-known and useful property of a convex function is that its sublevel sets are all convex. Many nonconvex functions also have this property and this property leads to the introduction of quasi-convex functions.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be quasi-convex if for all $x, y \in \mathbb{R}^n$ and $\alpha \in [0, 1]$ the following inequality holds

$$f((1-\alpha)x + \alpha y) \leq \max\{f(x), f(y)\},$$

or equivalently its sublevel sets

$$\{x \in \mathbb{R}^n : f(x) \leq \alpha\}, \ \forall \alpha \in \mathbb{R}$$

are convex.

The quasi-convex function is widely used in many application areas, such as economics, engineering, and geometric optimization (see e.g. [9, 35, 44, 109]). Due to the wide applications of quasi-convex programming, many properties and characteristics of quasi-convex functions have been studied. Greenberg and Pierskalla [42] studied the useful properties of quasi-convex functions and characterized the relationship between convex functions and quasi-convex functions. Crouzeix [33] gave a sufficient and necessary condition for a quasi-convex function to be convex. Luenberger [73] presented the existence of a multiplier and duality theory of quasi-convex programming.

Another interesting issue is the subdifferential of the quasi-convex function. Several notions of subdifferentials of quasi-convex functions have been proposed, such as Clarke-Rockafellar subdifferential, Dini subdifferential, Fréchet subdifferential (see [7] and references therein). The earliest one is the Greenberg-Pierskalla subdifferential proposed in [43]. The Greenberg-Pierskalla subdifferential is defined by the quasi-conjugate function based on the quasi-convexity structure. Since the conjugate relates to the support of epigraph, the quasi-conjugate relates to the support of sublevel set. Thus, the Greenberg-Pierskalla subdifferential, defined by quasi-conjugate functions, is actually supports of its sublevel set, i.e.,

$$\partial^* f(x) = \{g : \langle g, y - x\rangle < 0, \forall y \in S_f(x)\}, \tag{1.2.1}$$

where $S_f(x)$ is the strict sublevel set, given by $S_f(x) = \{y \in \mathbb{R}^n : f(y) < f(x)\}$.

To meet much more applications, Martínez-Legaz and Sach [77] introduced the Q-subdifferential. Given that the Q-subdifferential is a subset of the Greenberg-Pierskalla subdifferential, it shares with all other quasi-convex subdifferentials the property that its nonemptiness on the domain of a lower semi-continuous function implies the quasi-convexity of the function, which justifies the claim that it is a quasi-convex subdifferential.

Another similar notion is the normal cone to its strict sublevel set, defined by

$$N_f(x) = \{x^* : \langle x^*, y - x\rangle \leq 0, \forall y \in S_f(x)\}. \tag{1.2.2}$$

Borde and Crouzeix [20] described two important properties of the normal cone to the strict sublevel set of quasi-convex functions. The first is that $N_f(x)$ is an C-upper semi-continuous point-to-set mapping when f is a quasi-convex function. The second is that $N_f(x)$ can be expressed as the convex hull of the limits of $\{N_f(x_n)\}$, where $\{x_n\}$ is a sequence converging to x and contained in a dense set. Aussel and Daniilidis [8] slightly modified the definition of normal operator and studied the corresponding quasi-monotonicity, which actually characterizes the class of continuous and quasi-convex functions.

Subgradient method for quasi-convex programming

To meet much boarder class of problems, Kiwiel [60] proposed a quasi-subgradient method for quasi-convex optimization problems. The main idea of the quasi-subgradient method is similar to that of subgradient methods: proceed the subgradient iteration along the direction of the Greenberg-Pierskalla subgradient (see (1.2.1)) or normal vector (see (1.2.2)), and then project the resulting point onto the constraint set. Thus, the quasi-subgradient method is described as follows.

Quasi-subgradient method (QSGM)
Select a stepsize sequence $\{v_k\}$, start with an initial point $x_0 \in X$, and generate a sequence $\{x_k\} \in X$ via the iteration

$$x_{k+1} = P_X(x_k - v_k \hat{g}_k), \quad \hat{g}_k = g_k / \|g_k\|,$$

where $g_k \in \partial^* f(x_k)$ (or $N_f(x_k)$) is a quasi-subgradient of f at x_k.

Assuming the objective function is quasi-convex and upper semi-continuous, and using the diminishing stepsize rule, Kiwiel presented the convergence property in objective values in [60, Theorem 1]. In addition, under the assumption that the optimal solution set has a nonempty interior, the author also arrived at the finite convergence property. Furthermore, a surprising result is the convergence property in iterates described in [60, Theorem 4].

Modified dual subgradient method via sharp augmented Lagrangian

Subgradient method for nonconvex programming was also studied by virtue of the sharp augmented Lagrangian (see [23, 24, 40]). The following nondifferentiable and nonconvex optimization problem with equality constraints is considered, i.e.,

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) = 0 \\ & x \in X, \end{aligned} \tag{1.2.3}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are continuous functions, and X is a closed and convex set. The sharp augmented Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_+ \to \mathbb{R}$ associated with

problem (1.2.3) is defined by

$$L(x, u, c) = f(x) - \langle u, g(x)\rangle + c\|g(x)\|,$$

and the dual function $q : \mathbb{R}^m \times \mathbb{R}_+ \to \mathbb{R}$ is defined by (see [98, 99])

$$q(u, c) = \min_{x\in X} f(x) - \langle u, g(x)\rangle + c\|g(x)\|.$$

Hence, the dual problem arising from the sharp augmented Lagrangian relaxation is given by

$$\begin{array}{ll} \max & q(u, c) \\ \text{s.t.} & (u, c) \in \mathbb{R}^m_+ \times \mathbb{R}_+. \end{array} \tag{1.2.4}$$

Gasimov [40] and Burachik et al. [23, 24] proposed and developed an exact/inexact modified dual subgradient method via sharp augmented Lagrangian to solve nonconvex optimization problem (1.2.3). The modified dual subgradient method via sharp augmented Lagrangian is to utilize the classical subgradient method to solve the dual problem (1.2.4), and is described as follows.

Modified dual subgradient method via sharp augmented Lagrangian (SALMDSGM)

Select stepsize sequences $\{v_k\}$ and $\{s_k\}$, start with an initial point $(u_0, c_0) \in \mathbb{R}^m \times \mathbb{R}_+$, and generate sequences $\{x_k\} \in X$, $\{u_k\} \in \mathbb{R}^m$ and $\{c_k\} \in \mathbb{R}_+$ via the iteration

$$\begin{aligned} u_{k+1} &= u_k - v_k f(x_k), \\ c_{k+1} &= c_k + (v_k + s_k)\|f(x_k)\|, \end{aligned}$$

where the vector $(f(x_k), \|f(x_k)\|)$ is a subgradient of q at (u_k, c_k), and x_k is defined by

$$x_k \in \arg\min_{x\in X}\{f(x) + c\|g(x)\| - \langle u, g(x)\rangle\}.$$

The authors not only established primal and dual convergence results, but also generated a strictly increasing sequence of dual function values (see [40, Theorem 7] and [24, Proposition 3.1]). This monotone property is impossible in most versions of subgradient methods.

1.3 Notations and Preliminaries

The notation used in the thesis is standard. In particular, we consider the n-dimensional Euclidean space $\mathbb{R}^n$. We view a vector as a column vector, and denote by $\langle x, y\rangle$ the inner product of two vectors $x, y \in \mathbb{R}^n$. We use $\|x\|$ to denote the standard Euclidean norm, $\|x\| = \sqrt{\langle x, x\rangle}$. For a set Z in $\mathbb{R}^n$, we denote the closure (resp. interior, convex hull, convex cone hull, affine space, relative interior, relative boundary) of Z by cl$Z$ (resp. int$Z$, conv$Z$, cone$Z$, aff$Z$, ri$Z$, rbd$Z$). For each $x \in \mathbb{R}^n$ and $\delta \in \mathbb{R}_+$, $B(x, \delta)$ denotes the closed ball of radius δ centered at x, and specially B denotes the unit ball at the origin.

For a point x and a set Z, the Euclidean distance $\mathrm{dist}(x, Z)$ of x from Z and the projection $P_Z(x)$ of x onto Z are respectively defined by

$$\mathrm{dist}(x, Z) := \inf_{z \in Z} \|x - z\|,$$

and

$$P_Z(x) := \{z \in Z : \|x - z\| = \mathrm{dist}(x, Z)\} = \arg\min_{z \in Z} \|x - z\|.$$

The well-known nonexpansive property of the projection operator is described as follows.

Lemma 1.3.1 ([17, Proposition 2.2.1]) *Let C be a nonempty, closed and convex subset of $\mathbb{R}^n$. Then the projection operator $P_C : \mathbb{R}^n \to C$ is continuous and nonexpansive, i.e.,*

$$\|P_C(x) - P_C(y)\| \le \|x - y\|, \forall x, y \in C.$$

The normal cone to a convex set Z at x is defined by

$$N_Z(x) := \{\nu \in \mathbb{R}^n : \langle \nu, z - x\rangle \le 0, \forall z \in Z\}.$$

The indicator function δ_Z of Z is defined by

$$\delta_Z(x) := \begin{cases} 0, & x \in Z, \\ +\infty, & \text{otherwise.} \end{cases}$$

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

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

We call f a proper function if $f(x) < +\infty$ for at least one $x \in \mathbb{R}^n$, or in other words, if $\mathrm{dom} f$ is a nonempty set.

f is said to be convex if for all $x, y \in \mathbb{R}^n$ and $\alpha \in [0, 1]$ the following inequality holds

$$f((1-\alpha)x + \alpha y) \leq (1-\alpha)f(x) + \alpha f(y).$$

f is said to be quasi-convex if for all $x, y \in \mathbb{R}^n$ and $\alpha \in [0, 1]$ the following inequality holds

$$f((1-\alpha)x + \alpha y) \leq \max\{f(x), f(y)\}.$$

f is said to be upper semi-continuous on $\mathbb{R}^n$ if $f(x) = \overline{\lim_{y \to x}} f(y)$ for all $x \in \mathbb{R}^n$. It is said to be lower semi-continuous on $\mathbb{R}^n$ if $f(x) = \underline{\lim_{y \to x}} f(y)$ for all $x \in \mathbb{R}^n$. It is said to be continuous on $\mathbb{R}^n$ if f is both lower semi-continuous and upper semi-continuous on $\mathbb{R}^n$. In particular, f is said to be Lipschitz continuous of rank L over X if

$$|f(x) - f(y)| \leq L\|x - y\|, \forall x, y \in X.$$

Given $\epsilon \geq 0$, the subdifferential and ϵ-subdifferential of a convex function f at $x \in \mathbb{R}^n$ are defined respectively by

$$\partial f(x) := \{g : f(y) \geq f(x) + \langle g, y - x\rangle, \forall y \in \mathbb{R}^n\}, \tag{1.3.1}$$

and

$$\partial_\epsilon f(x) := \{g : f(y) \geq f(x) + \langle g, y - x\rangle - \epsilon, \forall y \in \mathbb{R}^n\}. \tag{1.3.2}$$

For each $\alpha \in \mathbb{R}$, we denote the (strict) sublevel sets and (strict) superlevel sets of f respectively by

$$\begin{aligned} S_{f,\alpha} &:= \{x \in \mathbb{R}^n : f(x) < \alpha\}, \quad S_f(x) := S_{f,f(x)}, \\ \bar{S}_{f,\alpha} &:= \{x \in \mathbb{R}^n : f(x) \leq \alpha\}, \quad \bar{S}_f(x) := \bar{S}_{f,f(x)}, \\ U_{f,\alpha} &:= \{x \in \mathbb{R}^n : f(x) > \alpha\}, \quad U_f(x) := U_{f,f(x)}, \\ \bar{U}_{f,\alpha} &:= \{x \in \mathbb{R}^n : f(x) \geq \alpha\}, \quad \bar{U}_f(x) := \bar{U}_{f,f(x)}. \end{aligned} \tag{1.3.3}$$

It is well-known that f is quasi-convex if and only if $S_{f,\alpha}$ $\left(\bar{S}_{f,\alpha}\right)$ is convex for all $\alpha \in \mathbb{R}$ (see [44, Lemma 1.27]), and that f is upper semi-continuous on $\mathbb{R}^n$ if and only if $S_{f,\alpha}$ is open for all $\alpha \in \mathbb{R}$ (see [47, Proposition 1.2.2]).

Given a square matrix A, if for some non-zero vector x and number λ, there holds

$$Ax = \lambda x,$$

then the vector x is called an eigenvector of A and the number λ is called the eigenvalue corresponding to that vector. More general, for a matrix $A \in \mathbb{R}^{m\times n}$, the singular values of A are defined by the square roots of the eigenvalues of the matrix A^*A, where A^* denotes the adjoint matrix of A.

The singular value decomposition (in short, SVD) is a factorization of a matrix, with many useful applications in signal processing and statistics. Formally, the singular value decomposition of a matrix $A \in \mathbb{R}^{m\times n}$ is a factorization of the form (see [48, Theorem 7.3.5])

$$A = U\Sigma V^T,$$

where U is an $m \times r$ unitary matrix, Σ is an $r \times r$ diagonal matrix of its singular values, and V^T (the transpose of V) is an $n \times r$ unitary matrix.

The nuclear norm of the matrix A is defined by the sum of its singular values, i.e.,

$$\|A\|_* = \sum_{i=1}^{r} \Sigma_{ii},$$

where Σ_{ii} denotes the i-th diagonal element of Σ. The nuclear norm is a convex function and its subdifferential at Z is given by (see e.g. [71, 96]),

$$\partial\|Z\|_* = \{UV^T + W : W \text{ and } Z \text{ have orthogonal row/column spaces and } \|W\| \leq 1\}.$$

When Z has no zero singular value (Z is full rank), the nuclear norm is differentiable and $\nabla\|Z\|_* = UV^T$.

We end this section by recalling the following well-known Separation Theorems and two well-known properties on real sequences, which are repeatedly used in this thesis, so as to make the thesis more self-contained.

Lemma 1.3.2 ([17, Proposition 2.4.5]) *Let C be a nonempty and convex subset of $\mathbb{R}^n$ and let x be a vector in $\mathbb{R}^n$. Then $x \notin \mathrm{ri}C$ if and only if there exists a hyperplane that properly separates C and x, that is, there exists a vector $a \neq 0$ such that*

$$\sup_{y\in C}\langle a, y\rangle \leq \langle a, x\rangle \quad \text{and} \quad \inf_{y\in C}\langle a, y\rangle < \langle a, x\rangle.$$

Lemma 1.3.3 ([17, Proposition 2.4.5]) *Let C_1 and C_2 be two nonempty and convex subsets of $\mathbb{R}^n$. Then $\mathrm{ri}C_1 \cap \mathrm{ri}C_2 = \emptyset$ if and only if there exists a hyperplane that properly separates C_1 and C_2, that is, there exists a vector $a \neq 0$ such that*

$$\sup_{x_1\in C_1}\langle a, x_1\rangle \leq \inf_{x_2\in C_2}\langle a, x_2\rangle \quad \text{and} \quad \inf_{x_1\in C_1}\langle a, x_1\rangle < \sup_{x_2\in C_2}\langle a, x_2\rangle.$$

Lemma 1.3.4 ([97, Theorem 11.7]) *Let C_1 and C_2 be two nonempty and convex subsets of $\mathbb{R}^n$, in additional, C_2 is a cone. If $\mathrm{ri}C_1 \cap \mathrm{ri}C_2 = \emptyset$, then there exists a hyperplane that properly separates C_1 and C_2 and passes through the origin, that is, there exists a vector $a \neq 0$ such that*

$$\langle a, x_1\rangle < 0, \forall x_1 \in \mathrm{ri}C_1, \quad \text{and } \langle a, x_2\rangle \geq 0, \forall x_2 \in \mathrm{cl}C_2.$$

Lemma 1.3.5 ([61, Lemma 2.1]) *Suppose $\{t_k\} \subset \mathbb{R}_+$ and $t_k^{sum} := \sum_{i=1}^{k} t_i \to \infty$ as $k \to \infty$. Given a scalar sequence $\{a_k\}$ and let the averaged sequence $\hat{a}_k := \big(\sum_{i=1}^{k} t_i a_i\big)/t_k^{sum}$ for all $k \in \mathbb{N}$. Then $\varliminf_{k\to\infty} a_k \leq \varliminf_{k\to\infty} \hat{a}_k \leq \varlimsup_{k\to\infty} \hat{a}_k \leq \varlimsup_{k\to\infty} a_k$. In particular, if $\lim_{k\to\infty} a_k = a$, then $\lim_{k\to\infty} \hat{a}_k = a$.*

Lemma 1.3.6 ([94, Lemma 2.2.2]) *Suppose $\{a_k\}$, $\{b_k\}$, and $\{c_k\}$ are positive scalar sequences satisfying*

$$a_{k+1} \leq a_k(1 + b_k) + c_k, \forall k \in \mathbb{N},$$

and

$$\sum_{k=1}^{\infty} b_k < \infty, \quad \sum_{k=1}^{\infty} c_k < \infty.$$

Then $\{a_k\}$ converges to some $a \in \mathbb{R}_+$.

1.4 Motivation and Outline of the Thesis

Due to its simple formula and low storage requirement, the subgradient method appears a popular and powerful method on large-scale nonsmooth and stochastic optimization problems.

In Part I, we propose highly efficient subgradient algorithms, based on gradient sampling technique, and analyze their convergence properties for convex programming.

Recently, a gradient sampling technique is popular in designing algorithms for optimization problems (see [29, 30, 31, 62, 63]). The gradient sampling technique was first presented by Burke et al. [30] and then used to design the steepest descent gradient sampling algorithm to minimize a locally Lipschitz function in [31].

In Chapter 2, we consider the nondifferentiable convex optimization problem with an extended real-valued objective function. Motivated by the philosophy, given by Burke et al. in [29, 30, 31], that the gradient is cheap to compute comparing with the subgradient, we incorporate the gradient sampling technique into the subgradient method, that is to construct the subgradient information via random sampling of relative gradients at nearby points. We demonstrate that the sequence, generated by the proposed algorithm, converges to the optimal value within some explicit tolerance when the constant stepsize rule is used, and converges to an optimal solution by using the divergent stepsize rule, with probability 1. In particular, in Section 2.5, we focus on the convex constrained optimization problem, and the algorithm reduces to a simple version, which skips the perturbation step and uses the gradient information instead of the relative gradient. In Section 2.6, we illustrate our proposed algorithms respectively on three examples. Our numerical experiments show that the gradient sampling procedure does not cost much time in the whole algorithm and our proposed algorithms always converge faster than existing subgradient algorithms. Especially for the low-rank recovery problem (i.e., the nuclear norm minimization problem), our algorithm only costs one third or half of time that is required for the classical subgradient method. This result might draw attention of the researchers to the application of subgradient methods on the low-rank recovery problem.

In Chapter 3, motivated by the convex constrained problem which has a favorable d-

ual structure, we incorporate the gradient sampling technique into the dual subgradient method and investigate its convergence propertiy. We indicate that this algorithm in the dual approach converges to the dual optimal value with probability 1. We also use an averaging scheme to recover approximate primal optimal solutions, in the presence of the Slater condition.

In Part II, we present inexact subgradient algorithms and investigate their properties for quasi-convex programming.

Motivated by practical reasons, the inexact algorithms is widely applied in solving optimization problems. The ϵ-subgradient method was widely studied in [2, 36, 46, 61, 68, 106]. Besides errors in ϵ-subgradient, Polyak [93, 94] proposed and studied the effect of noise, which is deterministic and bounded, on subgradient methods for convex programming.

In Chapter 4, we focus on an inexact subgradient algorithm for the quasi-convex optimization problem. Inspired by the idea in [61, 81, 93, 94] and references therein, we investigate the effect of inexact terms, including both computation errors and noise, on the inexact subgradient algorithm. We establish convergence properties in both objective values and iterates with the tolerance given explicitly in terms of the errors and noise. We also give finite convergence to approximate optimal solutions and efficiency estimates of iterates. Our investigation is divided into two cases: (i) X is compact and (ii) X is noncompact. When X is compact, we assume the Hölder condition of order $p > 0$, instead of the upper semi-continuity of the objective function used in [60]. When X is noncompact, we need to assume an additional generalized weak sharp minima condition. This condition extends the concept of weak sharp minima in [81] and is presented by using $\text{dist}(x, X^*)$, the distance of the decision variable x to X^*. We also illustrate the proposed algorithm on several numerical experiments in Section 4.6. Our numerical experiments show that the exact quasi-subgradient method arrives at a better solution in fewer iterates comparing with level function method proposed by Xu [109] on small-scale problems, and that inexact subgradient method is suitable for large-scale problems. We also indicate the sensitivity of inexact terms, which coincides with the convergence analysis proposed in Section 4.3.

In Chapter 5, motivated by the distributed optimization problem in networks, where

the data at each node and transmitted data required to reach some quantized level, we present the quantized approximate quasi-subgradient method, and investigate the influence of inexact items and convergence behavior.

Based on Chapter 2 and Chapter 4 respectively, the following papers have been written and submitted:

1. Y. H. Hu, C. K. Sim and X. Q. Yang. Subgradient methods for convex optimization problems based on gradient sampling. Submitted.

2. Y. H. Hu, X. Q. Yang and C. K. Sim. Inexact subgradient methods for quasi-convex optimization problems. Submitted.

Part I

Subgradient Methods for Convex Programming

Chapter 2

Subgradient Method Based on Gradient Sampling Technique

2.1 Introduction

Subgradient methods are popular and practical techniques used to minimize the nondifferentiable convex function. Subgradient methods originated with the works of Polyak [92] and Ermoliev [38] and were later developed by Shor [106] in the 1970s. In the last 40 years, many properties of subgradient methods have been discovered, generalizations and extensions have been proposed, and many applications have been found (see [16, 17, 46, 57, 58, 80, 82, 94, 106] and references therein). Nowadays, the subgradient method still remains an important tool for large-scale nonsmooth optimization and stochastic optimization problems, due to its simple formulation and low storage requirement.

Recently, a gradient sampling technique is popular in designing algorithms for optimization problems (see [29, 30, 31, 62, 63]). The gradient sampling (in short, GS) technique was first presented by Burke et al. [30] and used to solve typical matrix optimization problems. Extending their previous works, in [31], the authors designed the steepest descent GS algorithm to minimize a locally Lipschitz function. The steepest descent GS algorithm calculates the gradients of the objective function at random nearby points, uses this information to construct an approximate steepest descent direction,

and then produces the next iterate via an Armijo line search along this direction. They also demonstrated that the sequence generated by the steepest descent GS algorithm converges to a Clarke approximate stationary point when a fixed sampling radius was utilized (see [31, Theorem 3.4]), and converges to a stationary point when the sampling radius is reduced dynamically (see [31, Theorem 3.8]), with probability 1.

In this chapter, we consider the nondifferentiable convex optimization problem with an extended real-valued objective function, i.e.,

$$\min_{x\in\mathbb{R}^n} \quad f(x), \tag{2.1.1}$$

where $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is a proper, lower semi-continuous and convex function. Let $X = \mathrm{dom} f$, which may have an empty interior, and denote the optimal solution set and the optimal value of (2.1.1) respectively by X^* and f_*. Due to the lower semi-continuity and convexity of f, X is a closed and convex set.

Motivated by the philosophy, given by Burke et al. in [29, 31], that the gradient is cheap to compute comparing with the subgradient, we incorporate the GS technique into the subgradient method for problem (2.1.1). The main idea of the subgradient method based on the GS technique (in short, GS-SGM) is to construct a subgradient via a convex combination of relative gradients at random sampling points in a certain neighborhood of the current iterate. The use of relative gradients is due to the fact that the domain X dose not necessarily have a nonempty interior. Thus, the random sampling procedure cannot be carried out on the whole space $\mathbb{R}^n$, but on the affine space spanned by the domain X. Furthermore, as each iterate is not necessarily a relative interior point of the domain (might be on the relative boundary of X), we perform a perturbation step to shift the iterate to a relative interior point to ensure that the random sampling procedure can be carried out.

In the convergence analysis, we demonstrate that the sequence generated by the GS-SGM converges to the optimal value within the tolerance $vM^2/2$, where M is an upper bound on relative gradients of the objective function, when a constant stepsize v is used (see Theorem 2.3.1), and converges to an optimal solution when the divergent stepsize rule is used (see Theorem 2.3.2), with probability 1. It is worth mentioning that, without regard to the stochastic process and perturbation step, the GS-SGM is theoretically a version of the approximate subgradient method, e.g. in [61].

In Section 2.5, we propose a simple version of GS-SGM (in short, GS-SGM$^{\mathrm{S}}$) to solve the convex constrained optimization problem. The GS-SGM$^{\mathrm{S}}$ constructs the subgradient information via random sampling of gradients at nearby points and projects the point, obtained from the subgradient iteration along the constructed direction, onto the constraint set. Since the domain of the objective function is the whole space, all points are interior points of $\mathbb{R}^n$ and the perturbation step is then skipped. We also demonstrate convergence to an approximate optimal value when a fixed sampling radius or the constant stepsize rule is utilized (see Theorems 2.5.1-2.5.3), and indicate convergence to the optimal value when the sampling radius is reduced dynamically and the divergent stepsize rule is used (see Theorems 2.5.4), with probability 1.

We illustrate the GS-SGM/GS-SGM$^{\mathrm{S}}$ on three examples, including the nonsmooth convex optimization problem (see [100]), the assignment problem (see [80]) and the low-rank recovery problem (i.e., the nuclear norm minimization problem, see [96]). The numerical experiments show that the GS procedure does not cost much time in the whole algorithm and the GS-SGM/GS-SGM$^{\mathrm{S}}$ always converges faster than existing subgradient methods. Especially for the low-rank recovery problem, the GS-SGM$^{\mathrm{S}}$ only costs one third or half of the time that is required for the classical subgradient method. This result might draw attention of the researchers to the application of subgradient methods on the low-rank recovery problem.

This chapter is organized as follows. In Section 2.2, we present the subgradient method based on the GS technique by using a perturbation direction of the projection point. In Section 2.3, we demonstrate convergence properties of the GS-SGM. In Section 2.4, we calculate the perturbation direction of the projection point in two common cases of the domain. In Section 2.5, we focus on the convex constrained optimization problem and propose a simple version of GS-SGM to solve it. We also establish some corresponding convergence results as that of the GS-SGM. Finally, we exhibit several numerical results of the GS-SGM/GS-SGM$^{\mathrm{S}}$ in Section 2.6.

2.2 GS-SGM Algorithm

In this section, we propose a subgradient method based on the gradient sampling technique (GS-SGM) to solve problem (2.1.1). The main idea of the GS-SGM is to construct the subgradient information via the GS technique and then proceed the iteration of the classical subgradient method. No subgradient information is required in the GS-SGM algorithm. When $\text{int}(X) = \varnothing$, the domain of f has an empty interior and thus the gradient of f cannot be defined. Thus, we use the relative gradient instead. In the following, we introduce the definition of the relative gradient (see Hiriart-Urruty and Lemaréchal [47]).

Definition 2.2.1 (see [47]) *$f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to be relatively differentiable at $\bar{x}$, if there exists a vector $g \in V$, the subspace parallel to* aff$X$*, such that*

$$f(\bar{x} + h) = f(\bar{x}) + \langle g, h\rangle + o(\|h\|), \text{ for } h \in V.$$

When it exists, g is called the relative gradient of f at $\bar{x}$ and denoted by $\nabla_X f(\bar{x})$.

In order to make the above definition understood more easily, we introduce a new convex function $f_0(y) := f(x_0 + y)$, where x_0 is fixed in X and y varies in the subspace V. This transformation makes the domain of the new function f_0, which is also a proper, lower semi-continuous and convex function, full-dimensional in V, and hence f_0 is differentiable almost everywhere on $\text{int}(\text{dom} f_0) \subset V$, that is, f is relatively differentiable almost everywhere on ri$X$ (see [47, Page 117]). Actually, $\nabla_X f(x)$ is the gradient of $f_0(y)$ at $y := x - x_0$.

Since we need to calculate relative gradients at points in a certain neighborhood of the current iterate in the GS technique, we need all iterates to be relative interior points of the domain. However, if an iterate is not a relative interior point of X, we need to perform a perturbation step to guarantee the iterate to be a relative interior point and to ensure the GS technique can succeed. Therefore, the GS-SGM consists of generating a sequence $\{x_k\}$, by taking from x_k along the direction, which is constructed via random sampling of relative gradients at nearby points, and then projecting the resulting point onto X and finally performing a perturbation step.

In order to facilitate the reading and analysis of the algorithm, we provide a partial glossary of the notation, which is used in the GS-SGM, and present the GS-SGM algorithm as follows.

Table 2.1: Notation used in GS-SGM

k: Number of iterations.	s: Sample size.
v_k: Stepsize.	α_k: Perturbation weight.
μ_{ki}: Unit ball sample.	δ_k: Sampling radius.
x_k: Current iterate.	x_{ki}: Sampling point.
$\mathcal{D}$: Points of relative differentiability.	G_k: Approximate subdifferential.
λ_k: Iterative direction weight.	g_k: Iterative direction.
$\bar{x}_k$: Projection point.	y_k: Perturbation direction.

Subgradient method based on the gradient sampling technique (GS-SGM)

Step 1. (Initialization)

Start from $k = 0$, select an initial point $x_0 \in \mathrm{ri}X$ and parameters s, $\{v_k\}$ and $\{\alpha_k\}$.

Step 2. (Generate the approximate subdifferential via the GS technique)

Let $\mu_{k1}, \cdots, \mu_{ks}$ be sampled independently and uniformly from $B \cap V$, choose the sampling radius to satisfy $0 < \delta_k \leq \mathrm{dist}(x_k, \mathrm{rbd}X)$, and set

$$x_{ki} = x_k + \delta_k \mu_{ki}, \quad i = 1, \cdots, s.$$

If for some $i = 1, \cdots, s$, the point $x_{ki} \notin \mathcal{D}$, then STOP; otherwise, set

$$G_k = \mathrm{conv}\{\nabla_X f(x_{k1}), \cdots, \nabla_X f(x_{ks})\}.$$

Choose an arbitrary vector in G_k as the iterative direction, i.e.,

$$g_k = \sum_{i=1}^{s} \lambda_i \nabla_X f(x_{ki}), \text{ with } \sum_{i=1}^{s} \lambda_i = 1 \text{ and } \lambda_i \geq 0,$$

and go to Step 3.

Step 3. (Solution update and perturbation)

Compute $\bar{x}_{k+1} = P_X(x_k - v_k g_k)$ by solving the convex optimization problem

$$\begin{aligned} \min \quad & \|x - (x_k - v_k g_k)\|^2 \\ \text{s.t.} \quad & x \in X. \end{aligned}$$

Compute the perturbation direction of $\bar{x}_{k+1}$ as a certain relative interior point of X, which satisfies

$$y_k \in \{\bar{x}_{k+1} - N_X(\bar{x}_{k+1})\} \cap \mathrm{ri}X \cap B(\bar{x}_{k+1}, 1), \tag{2.2.1}$$

and set

$$x_{k+1} = (1 - \alpha_k)\bar{x}_{k+1} + \alpha_k y_k, \ 0 < \alpha_k < 1. \tag{2.2.2}$$

Set $k = k + 1$ and go back to Step 2.

The following remarks explain the choice of parameters and the designing of this algorithm.

Remark 2.2.1 *In Step 2, we choose $\delta_k \leq \mathrm{dist}(x_k, \mathrm{rbd}X)$ to keep all sampling points in X. By relating to α_{k-1}, the sampling radius δ_k in this form can be easily calculated in practical computation.*

Remark 2.2.2 *The perturbation direction plays a key role in the GS-SGM and its convergence analysis. In (2.2.1), we choose $y_k \in \mathrm{ri}X$ to guarantee $x_{k+1} \in \mathrm{ri}X$ and $y_k \in \{\bar{x}_{k+1} - N_X(\bar{x}_{k+1})\} \cap B(\bar{x}_{k+1}, 1)$ to obtain the convergence property of this algorithm. We will show the existence of such y_k in Lemma 2.3.3.*

In Step 2, if some random sampling point is not in $\mathcal{D}$, this algorithm will stop and it turns out to be failed. Fortunately, in the proof of Theorem 2.3.1, we show that the GS-SGM will not terminate finitely in Step 2, that is, a sequence of infinite points $\{x_k\}$ will be generated by the GS-SGM with probability 1.

The GS-SGM combine the subgradient method and the GS technique into one iteration. Thus, there are two main differences between the GS-SGM and the classical subgradient method. Firstly, it stems from the GS procedure. The classical subgradient method always assumes that the subgradient of the objective function can be obtained easily through a "black box", while the GS-SGM constructs the subgradient information

by calculating the convex combination of relative gradients sampled at random nearby points in Step 2. Secondly, it is the perturbation step. The classical subgradient method only proceeds a projection operator after solution updating. It makes each iterate x_k a feasible point that might not be a relative interior point of X. On the other hand, the GS-SGM performs a perturbation (2.2.2) after the projection operator. It makes each iterate x_{k+1} always be a relative interior point of X. However, it follows from (2.2.1) that $y_k = \bar{x}_{k+1}$ if $\bar{x}_{k+1}$ is a relative interior point of X, and thus the perturbation step is not necessary in this case. Therefore, when the optimal solution of problem (2.1.1) is a relative interior point of X, the GS-SGM differs little from the classical subgradient method in the perturbation step.

2.3 Convergence Analysis

In this section, we investigate convergence properties of the GS-SGM. First, we need to do some analysis on the procedures of this algorithm. The following Lemmas 2.3.1 and 2.3.2 tell us that each vector in G_k is an approximate subgradient of f at x_k.

Lemma 2.3.1 *Let g_1 be a subgradient of f at $x_1 \in X$. Then, for any $x_2 \in X$, g_1 is an ϵ-subgradient of f at x_2 with $\epsilon = f(x_2) - f(x_1) - \langle g_1, x_2 - x_1 \rangle$.*

Proof. By the definition of the subgradient (cf. (1.3.1)), we have

$$\langle g_1, x - x_1 \rangle \leq f(x) - f(x_1), \forall x \in \mathbb{R}^n.$$

Thus,

$$\begin{aligned} \langle g_1, x - x_2 \rangle &= \langle g_1, x - x_1 \rangle + \langle g_1, x_1 - x_2 \rangle \\ &\leq f(x) - f(x_1) - \langle g_1, x_2 - x_1 \rangle \\ &= f(x) - f(x_2) + f(x_2) - f(x_1) - \langle g_1, x_2 - x_1 \rangle \\ &= f(x) - f(x_2) + \epsilon, \forall x \in \mathbb{R}^n, \end{aligned}$$

with $\epsilon = f(x_2) - f(x_1) - \langle g_1, x_2 - x_1 \rangle$. ■

Lemma 2.3.2 *Let g_i be an ϵ_i-subgradient of f at $x \in X$ for $i = 1, \cdots, s$. Then, the convex combination $\sum_{i=1}^{s} \lambda_i g_i$ is also an ϵ-subgradient of f at x with $\epsilon = \sum_{i=1}^{s} \lambda_i \epsilon_i$.*

Proof. By the definition of the ϵ-subgradient (cf. (1.3.2)), we have

$$\langle g_i, y - x\rangle \leq f(y) - f(x) + \epsilon_i, \forall y \in \mathbb{R}^n, i = 1, \cdots, s.$$

Thus,

$$\begin{aligned}\langle \textstyle\sum_{i=1}^s \lambda_i g_i, y - x\rangle &= \textstyle\sum_{i=1}^s \lambda_i \langle g_i, y - x\rangle \\ &\leq f(y) - f(x) + \textstyle\sum_{i=1}^s \lambda_i \epsilon_i \\ &= f(y) - f(x) + \epsilon, \forall y \in \mathbb{R}^n,\end{aligned}$$

with $\epsilon = \sum_{i=1}^s \lambda_i \epsilon_i$. ■

From Lemmas 2.3.1-2.3.2, it follows that the iterative direction $g_k \in G_k$, generated in Step 2, is an approximate subgradient. Indeed, when $x_{ki} \in \mathcal{D}$, the relative gradient $\nabla_X f(x_{ki})$ is also a subgradient of f at x_{ki}. Thus, by using Lemma 2.3.1, we have that $\nabla_X f(x_{ki})$ is an ϵ_{ki}-subgradient of f at x_k with $\epsilon_{ki} = f(x_k) - f(x_{ki}) + \langle \nabla_X f(x_{ki}), \delta_k \mu_{ki}\rangle$. Furthermore, from Lemma 2.3.2, it follows that the convex combination $g_k = \sum\limits_{i=1}^{s} \lambda_i \nabla_X f(x_{ki})$ is an ϵ_k-subgradient of f at x_k with $\epsilon_k = \sum\limits_{i=1}^{s} \lambda_i \epsilon_{ki}$, that is,

$$\begin{aligned}\langle g_k, x - x_k\rangle &\leq f(x) - f(x_k) + \epsilon_k \\ &= f(x) - \sum_{i=1}^{s} \lambda_i f(x_{ki}) + \sum_{i=1}^{s} \lambda_i \langle \nabla_X f(x_{ki}), \delta_k \mu_{ki}\rangle, \forall x \in \mathbb{R}^n.\end{aligned} \tag{2.3.1}$$

The following lemma is very important for the GS-SGM. It demonstrates that Step 3 of the GS-SGM is well-defined in that it guarantees the existence of the perturbation direction.

Lemma 2.3.3 *Let $C \subset \mathbb{R}^n$ be a closed and convex set. Then for each $x \in C$, the intersection $\big(x - N_C(x)\big) \cap \mathrm{ri}C \cap B(x, 1)$ is nonempty.*

Proof. This lemma follows if we show that

$$\big(-N_C(x)\big) \cap \big(-x + \mathrm{ri}(C \cap B(x, 1))\big) \neq \varnothing,$$

by noting $\mathrm{ri}C \cap B(x, 1) = \mathrm{ri}(C \cap B(x, 1))$ (see [47, Proposition 2.1.10]).

By contradiction, suppose that $\big(-N_C(x)\big) \cap \big(-x + \mathrm{ri}(C \cap B(x, 1))\big) = \varnothing$. By the Separation Theorem (see Lemma 1.3.4), there exists some $s \neq 0$ such that

$$\langle s, -y\rangle \geq 0, \forall y \in N_C(x), \tag{2.3.2}$$

and

$$\langle s, -x + z\rangle < 0, \forall z \in \mathrm{ri}(C \cap B(x, 1)). \tag{2.3.3}$$

Taking z in the closure of $\mathrm{ri}(C \cap B(x, 1))$, the relation (2.3.3) implies that

$$\langle s, -x + z\rangle \leq 0, \forall z \in C \cap B(x, 1),$$

which is equivalent to $s \in N_C(x)$. Thus, it follows from (2.3.2) that $\langle s, -s\rangle \geq 0$, which implies $s = 0$. Hence, we arrive at the contradiction with the Separation Theorem $(s \neq 0)$. ■

Throughout the rest of this section, we use the following assumptions which are quite general in convex programming.

Assumption 2.3.1 *The optimal solution set X^* is nonempty.*

Assumption 2.3.2 *The relative gradients of f are bounded, i.e., there exists some scalar M such that $\|g\| \leq M$ for all $g \in \nabla_X f(\mathcal{D})$.*

In the convergence analysis, we start with the basic inequality, which shows a significant property of the sequence of iterates $\{x_k\}$.

Lemma 2.3.4 *Suppose Assumption 2.3.2 holds and the sequence $\{x_k\}$ is generated by the GS-SGM. Then for all $x \in X$, we have*

$$\|x_{k+1} - x\|^2 \leq \|x_k - x\|^2 - 2v_k\big(f(x_k) - f(x)\big) + 4v_k\delta_k M + v_k^2 M^2 + \alpha_k^2. \tag{2.3.4}$$

Proof. According to the GS-SGM, for all $x \in X$, we have

$$\begin{aligned} \|x_{k+1} - x\|^2 &= \|(1 - \alpha_k)\bar{x}_{k+1} + \alpha_k y_k - x\|^2 \\ &= \|\bar{x}_{k+1} - x - \alpha_k(\bar{x}_{k+1} - y_k)\|^2 \\ &= \|\bar{x}_{k+1} - x\|^2 - 2\alpha_k\langle \bar{x}_{k+1} - x, \bar{x}_{k+1} - y_k\rangle + \alpha_k^2\|\bar{x}_{k+1} - y_k\|^2. \end{aligned} \tag{2.3.5}$$

According to Lemma 2.3.3, the perturbation direction is well-defined. Due to the chosen rule of y_k (cf. (2.2.1)), $y_k \in \bar{x}_{k+1} - N_X(\bar{x}_{k+1})$ implies that

$$\langle \bar{x}_{k+1} - x, \bar{x}_{k+1} - y_k\rangle \geq 0, \forall x \in X,$$

and $y_k \in B(\bar{x}_{k+1}, 1)$ implies that $\|\bar{x}_{k+1} - y_k\| \leq 1$. Hence, the relation (2.3.5) reduces to

$$\begin{aligned} \|x_{k+1} - x\|^2 &\leq \|\bar{x}_{k+1} - x\|^2 + \alpha_k^2 \\ &\leq \|x_k - v_k g_k - x\|^2 + \alpha_k^2 \\ &= \|x_k - x\|^2 - 2v_k\langle g_k, x_k - x\rangle + v_k^2\|g_k\|^2 + \alpha_k^2, \forall x \in X, \end{aligned} \tag{2.3.6}$$

where the second inequality follows from the nonexpansive property of the projection operator (see Lemma 1.3.1).

It follows from Lemmas 2.3.1-2.3.2 that g_k is an ϵ_k-subgradient of f at x_k. Thus, substituting the subgradient inequality (2.3.1) into (2.3.6), we obtain

$$\begin{aligned} \|x_{k+1} - x\|^2 &\leq \|x_k - x\|^2 - 2v_k\Big(\sum_{i=1}^{s} \lambda_i f(x_{ki}) - f(x)\Big) \\ &\quad + 2v_k\Big(\sum_{i=1}^{s} \lambda_i\langle \nabla_X f(x_{ki}), \delta_k\mu_{ki}\rangle\Big) + v_k^2\|g_k\|^2 + \alpha_k^2 \\ &= \|x_k - x\|^2 - 2v_k\big(f(x_k) - f(x)\big) \\ &\quad + 2v_k\Big[\sum_{i=1}^{s} \lambda_i\Big(\langle \nabla_X f(x_{ki}), \delta_k\mu_{ki}\rangle + f(x_k) - f(x_{ki})\Big)\Big] + v_k^2\|g_k\|^2 + \alpha_k^2 \\ &\leq \|x_k - x\|^2 - 2v_k\big(f(x_k) - f(x)\big) \\ &\quad + 2v_k\Big(\sum_{i=1}^{s} \lambda_i\langle \nabla_X f(x_{ki}) - \nabla_X f(x_k), \delta_k\mu_{ki}\rangle\Big) + v_k^2\|g_k\|^2 + \alpha_k^2 \\ &\leq \|x_k - x\|^2 - 2v_k\big(f(x_k) - f(x)\big) \\ &\quad + 2v_k\delta_k\|\mu_{ki}\|\Big(\|\nabla_X f(x_{ki})\| + \|\nabla_X f(x_k)\|\Big) + v_k^2\|g_k\|^2 + \alpha_k^2 \\ &\leq \|x_k - x\|^2 - 2v_k\big(f(x_k) - f(x)\big) + 4v_k\delta_k M + v_k^2 M^2 + \alpha_k^2, \end{aligned}$$

where the second inequality follows from $f(x_k) - f(x_{ki}) \leq \langle \nabla_X f(x_k), -\delta_k\mu_{ki}\rangle$, and the fourth inequality follows from the bound on relative gradients M and sampling points in the unit ball. Thus, we arrive at the basic inequality (2.3.4). ■

Constant stepsize rule

We first describe the convergence property of the GS-SGM by using the constant stepsize rule.

Theorem 2.3.1 *Let Assumptions 2.3.1-2.3.2 hold. Suppose the sequence $\{x_k\}$ is generated by the GS-SGM with the constant stepsize rule, $\delta_k \leq \mathrm{dist}(x_k, \mathrm{rbd}X)$, and $\sum_{k=0}^{\infty} \alpha_k^2 < +\infty$. Then, $\underline{\lim}_{k\to\infty} f(x_k) \leq f_* + vM^2/2$ with probability 1.*

Proof. We begin the proof by making an observation concerning the stochastic structure of the GS-SGM. Although the algorithm specifies that the points $\mu_{k1}, \cdots, \mu_{km}$ are sampled from $B \cap V$ at each iteration, we may think of this sequence as a realization of a stochastic process $\{(\mu_{k1}, \cdots, \mu_{km})\}$ occurring before initiation of the algorithm. If the whole set has a strictly positive measure, by measure theory, x_{ki} lies outside any fixed subset of Lebesgue measure zero almost surely, which is independent of the sample size.

We first consider the case when the GS-SGM terminates finitely. Let $x \in \mathrm{ri}X$, $\delta > 0$, and z be a realization of a random variable that is uniformly distributed on $B \cap V$. Since f is relatively differentiable almost everywhere on $\mathrm{ri}X$ (also on X), by measure theory, the probability that $x + \delta z \notin \mathcal{D}$ is zero. Therefore, we have shown that the GS-SGM does not terminate finitely in Step 2 with probability 1.

We now restrict our attention to the case when the GS-SGM generates a sequence of infinite points $\{x_k\}$. According to the GS-SGM and Lemma 2.3.4 with $v_k \equiv v$, for all $x \in X$, we have

$$2v\big(f(x_k) - f(x)\big) \leq \|x_k - x\|^2 - \|x_{k+1} - x\|^2 + 4v\delta_k M + v^2 M^2 + \alpha_k^2. \tag{2.3.7}$$

Summing (2.3.7) over $k = 0, \cdots, n$, we arrive at

$$\frac{\sum\limits_{k=0}^{n} f(x_k)}{n} - f(x) \leq \frac{\|x_0 - x\|^2}{2nv} + 2M\frac{\sum\limits_{k=0}^{n} \delta_k}{n} + \frac{vM^2}{2} + \frac{\sum\limits_{k=0}^{n} \alpha_k^2}{2nv}. \tag{2.3.8}$$

In the GS-SGM, by the assumption, we have the following rules:

$$\begin{cases} \sum\limits_{k=0}^{\infty} \alpha_k^2 < +\infty, \\ \delta_k \leq \mathrm{dist}(x_k, \mathrm{rbd}X). \end{cases}$$

By the definitions of δ_k and α_k, we obtain $\delta_k \leq \mathrm{dist}(x_k, \mathrm{rbd}X) \leq \|x_k - \bar{x}_k\| \leq \alpha_{k-1}$. Furthermore, due to $\sum\limits_{k=0}^{\infty} \alpha_k^2 < +\infty$, we have $\lim\limits_{k\to\infty} \delta_k = 0$, and hence $\lim\limits_{n\to\infty} \frac{\sum_{k=0}^{n} \alpha_k^2}{2nv} = 0$ and $\lim\limits_{n\to\infty} \frac{\sum_{k=0}^{n} \delta_k}{n} = 0$ (cf. Lemma 1.3.5). Thus, by using Lemma 1.3.5, the relation (2.3.8) implies

$$\begin{aligned} \varliminf_{k\to\infty} f(x_k) &\leq \varliminf_{n\to\infty} \frac{\sum_{k=0}^{n} f(x_k)}{n} \\ &\leq f(x) + \frac{vM^2}{2}, \forall x \in X. \end{aligned}$$

Therefore, we arrive at that $\varliminf\limits_{k\to\infty} f(x_k) \leq f_* + vM^2/2$ with probability 1. ■

Divergent stepsize rule

The corresponding convergence result is indicated in the following theorem by using the divergent stepsize rule.

Theorem 2.3.2 *Let Assumptions 2.3.1-2.3.2 hold. Suppose the sequence $\{x_k\}$ is generated by the GS-SGM with the divergent stepsize rule* (1.1.6)*, $\delta_k \leq \mathrm{dist}(x_k, \mathrm{rbd}X)$, and $\sum\limits_{k=0}^{\infty} \alpha_k^2 < +\infty$. Then, x_k converges to some $x^* \in X^*$ and $\varliminf\limits_{k\to\infty} f(x_k) = f_*$, with probability 1.*

Proof. By the proof of Theorem 2.3.1, the GS-SGM does not terminate finitely in Step 2 with probability 1. We now focus on the case when the GS-SGM generates a sequence of infinite points $\{x_k\}$. According to the GS-SGM and Lemma 2.3.4, for all $x \in X$, we have

$$2v_k\big(f(x_k) - f(x)\big) \leq \|x_k - x\|^2 - \|x_{k+1} - x\|^2 + 4v_k\delta_k M + v_k^2 M^2 + \alpha_k^2. \tag{2.3.9}$$

Summing (2.3.9) over $k = 0, \cdots, n$, we obtain

$$\frac{\sum\limits_{k=0}^{n} v_k f(x_k)}{\sum\limits_{k=0}^{n} v_k} - f(x) \leq \frac{\|x_0 - x\|^2}{2\sum\limits_{k=0}^{n} v_k} + 2M\frac{\sum\limits_{k=0}^{n} v_k\delta_k}{\sum\limits_{k=0}^{n} v_k} + \frac{M^2\sum\limits_{k=0}^{n} v_k^2 + \sum\limits_{k=0}^{n} \alpha_k^2}{2\sum\limits_{k=0}^{n} v_k}. \tag{2.3.10}$$

In the GS-SGM, by the assumption, we have the following rules:

$$\begin{cases} \sum\limits_{k=0}^{\infty} v_k = +\infty \text{ and } \sum\limits_{k=0}^{\infty} v_k^2 < +\infty, \\ \sum\limits_{k=0}^{\infty} \alpha_k^2 < +\infty, \\ \delta_k \leq \mathrm{dist}(x_k, \mathrm{rbd}X). \end{cases} \tag{2.3.11}$$

By the definitions of δ_k and α_k, we have $\delta_k \leq \mathrm{dist}(x_k, \mathrm{rbd}X) \leq \|x_k - \bar{x}_k\| \leq \alpha_{k-1}$. Also, due to $\sum\limits_{k=0}^{\infty} \alpha_k^2 < +\infty$ and $\sum\limits_{k=0}^{\infty} v_k^2 < +\infty$, we have $\sum\limits_{k=1}^{\infty} v_k\alpha_{k-1} < +\infty$ and thus $\sum\limits_{k=0}^{\infty} v_k\delta_k < +\infty$. Therefore, we arrive at the following results:

$$\begin{cases} \lim\limits_{n\to\infty} \frac{\|x_0 - x\|^2}{\sum\limits_{k=0}^{n} v_k} = 0, \\ \lim\limits_{n\to\infty} \frac{\sum\limits_{k=0}^{n} v_k\delta_k}{\sum\limits_{k=0}^{n} v_k} = 0, \\ \lim\limits_{n\to\infty} \frac{M^2\sum\limits_{k=0}^{n} v_k^2 + \sum\limits_{k=0}^{n} \alpha_k^2}{\sum\limits_{k=0}^{n} v_k} = 0. \end{cases}$$

Substituting the above three relations into (2.3.10), we obtain

$$\begin{aligned}\underline{\lim}_{k\to\infty} f(x_k) &\leq \underline{\lim}_{n\to\infty} \frac{\sum_{k=0}^{n} v_k f(x_k)}{\sum_{k=0}^{n} v_k} \\ &\leq f(x), \forall x \in X, \end{aligned} \tag{2.3.12}$$

where the first inequality follows from Lemma 1.3.5. Therefore, we prove that $\underline{\lim}_{k\to\infty} f(x_k) = f_*$ with probability 1.

We next prove convergence in iterates $\{x_k\}$ as follows. By using (2.3.4) with any $x^* \in X^*$, we have

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 - 2v_k\big(f(x_k) - f_*\big) + 4v_k\delta_k M + v_k^2 M^2 + \alpha_k^2 \\ &\leq \|x_k - x^*\|^2 + 4v_k\delta_k M + v_k^2 M^2 + \alpha_k^2. \end{aligned} \tag{2.3.13}$$

The rules (2.3.11) immediately imply that

$$\sum_{k=0}^{\infty} \Big(4v_k\delta_k M + v_k^2 M^2 + \alpha_k^2\Big) < \infty.$$

Hence, it follows from Lemma 1.3.6 with $b_k = 0$ and $c_k = 4v_k\delta_k M + v_k^2 M^2 + \alpha_k^2$ that $a_k = \|x_k - x^*\|$ converges. Thus, the sequence $\{x_k\}$ is bounded. Furthermore, we have proved that $\underline{\lim}_{k\to\infty} f(x_k) = f_*$, which together with the boundedness of $\{x_k\}$ and the lower semi-continuity of f implies that $\{x_k\}$ has a cluster point $x^* \in X^*$. Finally, x_k converges to x^* from (2.3.13), noting that the tail sums $\sum_{i=k}^{\infty} \Big(4v_i\delta_i M + v_i^2 M^2 + \alpha_i^2\Big)$ vanishes as k tends to infinity. ■

2.4 Calculate the Perturbation Direction

The perturbation direction, $y_k \in \{\bar{x}_{k+1} - N_X(\bar{x}_{k+1})\} \cap \mathrm{ri}X \cap B(\bar{x}_{k+1}, 1)$, plays a key role in the GS-SGM to guarantee that each iterate is a relative interior point and to achieve the convergence property. We have proved its existence in Lemma 2.3.3. In this section, we show how this direction can be calculated in two common cases of the domain X: X is a convex polyhedron and the sublevel set of some convex quadratic functions. For the sake of simplicity, we denote $y_k = \bar{x} - \bar{s}$ with $\bar{x} = \bar{x}_{k+1}$ and $\bar{s} \in N_X(\bar{x}_{k+1})$ in what follows.

Case 1. Convex Polyhedron

If X is a convex polyhedron in $\mathbb{R}^n$, i.e., for some $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ $(i = 1, \cdots, m)$,

$$X := \{x : \langle a_i, x\rangle \leq b_i, \|a_i\| = 1, i = 1, 2, \cdots, m\},$$

which is assumed not to be a singleton. For each $\bar{x} \in X$, the active index set is defined by

$$J(\bar{x}) := \{i : \langle a_i, \bar{x}\rangle = b_i, i = 1, 2, \cdots, m\}.$$

Without loss of generality, deleting the redundant constraints, we assume that $\{a_i : i \in J(\bar{x})\}$ is linear independent. It is well-known that the normal cone to X at $\bar{x}$ is given by (see [98, Proposition 10.3])

$$\begin{aligned} N_X(\bar{x}) &= \text{cone}\{a_j : j \in J(\bar{x})\} \\ &= \{\sum_{j\in J(\bar{x})} \beta_j a_j : \beta_j \geq 0\}. \end{aligned}$$

Since $\bar{s} \in N_X(\bar{x})$, it can be represented as $\bar{s} = \sum_{j\in J(\bar{x})} \beta_j a_j$ with $\beta_j \geq 0$.

In order to calculate the perturbation direction, in the following, we will deduce the condition on parameters β_i such that $\bar{x} - \bar{s} \in \text{ri}X \cap B(\bar{x}, 1)$, that is $\|\bar{s}\| \leq 1$ and $\langle a_i, \bar{x} - \bar{s}\rangle < b_i$ for $i = 1, 2, \cdots, m$. The deduction is divided into two cases: (i) $\bar{x} \in \text{ri}X$ and (ii) $\bar{x} \in \text{rbd}X$.

(i) $\bar{x} \in \text{ri}X$.

In this case, $J(\bar{x}) = \varnothing$, $\bar{s} \in N_X(\bar{x}) = \{0\}$, and $\bar{x} - \bar{s} = \bar{x} \in \text{ri}X$.

(ii) $\bar{x} \in \text{rbd}X$.

Deduction of the condition on parameters β_i is split into two cases of index: (a) $i \notin J(\bar{x})$ and (b) $i \in J(\bar{x})$, as follows.

(a) $i \notin J(\bar{x})$.

In this case, we have $\langle a_i, \bar{x}\rangle < b_i$. Choose β_j for all $j \in J(\bar{x})$ satisfying

$$\beta_j < \frac{1}{|J(\bar{x})|} \min\{\min_{i\notin J(\bar{x})} \frac{b_i - \langle a_i, \bar{x}\rangle}{\max\{-\langle a_i, a_j\rangle, 0\}}, 1\}, \tag{2.4.1}$$

then we obtain the inequalities

$$\|\bar{s}\| \leq 1 \quad \text{and} \quad \langle a_i, \bar{x} - \bar{s}\rangle < b_i, \forall i \notin J(\bar{x}).$$

(b) $i \in J(\bar{x})$.

In this case, we have $\langle a_i, \bar{x}\rangle = b_i$. Hence, to guarantee $\langle a_i, \bar{x}-\bar{s}\rangle < b_i$ for all $i \in J(\bar{x})$, we need to choose β_j such that

$$\sum_{j\in J(\bar{x})} \beta_j \langle a_i, a_j\rangle > 0, \forall i \in J(\bar{x}). \tag{2.4.2}$$

According to what have been shown above, $\bar{s} = \sum_{j\in J(\bar{x})} \beta_j a_j$ with β_j satisfying (2.4.1) and (2.4.2) has the properties that $\bar{s} \in N_X(\bar{x})$ and $\bar{x} - \bar{s} \in \mathrm{ri}X \cap B(\bar{x}, 1)$, which is the perturbation direction.

Remark 2.4.1 *In practical computation, we always use the following equivalent linear system*

$$\sum_{j\in J(\bar{x})} \beta_j \langle a_i, a_j\rangle \geq 1, \forall i \in J(\bar{x}).$$

instead of (2.4.2) *to find parameters* β_j.

Remark 2.4.2 *The perturbation direction in the box constraint case is particularly easy to calculate. For example, if* $X = \mathbb{R}^n_+$, *we obtain*

$$\bar{s} = \frac{sign(\bar{x}) - e}{\sqrt{n}},$$

where $e = (1, 1, \cdots, 1)^T$ *and* $sign(\cdot)$ *denotes the sign function. It is a motivation factor to apply the GS-SGM to the Lagrangian dual of a convex constrained optimization problem, which will be shown in Chapter 3.*

Case 2. Sublevel set of some convex quadratic functions

Define

$$f_i(x) = \frac{1}{2}x^T Q_i x + c_i^T x + d_i, i \in I = \{1, 2, \cdots, m\},$$

where Q_i is symmetric, positive semi-definite matrix, $c_i \in \mathbb{R}^n$ and $d_i \in \mathbb{R}$ $(i = 1, \cdots, m)$.

Let X be the sublevel set of these convex quadratic functions, i.e.,

$$X := \{x : f_i(x) \leq 0,\ \forall i \in I\}.$$

For each $\bar{x} \in X$, the active index set is defined by

$$J(\bar{x}) := \{i : f_i(\bar{x}) = 0, i \in I\}.$$

Similarly, we will calculate the perturbation direction by looking at the two situations: (i) $\bar{x} \in \mathrm{ri}X$ and (ii) $\bar{x} \in \mathrm{rbd}X$.

(i) $\bar{x} \in \mathrm{ri}X$.

In this case, $J(\bar{x}) = \varnothing$, $\bar{s} \in N_X(\bar{x}) = \{0\}$, and $\bar{x} - \bar{s} = \bar{x} \in \mathrm{ri}X$.

(ii) $\bar{x} \in \mathrm{rbd}X$.

From [98, Proposition 10.3], by the regularity of f, the normal cone of X at $\bar{x}$ is given by

$$\begin{aligned} N_X(\bar{x}) &= \mathrm{cone}\{\nabla f_i(\bar{x}) : i \in J(\bar{x})\} \\ &= \mathrm{cone}\{Q_i\bar{x} + c_i : i \in J(\bar{x})\} \\ &= \{\sum_{j \in J(\bar{x})} \beta_j(Q_j\bar{x} + c_j) : \beta_j \geq 0\}. \end{aligned}$$

Since $\bar{s} \in N_X(\bar{x})$, it can be represented as $\bar{s} = \sum_{j \in J(\bar{x})} \beta_j(Q_j\bar{x} + c_j)$ with $\beta_j \geq 0$.

In short, denote $\beta = (\beta_1, \cdots, \beta_{|J(\bar{x})|})^T \in \mathbb{R}^{|J(\bar{x})|}$, $a_j = Q_j\bar{x} + c_j$ and $A = (a_1, \cdots, a_{|J(\bar{x})|}) \in \mathbb{R}^{n \times |J(\bar{x})|}$, and thus $\bar{s} = A\beta$.

In order to calculate the perturbation direction, in the following, we will deduce the condition on parameter β such that $\bar{x} - \bar{s} \in \mathrm{ri}X \cap B(\bar{x}, 1)$, that is, $\|\bar{s}\| \leq 1$ and $f_i(\bar{x} - \bar{s}) < 0$ for all $i \in I$, in two cases of index: (a) $i \notin J(\bar{x})$ and (b) $i \in J(\bar{x})$.

(a) $i \notin J(\bar{x})$.

In this case, we have $f_i(\bar{x}) < 0$. Thus,

$$\begin{aligned} f_i(\bar{x} - \bar{s}) &= \tfrac{1}{2}(\bar{x} - \bar{s})^T Q_i(\bar{x} - \bar{s}) + c_i^T(\bar{x} - \bar{s}) + d_i \\ &= \tfrac{1}{2}\bar{s}^T Q_i \bar{s} - \bar{x}^T Q_i \bar{s} - c_i^T \bar{s} + f_i(\bar{x}) \\ &= \tfrac{1}{2}\beta^T(A^T Q_i A)\beta - (\bar{x}^T Q_i + c_i^T)A\beta + f_i(\bar{x}). \end{aligned}$$

To guarantee $f_i(\bar{x} - \bar{s}) < 0$, we shall choose $\beta \in \mathbb{R}^{|J(\bar{x})|}$ such that

$$\frac{1}{2}\beta^T(A^T Q_i A)\beta - (\bar{x}^T Q_i + c_i^T)A\beta + f_i(\bar{x}) < 0, \forall i \notin J(\bar{x}). \tag{2.4.3}$$

Such β can easily be shown to exist since $f_i(\bar{x}) < 0$.

(b) $i \in J(\bar{x})$.

In this case, we have $f_i(\bar{x}) = 0$. Thus, as shown in (2.4.3), we need

$$\frac{1}{2}\beta^T(A^TQ_iA)\beta - (\bar{x}^TQ_i + c_i^T)A\beta < 0, \forall i \in J(\bar{x}). \tag{2.4.4}$$

According to what have been shown above, $\bar{s} = \sum_{j\in J(\bar{x})} \beta_j(Q_j\bar{x} + c_j)$ with β_j satisfying (2.4.3), (2.4.4) and $\|\bar{s}\| \leq 1$ has the properties that $\bar{s} \in N_X(\bar{x})$ and $\bar{x} - \bar{s} \in \mathrm{ri}X \cap B(\bar{x}, 1)$, which is the perturbation direction required.

For the sake of simplicity, let

$$F_i(\beta) = \frac{1}{2}\beta^T P_i\beta - h_i^T\beta + w_i(\bar{x}),$$

where $P_i = A^TQ_iA$, $h_i = (\bar{x}^TQ_i + c_i^T)A$ and

$$w_i(\bar{x}) = \begin{cases} 0, & \text{if } i \in J(\bar{x}), \\ f_i(\bar{x}), & \text{otherwise.} \end{cases}$$

Then inequalities (2.4.3) and (2.4.4) are equivalent to $F_i(\beta) < 0$ for all $i \in I$.

Ignoring terms involving β^2, by choosing β small enough, $F_i(\beta) < 0$ is equivalent to

$$(-h_i)^T\beta + w_i(\bar{x}) < 0, \forall i \in I. \tag{2.4.5}$$

Therefore, calculating such β_j that satisfy (2.4.3), (2.4.4) and $\|\bar{s}\| \leq 1$ is equivalent to finding β_j small enough and that satisfy (2.4.5), which can be done computationally as in Case 1 (Remark 2.4.1).

2.5 A Simple Version of GS-SGM

We have presented a GS-SGM for the nonsmooth convex optimization problem, proved its convergence to an optimal solution with probability 1, and calculated the perturbation direction in two common cases of the domain.

In previous sections, especially by the theoretical part, we consider the extended real-valued objective function, whose domain may have an empty interior. In this case, we need to use the relative gradient instead of the gradient and use the relative

gradients at random nearby points to construct the subgradient information. For the simple model of minimizing a convex function $f : \mathbb{R}^n \to \mathbb{R}$ over a constraint set (see [16, 17, 30, 31, 81, 80, 82, 87, 100, 106] and references therein), the GS-SGM reduces to a much simpler form (in short, GS-SGM$^{\mathrm{S}}$). Hence, in this section, we consider the following convex constrained optimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in X, \end{aligned} \tag{2.5.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, and X is a closed and convex set. Denote the optimal solution set and the optimal value of problem (2.5.1) respectively by X^* and f_*. It is worth mentioning that (2.5.1) is a special case of (2.1.1) as the objective function is replaced by $f + \delta_X$.

Since the domain of the objective function is the whole space, all points are interior points of $\mathbb{R}^n$, and thus we do not need to consider the relative gradient and perturbation step. Only the GS procedure is added in the iterative process of the classical subgradient method. Numerical tests on the low-rank recovery problem, in Section 2.6, show that the GS technique can improve the convergence behavior of the subgradient method.

2.5.1 GS-SGM$^{\mathrm{S}}$ Algorithm

In this subsection, we propose a simple version of GS-SGM (GS-SGM$^{\mathrm{S}}$) to solve problem (2.5.1). The GS-SGM$^{\mathrm{S}}$ is similar to the GS-SGM. The two main differences are that the gradients at random nearby points are used instead of the relative gradients, and that the perturbation step is skipped, because the domain of the objective function is the whole space. The GS-SGM$^{\mathrm{S}}$ constructs the subgradient information via random sampling of gradients at nearby points and projects the point, obtained from the subgradient iteration along the constructed direction, onto the constraint set.

In order to facilitate the reading and analysis of the algorithm, we provide a partial glossary of the notation, which is used in the GS-SGM$^{\mathrm{S}}$, and present the GS-SGM$^{\mathrm{S}}$ algorithm as follows.

A simple version of GS-SGM (GS-SGM$^{\mathrm{S}}$)

Table 2.2: Notation used in GS-SGM$^{\mathrm{S}}$

k: Number of iterations.	s: Sample size.
v_k: Stepsize.	δ_k: Sampling radius.
x_k: Current iterate.	x_{ki}: Sampling point.
μ_{ki}: Unit ball sample.	G_k: Approximate subdifferential.
$\mathcal{D}$: Points of differentiability.	g_k: Iterative direction.
λ_k: Iterative direction weight.	

Step 1. (Initialization)

Start from $k = 0$, select an initial pint $x_0 \in \mathrm{ri}X$ and parameters s, $\{v_k\}$ and $\{\delta_k\}$.

Step 2. (Generate the approximate subdifferential via the GS technique)

Let $\mu_{k1}, \cdots, \mu_{ks}$ be sampled independently and uniformly from B, and set

$$x_{ki} = x_k + \delta_k \mu_{ki}, \quad i = 1, \cdots, s.$$

If for some $i = 1, \cdots, s$, the point $x_{ki} \notin \mathcal{D}$, then STOP; otherwise, set

$$G_k = \mathrm{conv}\{\nabla f(x_{k1}), \cdots, \nabla f(x_{ks})\}.$$

Choose an arbitrary vector in G_k as the iterative direction, i.e.,

$$g_k = \sum_{i=1}^{s} \lambda_i \nabla f(x_{ki}), \text{ with } \sum_{i=1}^{s} \lambda_i = 1 \text{ and } \lambda_i \geq 0,$$

and go to Step 3.

Step 3. (Solution update)

Compute $x_{k+1} = P_X(x_k - v_k g_k)$ by solving the convex optimization problem

$$\begin{aligned} \min \quad & \|x - (x_k - v_k g_k)\|^2 \\ \text{s.t.} \quad & x \in X. \end{aligned}$$

Set $k = k + 1$ and go back to Step 2.

It has been illustrated in the proof of Theorem 2.3.1 that the GS-SGM$^{\mathrm{S}}$ will not terminate finitely in Step 2, that is, a sequence of infinite points $\{x_k\}$ will be generated by the GS-SGM$^{\mathrm{S}}$ with probability 1.

The only difference between the GS-SGM$^{\mathrm{S}}$ and the classical subgradient method is the GS procedure. The classical SGM always assumes that the subgradient of the objective function can be obtained through a "black box", while the GS-SGM$^{\mathrm{S}}$ constructs the subgradient information by calculating the convex combination of gradients sampled at random nearby points in Step 2.

2.5.2 Convergence Analysis

In this subsection, we investigate convergence properties of the GS-SGM$^{\mathrm{S}}$ for different types of stepsize rules and sampling radius rules.

Similar to what have been described in Section 2.3, from Lemmas 2.3.1-2.3.2, it follows that the iterative direction $g_k \in G_k$, generated in Step 2, is an approximate subgradient, that is

$$\begin{aligned} \langle g_k, x - x_k \rangle &\leq f(x) - f(x_k) + \epsilon_k \\ &= f(x) - \sum_{i=1}^{s} \lambda_i f(x_{ki}) + \sum_{i=1}^{s} \lambda_i \langle \nabla f(x_{ki}), \delta_k \mu_{ki} \rangle, \forall x \in \mathbb{R}^n. \end{aligned} \tag{2.5.2}$$

Throughout the rest of this section, we use the following assumption which is quite natural in convex programming.

Assumption 2.5.1 *The gradients of f are bounded, i.e., there exists a scalar M such that $\|g\| \leq M$ for all $g \in \nabla f(\mathcal{D})$.*

In the convergence analysis, we start with the basic inequality, which shows a significant property of the iterate sequence $\{x_k\}$.

Lemma 2.5.1 *Suppose Assumption 2.5.1 holds and the sequence $\{x_k\}$ is generated by the GS-SGM^S. Then for all $x \in X$, we have*

$$\|x_{k+1} - x\|^2 \leq \|x_k - x\|^2 - 2v_k\big(f(x_k) - f(x)\big) + 4v_k\delta_k M + v_k^2 M^2. \tag{2.5.3}$$

Proof. According to the GS-SGM$^{\mathrm{S}}$, it follows from the nonexpansive property of the

projection operator that

$$\begin{aligned}\|x_{k+1}-x\|^2 &= \|P_X(x_k - v_k g_k) - x\|^2 \\ &\le \|x_k - v_k g_k - x\|^2 \\ &= \|x_k - x\|^2 - 2v_k\langle g_k, x_k - x\rangle + v_k^2\|g_k\|^2 \\ &\le \|x_k - x\|^2 - 2v_k\big(f(x_k) - f(x)\big) \\ &\quad + 2v_k\Big[\sum_{i=1}^{s}\lambda_i\Big(\langle\nabla f(x_{ki}), \delta\mu_{ki}\rangle + f(x_k) - f(x_{ki})\Big)\Big] + v_k^2\|g_k\|^2 \\ &\le \|x_k - x\|^2 - 2v_k\big(f(x_k) - f(x)\big) + 4v_k\delta_k M + v_k^2 M^2, \forall x \in X,\end{aligned}$$

where the second inequality follows from (2.5.2), and the third inequality follows from $f(x_k) - f(x_{ki}) \le \langle\nabla f(x_k), -\delta_k\mu_{ki}\rangle$, the bound on gradients M and sampling points in the unit ball. Thus, we arrive at the basic inequality (2.5.3). ■

Constant stepsize rule

We first describe the convergence property of the GS-SGM$^{\mathrm{S}}$ by using the constant stepsize rule and the constant sampling radius.

Theorem 2.5.1 *Let Assumption 2.5.1 hold. Suppose the sequence $\{x_k\}$ is generated by the GS-SGM^S with the constant stepsize rule and the constant sampling radius δ. Then, $\underline{\lim}_{k\to\infty} f(x_k) \le f_* + 2M\delta + vM^2/2$ with probability 1.*

Proof. By the proof of Theorem 2.3.1, the GS-SGM$^{\mathrm{S}}$ does not terminate finitely in Step 2 with probability 1. We now focus on the case when the GS-SGM$^{\mathrm{S}}$ generates a sequence of infinite points $\{x_k\}$.

According to the GS-SGM$^{\mathrm{S}}$ and Lemma 2.5.1 with $v_k \equiv v$ and $\delta_k \equiv \delta$, for all $x \in X$, we have

$$\|x_{k+1} - x\|^2 \le \|x_k - x\|^2 - 2v\big(f(x_k) - f(x)\big) + 4v\delta M + v^2M^2. \tag{2.5.4}$$

Summing (2.5.4) over $k = 0, \cdots, n$, we obtain

$$\frac{\sum_{k=0}^{n} f(x_k)}{n} - f(x) \le \frac{\|x_0 - x\|^2}{2nv} + 2M\delta + \frac{vM^2}{2}.$$

Thus, by using Lemma 1.3.5, we arrive at

$$\begin{aligned}\underline{\lim}_{k\to\infty} f(x_k) &\leq \underline{\lim}_{n\to\infty} \frac{\sum_{k=0}^{n} f(x_k)}{n} \\ &\leq f(x) + 2M\delta + \frac{vM^2}{2}, \forall x \in X.\end{aligned}$$

Therefore, we arrive at that $\underline{\lim}_{k\to\infty} f(x_k) \leq f_* + 2M\delta + vM^2/2$ with probability 1. ■

When the sampling radius is reduced dynamically, the error term involving the sampling radius δ vanishes.

Theorem 2.5.2 *Let Assumption 2.5.1 hold. Suppose the sequence $\{x_k\}$ is generated by the GS-SGM^S with the constant stepsize rule and $\lim_{k\to\infty} \delta_k = 0$. Then, $\underline{\lim}_{k\to\infty} f(x_k) \leq f_* + vM^2/2$ with probability 1.*

Proof. By the proof of Theorem 2.3.1, the GS-SGM$^{\mathrm{S}}$ does not terminate in Step 2 with probability 1. We now focus on the case when the GS-SGM$^{\mathrm{S}}$ generates a sequence of infinite points $\{x_k\}$.

Summing (2.5.3) over $k = 0, \cdots, n$, we obtain

$$\frac{\sum_{k=0}^{n} f(x_k)}{n} - f(x) \leq \frac{\|x_0 - x\|^2}{2nv} + 2M\frac{\sum_{k=0}^{n} \delta_k}{n} + \frac{vM^2}{2}.$$

Thus, by using Lemma 1.3.5, we arrive at

$$\begin{aligned}\underline{\lim}_{k\to\infty} f(x_k) &\leq \underline{\lim}_{n\to\infty} \frac{\sum_{k=0}^{n} f(x_k)}{n} \\ &\leq f(x) + \frac{vM^2}{2}, \forall x \in X,\end{aligned}$$

where the second inequality holds since $\lim_{n\to\infty} \frac{\sum_{k=0}^{n} \delta_k}{n} = 0$, which follows from Lemma 1.3.5 and $\lim_{k\to\infty} \delta_k = 0$. Therefore, we arrive at that $\underline{\lim}_{k\to\infty} f(x_k) \leq f_* + vM^2/2$ with probability 1. ■

Divergent stepsize rule

The corresponding convergence results are indicated in the following theorems by using the diminishing stepsize rule.

Theorem 2.5.3 *Let Assumption 2.5.1 hold. Suppose the sequence* $\{x_k\}$ *is generated by the GS-SGM*^{S} *with the divergent stepsize rule* (1.1.6) *and the constant sampling radius* δ. *Then,* $\underline{\lim}_{k\to\infty} f(x_k) \leq f_* + 2M\delta$ *with probability 1.*

Proof. By the proof of Theorem 2.3.1, the GS-SGM$^{\rm S}$ does not terminate in Step 2 with probability 1. We now focus on the case when the GS-SGM$^{\rm S}$ generates a sequence of infinite points $\{x_k\}$.

Summing (2.5.3) over $k = 0, \cdots, n$, we obtain

$$\frac{\sum\limits_{k=0}^{n} v_k f(x_k)}{\sum\limits_{k=0}^{n} v_k} - f(x) \leq \frac{\|x_0 - x\|^2}{2\sum\limits_{k=0}^{n} v_k} + 2M\delta + M^2 \frac{\sum\limits_{k=0}^{n} v_k^2}{2\sum\limits_{k=0}^{n} v_k}.$$

Thus, by using Lemma 1.3.5, we obtain

$$\begin{aligned} \underline{\lim}_{k\to\infty} f(x_k) &\leq \underline{\lim}_{n\to\infty} \frac{\sum\limits_{k=0}^{n} v_k f(x_k)}{\sum\limits_{k=0}^{n} v_k} \\ &\leq f(x) + 2M\delta, \forall x \in X, \end{aligned}$$

where the second inequality follows from the properties of the divergent stepsize rule (cf. (1.1.6)). Therefore, we arrive at that $\underline{\lim}_{k\to\infty} f(x_k) \leq f_* + 2M\delta$ with probability 1. ■

Theorem 2.5.4 *Let Assumption 2.5.1 hold. Suppose the sequence* $\{x_k\}$ *is generated by the GS-SGM*^{S} *with the divergent stepsize rule* (1.1.6) *and* $\lim_{k\to\infty} \delta_k = 0$. *Then,* $\underline{\lim}_{k\to\infty} f(x_k) = f_*$ *with probability 1.*

Proof. By the proof of Theorem 2.3.1, the GS-SGM$^{\rm S}$ does not terminate in Step 2 with probability 1. We now focus on the case when the GS-SGM$^{\rm S}$ generates a sequence of infinite points $\{x_k\}$.

Summing (2.5.3) over $k = 0, \cdots, n$, we obtain

$$\frac{\sum\limits_{k=0}^{n} v_k f(x_k)}{\sum\limits_{k=0}^{n} v_k} - f(x) \leq \frac{\|x_0 - x\|^2}{2\sum\limits_{k=0}^{n} v_k} + 2M \frac{\sum_{k=0}^{n} v_k \delta_k}{\sum\limits_{k=0}^{n} v_k} + M^2 \frac{\sum\limits_{k=0}^{n} v_k^2}{2\sum\limits_{k=0}^{n} v_k}.$$

Thus, by using Lemma 1.3.5, we obtain

$$\begin{aligned}\underline{\lim}_{k\to\infty} f(x_k) &\leq \underline{\lim}_{n\to\infty} \frac{\sum_{k=0}^{n} v_k f(x_k)}{\sum_{k=0}^{n} v_k} \\ &\leq f(x), \forall x \in X,\end{aligned}$$

where the second inequality follows from the properties of the divergent stepsize rule (cf. (1.1.6)) and $\lim_{k\to\infty} \delta_k = 0$. Therefore, we arrive at that $\underline{\lim}_{k\to\infty} f(x_k) = f_*$ with probability 1. ■

2.6 Numerical Experiments

In this section, we show some numerical experiments to illustrate that the GS-SGM/GS-SGM$^{\mathrm{S}}$ is comparable with some existing subgradient algorithms. In the first experiment, we compare the GS-SGM with the algorithm proposed by Ruszczyński [100] on a non-smooth convex optimization. In the second experiment, cited in [80], we apply the GS-SGM to solve the dual problem arising from the assignment problem and compare with the incremental subgradient method used in [80]. In the third experiment, we use the GS-SGM$^{\mathrm{S}}$ to deal with the low rank-recovery problem. Applying the GS-SGM$^{\mathrm{S}}$, we can recover the MIT logo and PolyU logo clearly.

Before we present the numerical experiments in detail, we need to clarify some points in the numerical experiments.

Since the subgradient method is not a descent method, it is common to keep track of the best point found so far, i.e., the one with the least function value so far. Therefore, at each iteration, we set the record value

$$f_k^{rec} := \min\{f_k, f_{k-1}^{rec}\}.$$

This technique makes the sequence $\{f_k^{rec}\}$ to be nonincreasing.

We have presented the convergence theory of the GS-SGM for minimizing an extended real-valued convex function, whose domain might have an empty interior. In that case, we need to introduce the relative gradient. In the following three numerical

experiments, domains of objective functions are all full dimensional, hence relative gradients reduce to gradients for these cases. Following [31], we do not attempt to check whether the iterates lie in the set $\mathcal{D}$ where f is differentiable in Step 2. Moreover, we have demonstrated that the GS-SGM does not terminate finitely in Step 2 and thus generates a sequence of infinite points $\{x_k\}$ with probability 1. Therefore, we skip the differentiability check and assume that we have the information whether the gradient of the objective function exists or not at a given point.

Another issue is the stopping criterion. Besides the nondifferentiable information, we do not set any stopping criterion in the GS-SGM/GS-SGM$^{\mathrm{S}}$. Lack of implementable stopping criterion is a traditional drawback of subgradient methods. This drawback originated from the nondescent property of the subgradient direction. If we cannot obtain or estimate the optimal value, it is really hard to set an effective stopping criterion. One common trick is to check whether there is any improvement in the last 100 iterations. If f_k^{rec} does not decrease in the last 100 iterations, then we stop and obtain the optimal value so far. Another idea is to use the primal-dual subgradient method. The natural stopping criterion is the gap between the primal function value and the dual function value. In the following numerical experiments, we do not set any stopping criterion and just illustrate the performance of the GS-SGM/GS-SGM$^{\mathrm{S}}$, comparing with other algorithms in the specified number of iterations.

Nonsmooth convex optimization

Consider the nonsmooth convex optimization problem (see [100])

$$\min_{x\in\mathbb{R}^n} f(x), \tag{2.6.1}$$

where $f:\mathbb{R}^n\to\bar{\mathbb{R}}$ is defined by

$$f(x)=\begin{cases}\max\{f_1(x),f_2(x)\}, & \text{if } x\in\mathbb{R}^n_+,\\ +\infty, & \text{otherwise,}\end{cases}$$

with

$$f_1(x)=\alpha-c^Tx,\qquad f_2(x)=\frac{1}{2}x^TDx.$$

As in [100], we set $n=100$ and

$$\alpha=n\cdot\mathrm{rand}(),\quad c=2\cdot\mathrm{rand}(n,1)-e,\quad D=\mathrm{diag}(\mathrm{rand}(n,1)).$$

Here, rand() denotes a random value drawn from an uniform distribution on the unit interval, rand$(n, 1)$ denotes a column vector with n elements that all take random values on the unit interval, e denotes the vector in $\mathbb{R}^n$ with all elements 1, and D is a diagonal matrix with random diagonal entries.

Solved using CVX[1], the optimal value for an instance of the above problem is $f_* = 18.5166$.

Ruszczyński [100] designed the subgradient algorithm based on a merit function approach (MFA-SGM) (see Section 1.1, Page 13). The MFA-SGM differs from the GS-SGM in two main ways. The first difference is the GS procedure. The MFA-SGM updates the subgradient information by calculating the convex combination of current subgradient and successive direction, while the GS-SGM constructs the subgradient information via calculating the convex combination the relative gradients at random nearby points. The second difference is the updating and projection steps. The MFA-SGM uses stepsize τ_k in the updating step after the projection operation while the GS-SGM uses stepsize v_k in the updating step before the projection operation. This is the essential difference between the MFA-SGM and the GS-SGM. Note that, if the MFA-SGM starts from an (relative) interior point of X, then all iterates are relative interior points, which is the same property with the GS-SGM.

In the numerical computation, we use the same parameter $a = 0.1$ and stepsize $\tau_k = \tau/(1 + 0.01k)$ in the MFA-SGM as in [100]. For comparison, we choose the divergent stepsize rule $v_k = v/(1 + 0.01k)$ and other parameters $s = \frac{n}{2} = 50$, $\alpha_k = v_k$, $\delta_k = \alpha_k/2$, $\lambda_i = 1/s$ in the GS-SGM algorithm. Figure 2.1 plots the difference $f^k_{rec} - f_*$ when $\tau = 0.5$ in the MFA-SGM and when $v = 0.3$, 1, and 1.5 respectively in the GS-SGM until 3000 iterations. It is illustrated that the GS-SGM when $v = 1.5$ converges faster than the MFA-SGM when $\tau = 0.5$, but slower than the GS-SGM when $v = 1$.

[1]CVX, designed by Michael Grant and Stephen Boyd, is a Matlab-based modeling system for convex optimization. Detailed information is available at the website http://cvxr.com/cvx/.

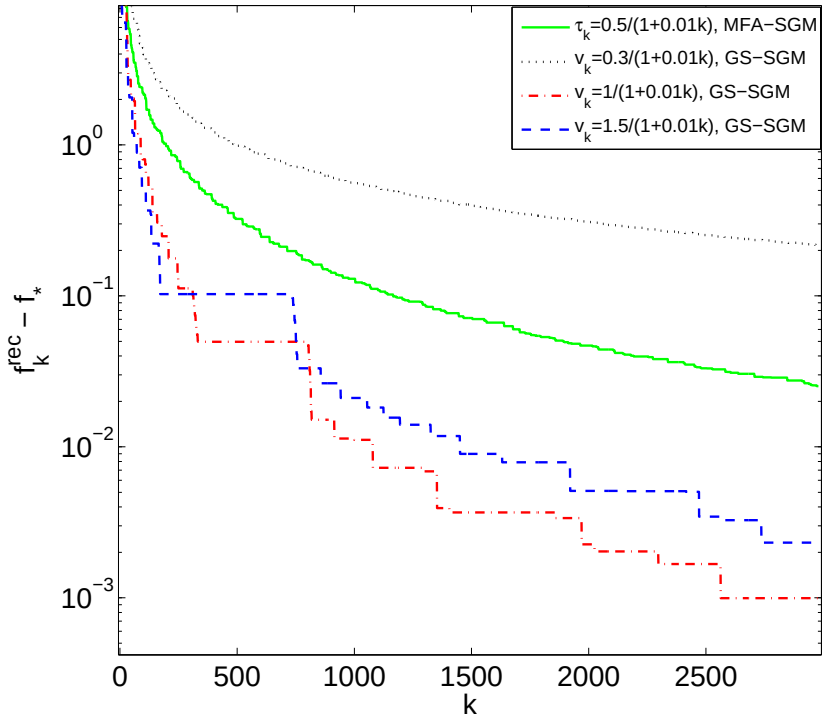


Figure 2.1: Comparison of MFA-SGM and GS-SGM

Assignment problem

The assignment problem is to assign m jobs to n machines such that the total cost is minimal (see [66, 80]). Job i, performed at machine j, costs a_{ij} and requires p_{ij} time units. Given the total available time t_j at each machine j, we want to find the minimum cost assignment of the jobs to the machines. Formally the problem can be written as

$$\begin{aligned} \min \quad & \textstyle\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} y_{ij} \\ \text{s.t.} \quad & \textstyle\sum_{j=1}^{n} y_{ij} = 1, \quad i = 1, \ldots, m, \\ & \textstyle\sum_{i=1}^{m} p_{ij} y_{ij} \leq t_j, \quad j = 1, \ldots, n, \\ & y_{ij} = 0 \text{ or } 1, \quad \text{for all } i, j, \end{aligned}$$

where y_{ij} is the assignment variable, which equals to 1 if the i-th job is assigned to the j-th machine and equals to 0 otherwise.

In the numerical experiment we choose n equal to 4 and m equal to 100. The data of the problem are randomly drawn from an uniform distribution on the unit interval, i.e.,

$$A = (a_{ij}) = \text{rand}(m, n), \quad P = (p_{ij}) = \text{rand}(m, n).$$

The value t_j is calculated according to the formula

$$t_j = \frac{\bar{t}}{n} \sum_{i=1}^{m} p_{ij}, \ j = 1, \ldots, n,$$

with $\bar{t}$ taking the value 0.4.

By relaxing the time constraints of machines, the dual problem is given by (see [80])

$$\begin{aligned} \max \quad & f(x) = \textstyle\sum_{i=1}^{m} f_i(x) \\ \text{s.t.} \quad & x \geq 0, \end{aligned} \tag{2.6.2}$$

where

$$f_i(x) = \min_{\sum_{j=1}^{n} y_{ij}=1, y_{ij}=0 \text{ or } 1} \sum_{j=1}^{n} (a_{ij} + x_j p_{ij}) y_{ij} - \frac{1}{m} \sum_{j=1}^{n} t_j x_j. \tag{2.6.3}$$

This is a concave maximization problem with a piecewise differentiable objective function. Since $a_{ij} + x_j p_{ij} \geq 0$ for all i, j, the objective value $f_i(x)$ is easily evaluated for each $x \geq 0$ by

$$f_i(x) = a_{ij^*} + x_j p_{ij^*} - \frac{1}{m} \sum_{j=1}^{n} t_j x_j,$$

where j^* is the index such that

$$a_{ij^*} + x_j p_{ij^*} = \min_{1 \leq j \leq n} \{a_{ij} + x_j p_{ij}\}.$$

Without additional cost, we obtain a subgradient g_i of f_i at x, whose j-th element is given by

$$(g_i)_j = \begin{cases} -\frac{t_j}{m}, & \text{if } j \neq j^*, \\ p_{ij^*} - \frac{t_{j^*}}{m}, & \text{if } j = j^*. \end{cases}$$

The subgradient method for solving the dual problem (2.6.2) is given by

$$x_{k+1} = P_{\mathbb{R}^m_+}[x_k + v_k \sum_{i=1}^{m} g_{i,k}],$$

where $g_{i,k}$ is a subgradient of f_i at x_k and v_k is the stepsize. While Nedić and Bertsekas [80] proposed an incremental subgradient method (IncSGM) to solve the dual problem (2.6.2) (see Section 1.1, Page 6).

In the numerical computation, we choose the divergent stepsize rule $v_k = 0.05/(1 + 0.01k)$ in both the IncSGM and the GS-SGM, and parameters $\alpha_k = v_k$, $\delta_k = \alpha_k/2$, $\lambda_i = 1/s$ and different sample size $s = 1, 5, 10$ in the GS-SGM. Figure 2.2 shows the

values of f_k in the IncSGM and in the GS-SGM when $s = 1, 5, 10$ until 300 iterates. It is illustrated that the GS-SGM with $s = 1$ performs almost the same as the IncSGM, while the GS-SGM with $s = 5, 10$ have better convergence behavior and obtain the better optimal value than the IncSGM.

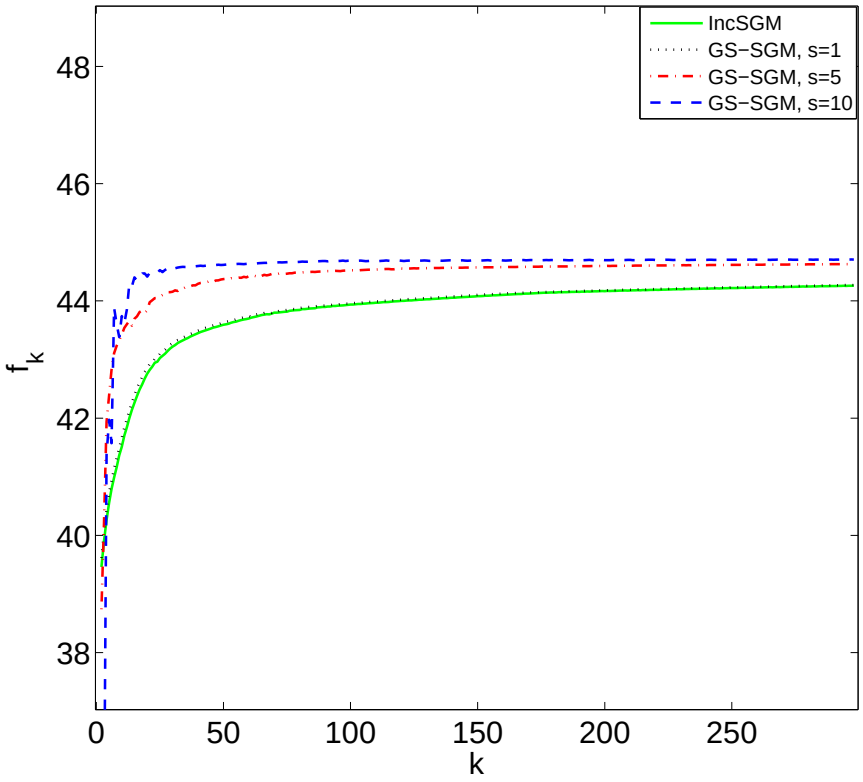


Figure 2.2: Comparison of IncSGM and GS-SGM

Low-rank recovery

In many engineering applications, the underlying data lies approximately on a low-dimensional linear subspace, hence the low-rank recovery problem has become an important issue in many applications in recent years. This problem can be stated as (see [96])

$$\begin{aligned} \min \quad & \mathrm{rank} Z \\ \text{s.t.} \quad & \mathcal{A}(Z) = b, \end{aligned} \tag{2.6.4}$$

where $Z \in \mathbb{R}^{m\times n}$ is the decision variable, the linear mapping $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ and vector $b \in \mathbb{R}^p$ are given. Denote $K := mn$, the linear mapping $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ can always be

written as its matrix representation, i.e.,

$$\mathcal{A}(Z) = AZ_{vec},$$

where $Z_{vec} \in \mathbb{R}^K$ denotes the "vectorized" Z with its columns stacked in order on top of one another, and A is a $p \times K$ matrix.

The basic idea for the low-rank recovery problem is to reformulate (2.6.4) as a nuclear norm minimization problem and solve it efficiently as a convex optimization problem. The corresponding nuclear norm minimization problem is given by

$$\begin{aligned} \min \quad & \|Z\|_* \\ \text{s.t.} \quad & \mathcal{A}(Z) = b. \end{aligned} \tag{2.6.5}$$

It is recalled that the nuclear norm of Z, denoted by $\|Z\|_*$, is defined as the sum of its singular values (see Section 1.3, Page 20). Let $Z = U\Sigma V^T$ be an SVD where $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$, and Σ is an $r \times r$ diagonal matrix of singular values. The subdifferential of the nuclear norm at Z is given by (see [71, 96])

$$\partial\|Z\|_* = \{UV^T + W : W \text{ and } Z \text{ have orthogonal row/column spaces and } \|W\| \leq 1\}.$$

When Z has no zero singular value (Z is full rank), the nuclear norm is differentiable and $\nabla\|Z\|_* = UV^T$.

(a) MIT logo (b) PolyU logo

Figure 2.3: The original MIT and PolyU logos.

We are interested in the logos of Massachusetts Institute of Technology (MIT) and The Hong Kong Polytechnic University (PolyU), which are shown in Figure 2.3. To make these logos low rank, we need to do some modifications on the two logos. The modified MIT logo has three distinct colors white, gray, and black, with rank equal

to 5, while the modification of PolyU logo is a little more complex. Since the original PolyU logo is almost full rank, we rotate it by 45 degrees and then make it low rank. We can see them in Figure 2.4. The modified PolyU logo has two distinct colors white and black, with rank equal to 9.

(a) MIT logo (b) PolyU logo

Figure 2.4: The modified MIT and PolyU logos.

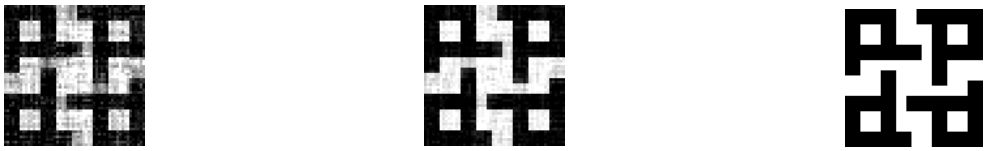
Consider the modified MIT and PolyU logos presented in Figure 2.4. The modified MIT logo has 46 rows and 81 columns (3726 elements), with three distinct values corresponding to white, gray, and black, while the modified PolyU logo has 60 rows and 60 columns (3600 elements), with two distinct values corresponding to white and black. For the linear mapping $\mathcal{A}$, we use the Gaussian ensemble and sample constraint matrice A with p ranging between 700 and 2400 in the numerical experiments.

Here, we use the classical subgradient method (SGM) and the GS-SGM$^{\mathrm{S}}$ to solve the nuclear norm minimization problem (2.6.5). In the numerical computation, we choose divergent stepsize rule $v_k = 1/(1 + 0.1k)$, $\alpha_k = v_k$, $\delta_k = \alpha_k/2$, $\lambda_i = 1/s$ and different sample size $s = 1, 5, 20, 50, 200$ in the GS-SGM$^{\mathrm{S}}$.

Figure 2.5 shows the recovered images for the modified MIT logo respectively using the Gaussian ensemble with $p = 800$, $p = 1100$ and $p = 1400$, while Figure 2.6 shows the recovered images for the modified PolyU logo respectively using the Gaussian ensemble with $p = 1000$, $p = 1500$ and $p = 2000$. Moreover, Figure 2.7 illustrates the errors, measured by the Frobenius norm, between the recovered image and the truth image for both MIT and PolyU logos until 1000 iterations in the GS-SGM$^{\mathrm{S}}$, as the number of constraints p changes from 700 to 2100. The numerical result verifies the conclusion, given by Recht in [96], that the perfect recovery is always attained when $p > 2r(m + n - r)$ for a rectangle image (like MIT logo), and when $p > n\sqrt{r(2n - r)}$ for a square image (like PolyU logo).

(a) 800 constraints (b) 1100 constraints (c) 1300 constraints

Figure 2.5: Recovered images for the modified MIT logo.



(a) 1000 constraints (b) 1500 constraints (c) 2000 constraints

Figure 2.6: Recovered images for the modified PolyU logo.

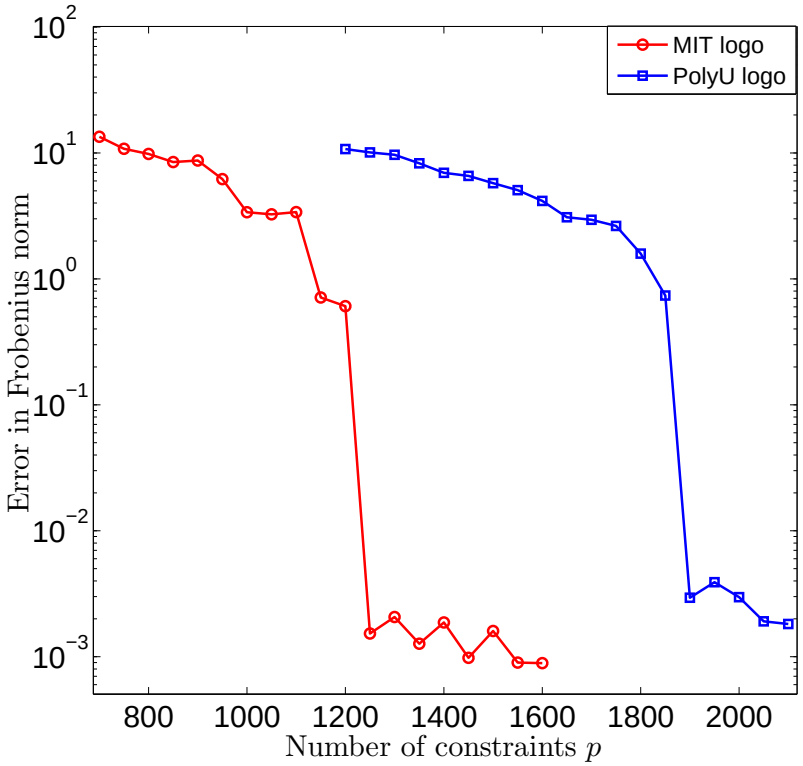


Figure 2.7: Error between the recovered image and the ground truth for both the modified MIT and PolyU logos.

Choosing numbers of constraints p of the MIT and PolyU logos described above, we display the numerical results for recovering the modified MIT logo as $p \geq 1300$ in Table 2.3 and the modified PolyU logo as $p \geq 2000$ in Table 2.4. In these tables, Δf denotes the required error of the objective value, and NIT and time denote the corresponding number and time of iterations needed to reach the specified precision of Δf respectively. It is illustrated in Tables 2.3-2.4 that the GS-SGM$^{\mathrm{S}}$ arrives at the required level in fewer iterations and less time than that of SGM. The GS-SGM$^{\mathrm{S}}$ meets the requirement, called for by Recht in [96], what extent subgradient methods can be efficiently applied to the nuclear norm minimization problem. When $s = 50$, it only costs one third or half of the time that is required for the SGM. Bigger sample size, better result? The answer is negative. From Tables 2.3-2.4, we observe that the number of iterations is less as the sample size increases. However, it costs more time to compute gradients when $s = 200$ and 500. As such the total computational time becomes large again when s is over 200.

Figure 2.8 and Figure 2.9 show the convergence behavior in objective values and iterates respectively for recovering the modified MIT logo as $p = 1500$, while Figure 2.10 and Figure 2.11 show the convergence behavior in objective values and iterates respectively for recovering the modified PolyU logo as $p = 2200$, using the SGM and the GS-SGM$^{\mathrm{S}}$. In these figures, $\Delta f_k = f_k - f_*$ denotes the error of objective values and $\Delta Z_k = \|Z_k - Z^*\|_F$ denotes the error of iterates in Frobenius norm.

Table 2.3: Computation results for recovering the modified MIT logo.

	$p = 1300$			$p = 1400$		
Algorithms/ sample size	Δf	NIT	time	Δf	NIT	time
SGM	0.3	369	11min	0.3	427	14min
GS-SGM$^{\mathrm{S}}$/ $s = 1$	0.3	427	13min	0.3	471	15min
GS-SGM$^{\mathrm{S}}$/ $s = 5$	0.3	295	9min	0.3	298	10min
GS-SGM$^{\mathrm{S}}$/ $s = 20$	0.3	254	8min	0.3	226	8min
GS-SGM$^{\mathrm{S}}$/ $s = 50$	0.3	241	8min	0.3	206	7min
GS-SGM$^{\mathrm{S}}$/ $s = 200$	0.3	234	10min	0.3	191	9min
GS-SGM$^{\mathrm{S}}$/ $s = 500$	0.3	212	13min	0.3	186	13min
	$p = 1500$			$p = 1600$		
SGM	0.3	460	16min	0.3	486	18min
GS-SGM$^{\mathrm{S}}$/ $s = 1$	0.3	449	16min	0.3	510	20min
GS-SGM$^{\mathrm{S}}$/ $s = 5$	0.3	313	11min	0.3	320	12min
GS-SGM$^{\mathrm{S}}$/ $s = 20$	0.3	223	8min	0.3	236	9min
GS-SGM$^{\mathrm{S}}$/ $s = 50$	0.3	207	8min	0.3	211	8min
GS-SGM$^{\mathrm{S}}$/ $s = 200$	0.3	191	10min	0.3	186	10min
GS-SGM$^{\mathrm{S}}$/ $s = 500$	0.3	187	13min	0.3	185	14min

Table 2.4: Computation results for recovering the modified PolyU logo.

	$p = 2000$			$p = 2100$		
Algorithms/ sample size	Δf	NIT	time	Δf	NIT	time
SGM	0.3	472	18min	0.3	495	19min
GS-SGM$^{\mathrm{S}}$/ $s = 1$	0.3	424	16min	0.3	421	17min
GS-SGM$^{\mathrm{S}}$/ $s = 5$	0.3	252	10min	0.3	225	9min
GS-SGM$^{\mathrm{S}}$/ $s = 20$	0.3	193	8min	0.3	188	8min
GS-SGM$^{\mathrm{S}}$/ $s = 50$	0.3	168	7min	0.3	163	7min
GS-SGM$^{\mathrm{S}}$/ $s = 200$	0.3	153	9min	0.3	149	8min
GS-SGM$^{\mathrm{S}}$/ $s = 500$	0.3	150	12min	0.3	143	11min
	$p = 2200$			$p = 2300$		
SGM	0.3	520	22min	0.3	509	23min
GS-SGM$^{\mathrm{S}}$/ $s = 1$	0.3	461	20min	0.3	397	17mim
GS-SGM$^{\mathrm{S}}$/ $s = 5$	0.3	252	11min	0.3	244	11min
GS-SGM$^{\mathrm{S}}$/ $s = 20$	0.3	183	8min	0.3	173	8min
GS-SGM$^{\mathrm{S}}$/ $s = 50$	0.3	158	7min	0.3	148	7min
GS-SGM$^{\mathrm{S}}$/ $s = 200$	0.3	141	8min	0.3	135	8min
GS-SGM$^{\mathrm{S}}$/ $s = 500$	0.3	138	11min	0.3	131	11min

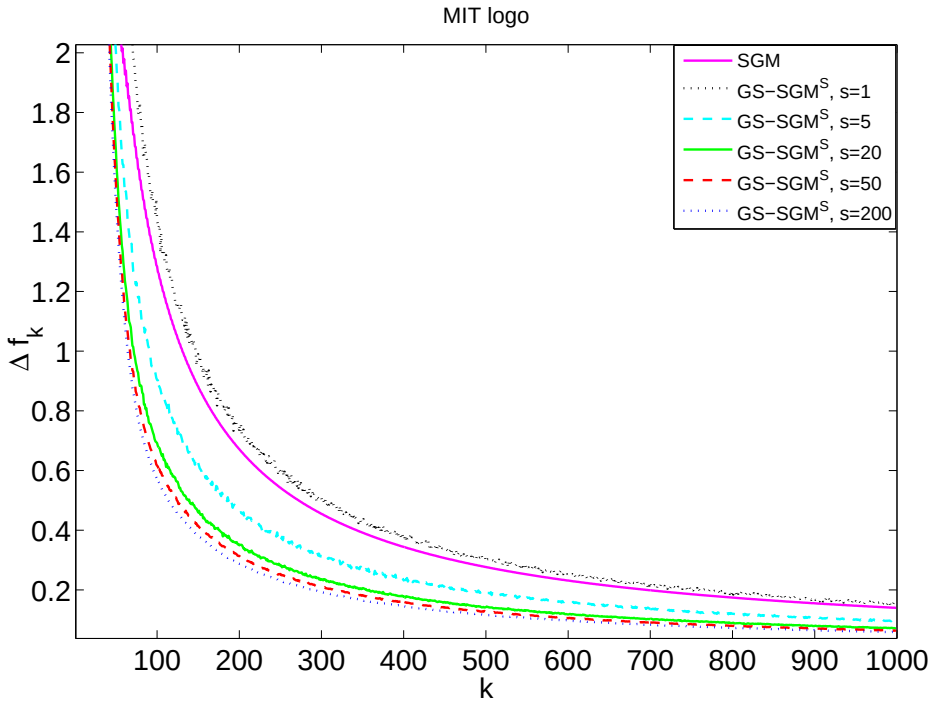


Figure 2.8: Convergence in objective values of the SGM and the GS-SGM$^{\mathrm{S}}$ for recovering the modified MIT logo as $p = 1500$.

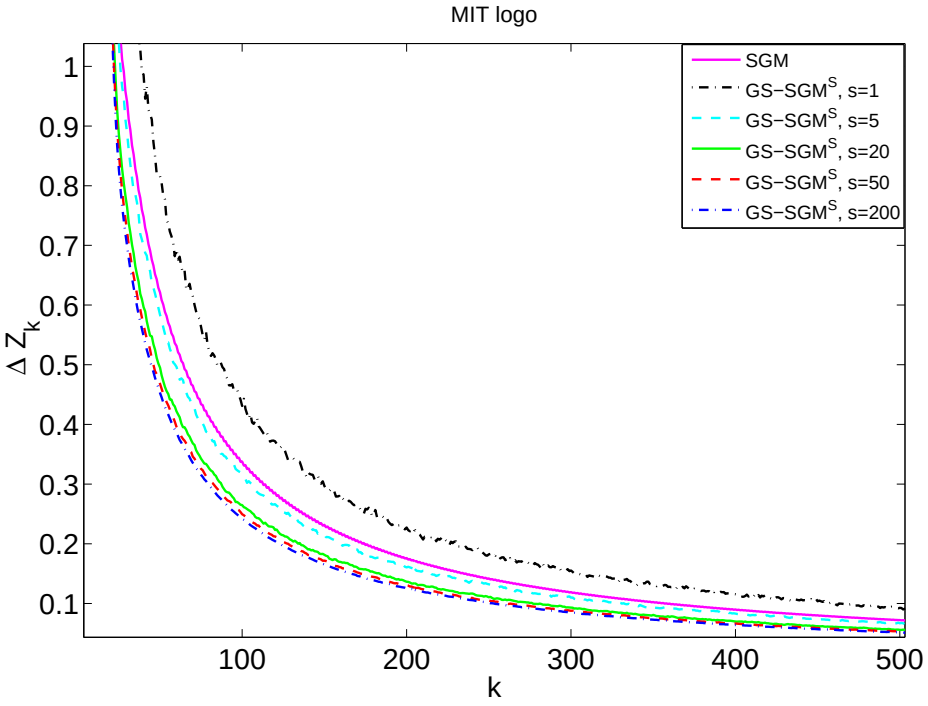


Figure 2.9: Convergence in iterates of the SGM and the GS-SGM$^{\mathrm{S}}$ for recovering the modified MIT logo as $p = 1500$.

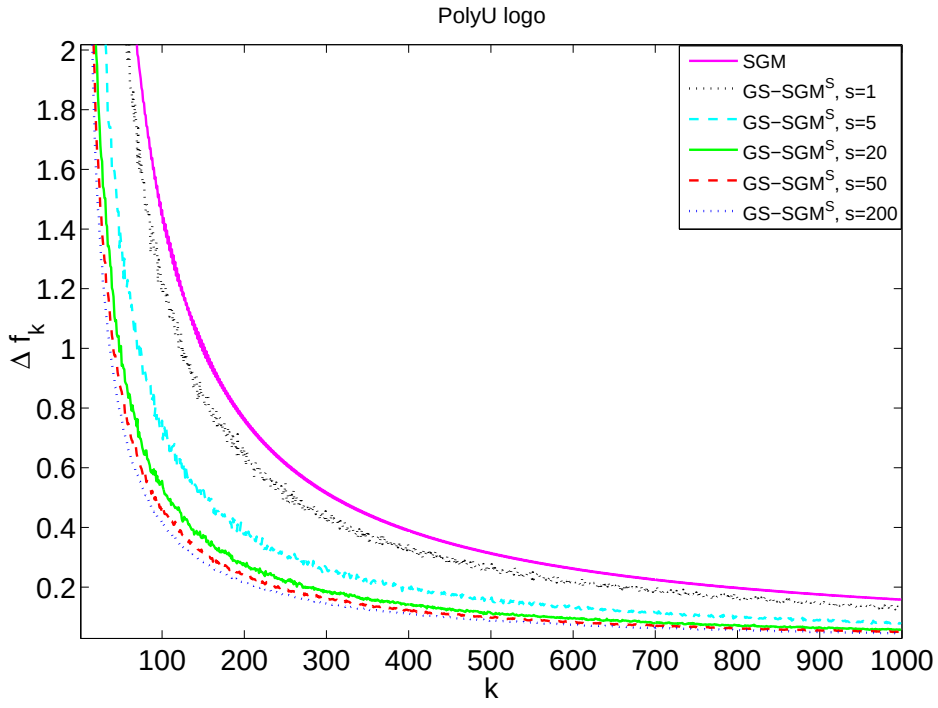


Figure 2.10: Convergence in objective values of the SGM and the GS-SGM$^{\mathrm{S}}$ for recovering the modified PolyU logo as $p = 2200$.

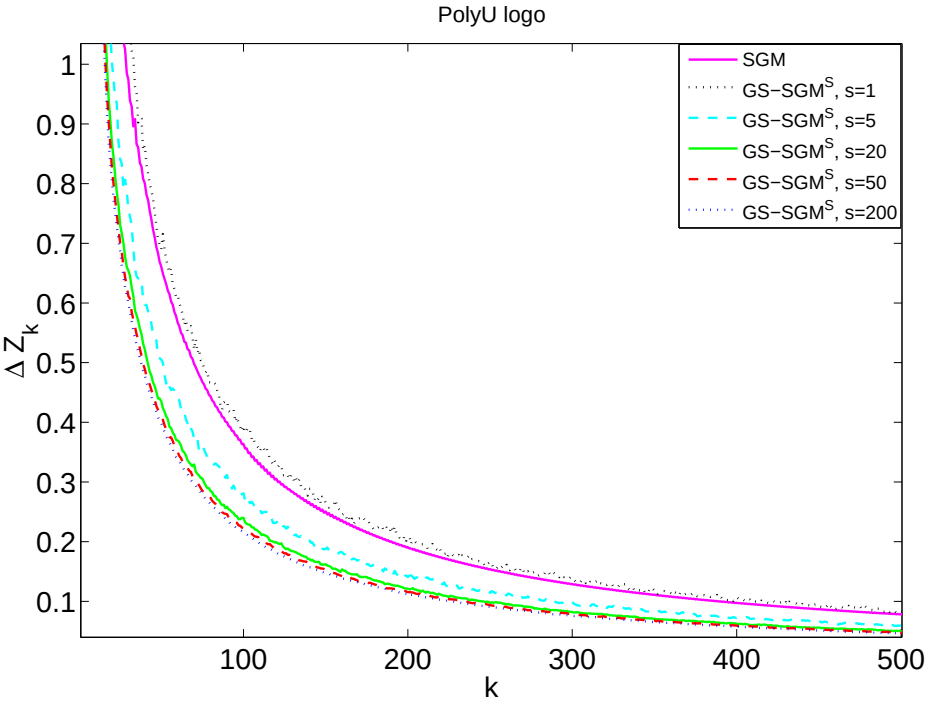


Figure 2.11: Convergence in iterates of the SGM and the GS-SGM$^{\mathrm{S}}$ for recovering the modified PolyU logo as $p = 2200$.

Chapter 3

Dual Subgradient Method Based on Gradient Sampling Technique

3.1 Introduction

Lagrangian relaxation and duality are powerful and effective tools for solving convex optimization problems and providing lower bounds on the optimal value of nonconvex optimization problems. Subgradient methods are popular and efficent techniques, used in this framework, to achieve dual optimal solutions and lower bounds on the primal optimal value. By combining the subgradient method with the dual approach, the dual subgradient method was proposed in Shor [106] and widely studied in [67, 78, 82, 104]. In these works, Sherali and Choi [104] focused on the linear programming, while Mijangos [78], Larsson [67], and Nedić and Ozdaglar [82] investigated the convex programming applied the dual subgradient method to solve the convex resource allocation problems in large-scale networks.

In this chapter, we consider the following convex constrained optimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq 0, \\ & x \in X, \end{aligned} \tag{3.1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, $g = (g_1, \cdots, g_m)^T$ with each $g_i : \mathbb{R}^n \to \mathbb{R}$ being convex, and X is a closed and convex subset of $\mathbb{R}^n$. We denote the optimal solution set

and the optimal value of problem (3.1.1) respectively by X^* and f_*.

Arising from the Lagrangian relaxation of the inequality constraints $g(x) \leq 0$, the dual problem of (3.1.1) is given by

$$\begin{array}{ll} \max & q(u) \\ \text{s.t.} & u \in \mathbb{R}^m_+, \end{array} \tag{3.1.2}$$

where $q : \mathbb{R}^m_+ \to \mathbb{R}$ is the dual function defined by

$$q(u) = \inf_{x \in X} \{f(x) + \langle u, g(x) \rangle\}. \tag{3.1.3}$$

We refer to the vector $u \in \mathbb{R}^m_+$ in (3.1.3) as a multiplier, and denote the dual optimal solution set and the dual optimal value of dual problem (3.1.2) respectively by U^* and q^*.

Due to the formula of the dual function (cf. (3.1.3)), q is a concave function on $\mathbb{R}^m$ and its subgradients at u are related to primal vectors x_u, which attain the minimum in (3.1.3). Formally, the subdifferential of q at $u \geq 0$ is given by

$$\partial q(u) = \text{conv}\{g(x_u), x_u \in X_u\}, \tag{3.1.4}$$

where X_u is the optimal solution set of the problem of minimizing $f(x) + \langle u, g(x) \rangle$ over $x \in X$, i.e.,

$$X_u = \{x_u \in X, q(u) = f(x_u) + \langle u, g(x_u) \rangle\}.$$

Incorporating the GS technique into the dual subgradient method, we propose a dual subgradient method based on the GS technique (in short, GS-DSGM) to solve dual problem (3.1.2). The main idea of the GS-DSGM is to construct the subgradient information in the dual space via the GS technique, that is to approach the subgradient by the convex combination of gradients at random nearby points. After the subgradient iteration along the constructed direction, the resulting point is projected onto the nonnegative orthant $\mathbb{R}^m_+$ and then performed a perturbation step.

The main objective of this chapter is to investigate convergence properties of the GS-DSGM by using both the constant and divergent stepsize rules, and to recover the approximate primal optimal solutions by using an averaging scheme. We indicate that the sequence generated by the GS-DSGM converges to the dual optimal value within the

tolerance $vM^2/2$, where M is an upper bound on gradients of the dual function, when a constant stepsize v is used (see Theorem 3.3.1), and converges to the dual optimal value by using the divergent stepsize rule (see Theorem 3.3.2), with probability 1. Moreover, we use an averaging scheme to recover primal optimal solutions by forming running averages of the primal vectors, which are generated by evaluating gradients of the dual function. Focusing on using the constant stepsize rule and the constant sampling radius, we provide the upper and lower bounds on the primal function value at the averaged vector, and give the upper and lower bounds on the amount of constraint violation at the averaged vector, per iteration.

This chapter is organized as follows. In Section 3.2, we propose the dual subgradient method based on the GS technique. In Section 3.3, we demonstrate convergence properties of the GS-DSGM by using both the constant and divergent stepsize rules. Finally, in Section 3.4, we recover the approximate primal optimal solutions by using an averaging scheme and provide bounds on their feasibility violation and primal function values.

3.2 GS-DSGM Algorithm

In this section, we propose a dual subgradient method based on the gradient sampling technique (GS-DSGM) to solve dual problem (3.1.2). This work is motivated by the constrained primal problems, which have a favorable dual problem structure leading to efficient implementation of dual subgradient methods, and by Remark 2.4.2 that the perturbation direction in the box constraint case is particularly easy to calculate. The main idea of the GS-DSGM is to construct the subgradient information in the dual space via the GS technique and then proceed the iterative process of the subgradient method. A perturbation step is needed to guarantee each iterate to be an interior point of $\mathbb{R}^m_+$. Therefore, the GS-DSGM consists of generating a multiplier sequence $\{u_k\}$, by carrying out the subgradient iteration, where the iterative direction is constructed via random sampling of gradients at nearby points in the dual space, and then projecting the resulting point onto $\mathbb{R}^m_+$ and finally performing a perturbation step.

In order to facilitate the reading and analysis of the algorithm, we provide a partial

glossary of the notation, used in the GS-DSGM, and present the GS-DSGM algorithm as follows.

Table 3.1: Notation used in GS-DSGM

k: Number of iterations.	s: Sample size.
v_k: Stepsize.	α_k: Perturbation weight.
ν_{ki}: Unit ball sample.	δ_k: Sampling radius.
u_k: Current iterate in dual space.	u_{ki}: Sampling point in dual space.
$\mathcal{D}$: Points of differentiability.	x_{ki}: Corresponding vector in primal space.
G_k: Approximate subdifferential.	g_k: Iterative direction.
λ_k: Iterative direction weight.	u_k^+: Projection point.
ω_k: Perturbation direction.	

Dual subgradient method based on the gradient sampling technique(GS-DSGM)

Step 1. (Initialization)

Start from $k = 0$, select an initial point $u_0 \in \mathbb{R}^m_{++}$ and parameters s, $\{v_k\}$ and $\{\alpha_k\}$.

Step 2. (Generate the approximate subdifferential via the GS technique)

Let $\nu_{k1}, \cdots, \nu_{ks}$ be sampled independently and uniformly from B, choose the sampling radius to satisfy $0 < \delta_k \leq \alpha_{k-1}$, and set

$$u_{ki} = u_k + \delta_k \nu_{ki}, \quad i = 1, \cdots, s.$$

If for some $i = 1, \cdots, s$, the point $u_{ki} \notin \mathcal{D}$, then STOP; otherwise, set

$$G_k = \text{conv}\{g(x_{k1}), \cdots, g(x_{ks})\},$$

where x_{ki} is the optimal solution of $q(u_{ki})$, i.e.,

$$x_{ki} = \arg\min_{x \in X}\{f(x) + \langle u_{ki}, g(x)\rangle\}.$$

Choose an arbitrary vector in G_k as the iterative direction, i.e.,

$$g_k = \sum_{i=1}^{s} \lambda_i g(x_{ki}),$$

and go to Step 3.

Step 3. (Solution update and perturbation)

Compute $u_{k+1}^+ = \max\{0, u_k + v_k g_k\}$, the projection onto the nonnegative orthant $\mathbb{R}_+^m$, and then proceed to perturb u_{k+1}^+ by an interior point of $\mathbb{R}_+^m$, i.e.,

$$u_{k+1} = (1 - \alpha_k)u_{k+1}^+ + \alpha_k \omega_k,$$

where

$$\omega_k \in \{u_{k+1}^+ - N_{\mathbb{R}_+^m}(u_{k+1}^+)\} \cap B(u_{k+1}^+, 1). \tag{3.2.1}$$

Set $k = k + 1$ and go back to Step 2.

Remark 3.2.1 *It follows from Remark 2.4.2 that the perturbation direction ω_k has the following analytical formula:*

$$\omega_k = \frac{sign(u_{k+1}^+) - e}{\sqrt{m}}.$$

Since the dual function q is concave, it is differentiable almost everywhere on $\mathbb{R}_+^m$. Thus, by using the same argument as in the proof of Theorem 2.3.1, it is easy to see that the GS-DSGM will not terminate finitely in Step 2, that is, a sequence of infinite multipliers $\{u_k\}$ will be generated by the GS-DSGM, with probability 1.

3.3 Convergence Analysis

In this section, we investigate convergence properties of the GS-DSGM.

Since the dual function q is a concave function, it follows from Lemmas 2.3.1-2.3.2 that the iterative direction $g_k \in G_k$, generated in Step 2, is an approximate subgradient of q at u_k. Indeed, when $u_{ki} \in \mathcal{D}$, it follows from (3.1.4) that $g(x_{ki})$ is a gradient (subgradient) of f at u_{ki}. By using Lemma 2.3.1, we obtain that $g(x_{ki})$ is an ϵ_{ki}-subgradient of q at u_k with $\epsilon_{ki} = q(u_{ki}) - q(u_k) - \langle g(x_{ki}), \delta_k \nu_{ki} \rangle$. Furthermore, from Lemma 2.3.2, it follows that the convex combination $g_k = \sum_{i=1}^{s} \lambda_i g(x_{ki})$ is an ϵ_k-subgradient of q at u_k with $\epsilon_k = \sum_{i=1}^{s} \lambda_i \epsilon_{ki}$, that is,

$$\begin{aligned} \langle g_k, u - u_k \rangle &\geq q(u) - q(u_k) - \epsilon_k \\ &= q(u) - \sum_{i=1}^{s} \lambda_i q(u_{ki}) + \sum_{i=1}^{s} \lambda_i \langle g(x_{ki}), \delta_k \nu_{ki} \rangle, \forall u \in \mathbb{R}_+^m. \end{aligned} \tag{3.3.1}$$

Throughout the rest of this section, we use the following assumption to investigate convergence properties of the GS-DSGM.

Assumption 3.3.1 *The constraint set X in the primal problem* (3.1.1) *is bounded.*

Since constraint functions g_i are all convex on $\mathbb{R}^m$, they are all continuous on $\mathbb{R}^m$. Thus, the boundedness of X implies that the gradient of q is uniformly bounded on $\mathbb{R}^m_+$, i.e.,

$$\|g_{ki}\| \leq M, \forall k, i \in \mathbb{N}, \quad \text{with } M := \max_{x \in X} \|g(x)\|. \tag{3.3.2}$$

In the convergence analysis, we start with the basic inequality, which shows a significant property of the multiplier sequence $\{u_k\}$.

Lemma 3.3.1 *Suppose Assumption 3.3.1 holds and the multiplier sequence $\{u_k\}$ is generated by the GS-DSGM. Then for all $u \in \mathbb{R}^m_+$, we have*

$$\|u_{k+1} - u\|^2 \leq \|u_k - u\|^2 + 2v_k\big(q(u_k) - q(u)\big) + 4v_k\delta_k M + v_k^2 M^2 + \alpha_k^2. \tag{3.3.3}$$

Proof. According to the GS-DSGM, for all $u \in \mathbb{R}^m_+$, we have

$$\begin{aligned} \|u_{k+1} - u\|^2 &= \|(1-\alpha_k)u_{k+1}^+ + \alpha_k\omega_k - u\|^2 \\ &= \|u_{k+1}^+ - u - \alpha_k(u_{k+1}^+ - \omega_k)\|^2 \\ &= \|u_{k+1}^+ - u\|^2 - 2\alpha_k\langle u_{k+1}^+ - u, u_{k+1}^+ - \omega_k\rangle + \alpha_k^2\|u_{k+1}^+ - \omega_k\|^2. \end{aligned} \tag{3.3.4}$$

Due to the chosen rule (3.2.1), $\omega_k \in u_{k+1}^+ - N_{\mathbb{R}^m_+}(u_{k+1}^+)$ implies that

$$\langle u_{k+1}^+ - u, u_{k+1}^+ - \omega_k\rangle \geq 0, \forall u \in \mathbb{R}^m_+,$$

and $\omega_k \in B(u_{k+1}^+, 1)$ implies that $\|u_{k+1}^+ - \omega_k\| \leq 1$. Hence, the relation (3.3.4) reduces to

$$\begin{aligned} \|u_{k+1} - u\|^2 &\leq \|u_{k+1}^+ - u\|^2 + \alpha_k^2 \\ &\leq \|u_k + v_k g_k - u\|^2 + \alpha_k^2 \\ &= \|u_k - u\|^2 + 2v_k\langle g_k, u_k - u\rangle + v_k^2\|g_k\|^2 + \alpha_k^2, \forall u \in \mathbb{R}^m_+, \end{aligned} \tag{3.3.5}$$

where the second inequality follows from the nonexpansive property of the projection operator.

It follows from the preceding declaration that g_k is an ϵ_k-subgradient of q at u_k. Thus, by using (3.3.1) and (3.3.5), we obtain

$$\begin{aligned}
\|u_{k+1}-u\|^2 &\le \|u_k-u\|^2+2v_k\Big(\sum_{i=1}^{s}\lambda_i q(u_{ki})-q(u)\Big)\\
&\quad +2v_k\Big(\sum_{i=1}^{s}\lambda_i\langle g(x_{ki}),\delta_k\nu_{ki}\rangle\Big)+v_k^2\|g_k\|^2+\alpha_k^2\\
&\le \|u_k-u\|^2+2v_k\Big(\sum_{i=1}^{s}\lambda_i q(u_{ki})-q(u)\Big) \qquad (3.3.6)\\
&\quad +2v_k\delta_k M+v_k^2M^2+\alpha_k^2\\
&\le \|u_k-u\|^2+2v_k\big(q(u_k)-q(u)\big)\\
&\quad +2v_k\sum_{i=1}^{s}\lambda_i\big(q(u_{ki})-q(u_k)\big)+2v_k\delta_k M+v_k^2M^2+\alpha_k^2\\
&\le \|u_k-u\|^2+2v_k\big(q(u_k)-q(u)\big)\\
&\quad +2v_k\Big(\sum_{i=1}^{s}\lambda_i\langle g(x_k),\delta_k\nu_{ki}\rangle\Big)+2v_k\delta_k M+v_k^2M^2+\alpha_k^2\\
&\le \|u_k-u\|^2+2v_k\big(q(u_k)-q(u)\big)+4v_k\delta_k M+v_k^2M^2+\alpha_k^2,
\end{aligned}$$

where the second and fifty inequalities both follow from the bound on gradients M and sampling points in the unit ball, and the fourth inequality follows from $q(u_{ki})-q(u_k)\le\langle g(x_k),\delta_k\nu_{ki}\rangle$. Thus, we arrive at the basic inequality (3.3.3). ■

Constant stepsize rule

The convergence analysis of the GS-DSGM is similar to that of the GS-SGM in Section 2.3. We first describe the convergence property of the GS-DSGM by using the constant stepsize rule.

Theorem 3.3.1 *Let Assumption 3.3.1 hold. Suppose the multiplier sequence $\{u_k\}$ is generated by the GS-DSGM with the constant stepsize rule, $\delta_k\le\alpha_k$, and $\sum_{k=0}^{\infty}\alpha_k^2<+\infty$. Then, $\overline{\lim_{k\to\infty}}\,q(u_k)\ge q^*-vM^2/2$ with probability 1.*

Proof. By the proof of Theorem 2.3.1, GS-DSGM does not terminate finitely in Step 2 with probability 1. We now focus on the case when the GS-DSGM generates a sequence of infinite multipliers $\{u_k\}$.

According to the GS-DSGM and Lemma 3.3.1 with $v_k \equiv v$, for each $u \in \mathbb{R}^m_+$, we have

$$2v\big(q(u) - q(u_k)\big) \leq \|u_k - u\|^2 - \|u_{k+1} - u\|^2 + 4v\delta_k M + v^2M^2 + \alpha_k^2. \tag{3.3.7}$$

Summing (3.3.7) over $k = 0, \cdots, n$, we obtain

$$q(u) - \frac{\sum\limits_{k=0}^{n} q(u_k)}{n} \leq \frac{\|u_0 - u\|^2}{2nv} + 2M\frac{\sum\limits_{k=0}^{n}\delta_k}{n} + \frac{vM^2}{2} + \frac{\sum\limits_{k=0}^{n}\alpha_k^2}{2nv}. \tag{3.3.8}$$

By the assumptions, $\sum\limits_{k=0}^{\infty}\alpha_k^2 < +\infty$ implies $\lim\limits_{k\to\infty}\alpha_k = 0$. Furthermore, due to $\delta_k \leq \alpha_k$, we have $\lim\limits_{k\to\infty}\delta_k = 0$ and hence $\lim\limits_{n\to\infty}\frac{\sum_{k=0}^{n}\delta_k}{n} = 0$ (cf. Lemma 1.3.5). Thus, by using Lemma 1.3.5, the relation (3.3.8) implies

$$\begin{aligned}\overline{\lim_{k\to\infty}}\, q(u_k) &\geq \overline{\lim_{k\to\infty}}\,\frac{\sum\limits_{k=1}^{n} q(u_k)}{n}\\ &\geq q(u) - \frac{vM^2}{2}, \forall u \in \mathbb{R}^m_+.\end{aligned}$$

Therefore, we arrive at that $\overline{\lim\limits_{k\to\infty}}\, q(u_k) \geq q^* - vM^2/2$ with probability 1. ■

Divergent stepsize rule

The corresponding convergence result is indicated in the following theorem by using the divergent stepsize rule.

Theorem 3.3.2 *Let Assumption 3.3.1 hold. Suppose the multiplier sequence $\{u_k\}$ is generated by the GS-DSGM with divergent stepsize rule* (1.1.6)*, $\delta_k \leq \alpha_k$, and $\sum\limits_{k=0}^{\infty}\alpha_k^2 < +\infty$. Then, $\overline{\lim\limits_{k\to\infty}}\, q(u_k) = q^*$ with probability 1.*

Proof. By the proof of Theorem 2.3.1, GS-DSGM does not terminate finitely in Step 2 with probability 1. We now focus on the case when the GS-DSGM generates a sequence of infinite multipliers $\{u_k\}$.

According to the GS-DSGM and Lemma 3.3.1, for each $u \in \mathbb{R}^m_+$, we have

$$2v_k\big(q(u) - q(u_k)\big) \leq \|u_k - u\|^2 - \|u_{k+1} - u\|^2 + 4v_k\delta_k M + v_k^2M^2 + \alpha_k^2. \tag{3.3.9}$$

Summing (3.3.9) over $k = 0, \cdots, n$, we obtain

$$q(u) - \frac{\sum\limits_{k=0}^{n} v_k q(u_k)}{\sum\limits_{k=0}^{n} v_k} \leq \frac{\|u_0 - u\|^2}{2\sum\limits_{k=0}^{n} v_k} + 2M\frac{\sum\limits_{k=0}^{n} v_k \delta_k}{\sum\limits_{k=0}^{n} v_k} + \frac{M^2 \sum\limits_{k=0}^{n} v_k^2 + \sum\limits_{k=0}^{n} \alpha_k^2}{2\sum\limits_{k=0}^{n} v_k}. \tag{3.3.10}$$

Similar to the proof of Theorem 2.3.2, by the assumptions of parameter rules, we obtain the following results:

$$\begin{cases} \lim\limits_{n\to\infty} \frac{\|u_0-u\|^2}{\sum\limits_{k=0}^{n} v_k} = 0, \\ \lim\limits_{n\to\infty} \frac{\sum\limits_{k=0}^{n} v_k \delta_k}{\sum\limits_{k=0}^{n} v_k} = 0, \\ \lim\limits_{n\to\infty} \frac{M^2 \sum\limits_{k=0}^{n} v_k^2 + \sum\limits_{k=0}^{n} \alpha_k^2}{\sum\limits_{k=0}^{n} v_k} = 0. \end{cases}$$

Substituting the above three relations into (3.3.10), we obtain

$$\begin{aligned} \overline{\lim_{k\to\infty}}\, q(u_k) &\geq \overline{\lim_{k\to\infty}} \frac{\sum\limits_{k=1}^{n} v_k q(u_k)}{\sum\limits_{k=1}^{n} v_k} \\ &\geq q(u), \forall u \in \mathbb{R}_+^m, \end{aligned}$$

where the first inequality follows from Lemma 1.3.5. Therefore, we arrive at that $\overline{\lim\limits_{k\to\infty}}\, q(u_k) = q^*$ with probability 1. ■

3.4 Approximate Primal Optimal Solutions

We have demonstrated the convergence property of the sequence $q(u_k)$ in the preceding Theorems. However, a defect of the GS-DSGM is that it cannot directly generate the primal optimal solution, which is inherited from the dual subgradient method.

In this section, to conquer this obstacle, we use an averaging scheme to recover the approximate optimal solutions of problem (3.1.1). Using the averaging scheme to recover the primal optimal solution was proposed by Nemirovskii and Yudin [84] and then developed in many works (see e.g. [67, 104, 82]). Due to the convexity of the objective function and constraint functions, the averaging scheme can reduce the primal function value and the amount of constraint violation at primal vectors.

In view of simplicity and practical significance, we focus on the constant stepsize, the constant sampling radius and the constant perturbation weight rules throughout this section, i.e.,

$$v_k \equiv v, \quad \alpha_k \equiv \alpha, \quad \text{and} \quad \delta_k \equiv \alpha.$$

Thus, for all $u \in \mathbb{R}^m_+$, the basic inequality (3.3.6) reduces to

$$\|u_{k+1} - u\|^2 \leq \|u_k - u\|^2 + 2v\Big(\sum_{i=1}^{s} \lambda_i q(u_{ki}) - q(u)\Big) + (vM + \alpha)^2. \tag{3.4.1}$$

For sequences $\{\nu_{ki}\}$, $\{u_{ki}\}$, $\{x_{ki}\}$ generated in the GS-DSGM, we denote

$$\bar{\nu}_k := \sum_{i=1}^{s} \lambda_i \nu_{ki}, \quad \bar{u}_k := \sum_{i=1}^{s} \lambda_i u_{ki}, \quad \text{and} \quad \bar{x}_k := \sum_{i=1}^{s} \lambda_i x_{ki}, \tag{3.4.2}$$

where $\{\lambda_k\}$ is the iterative direction weight, used in the GS-DSGM.

In the averaging scheme, the average of vectors $\bar{x}_0, \cdots, \bar{x}_{k-1}$ is defined by

$$\hat{x}_k = \frac{1}{k} \sum_{j=0}^{k-1} \bar{x}_j, \forall k \geq 1. \tag{3.4.3}$$

Since X is convex and $x_{ki} \in X$ for all i and k, $\bar{x}_k$ lies in X and thus the averaged vector $\hat{x}_k$ lies in X. However, the averaged vector $\hat{x}_k$ may not satisfy the constraint inequalities $g(x) \leq 0$, and therefore, it can be primal infeasible.

In the following theorem, we provide the upper and lower bounds on the amount of constraint violation and the primal function value at averaged vector per iteration.

Theorem 3.4.1 *Let Assumption 3.3.1 hold. Suppose the multiplier sequence $\{u_k\}$ is generated by the GS-DSGM, and $\{\hat{x}_k\}$ is the sequence of averaged vectors given in* (3.4.3)*. Then, for all $k \geq 1$, the following statements are true:*

(i) *an upper bound on the amount of constraint violation at the averaged vector $\hat{x}_k$ is given by*

$$\|g(\hat{x}_k)^+\| \leq \frac{\|u_k\|}{kv},$$

(ii) *an upper bound on the primal function value at the averaged vector $\hat{x}_k$ is given by*

$$f(\hat{x}_k) \leq q^* + \frac{\|u_0\|^2}{2kv} + \frac{(vM + \alpha)^2}{2v},$$

(iii) *a lower bound on the primal function value (a lower bound on the amount of constraint violation) at the averaged vector $\hat{x}_k$ is given by*

$$f(\hat{x}_k) \geq q^* - \|u^*\|\|g(\hat{x}_k)^+\|.$$

Proof.

(i) According to the GS-DSGM, for all $k \in \mathbb{N}$, we have

$$\begin{aligned} u_{k+1} &= (1-\alpha)u_{k+1}^+ + \alpha\omega_k \\ &= u_{k+1}^+ + \alpha(\omega_k - u_{k+1}^+) \\ &\geq u_k + vg_k + \alpha(\omega_k - u_{k+1}^+). \end{aligned}$$

Since $g_k = \sum_{i=1}^{s} \lambda_i g(x_{ki})$ and $\bar{x}_k = \sum_{i=1}^{s} \lambda_i x_{ki}$ (see (3.4.2)), it follows from the convexity of constraint functions that

$$\begin{aligned} vg(\bar{x}_k) &\leq vg_k \\ &\leq u_{k+1} - u_k - \alpha(\omega_k - u_{k+1}^+), \forall k \geq 0. \end{aligned}$$

Thus,

$$\begin{aligned} v\sum_{j=0}^{k-1} g(\bar{x}_j) &\leq u_k - u_0 - \alpha\sum_{j=0}^{k-1}(\omega_j - u_{j+1}^+) \\ &\leq u_k, \end{aligned}$$

where the second inequality holds due to $u_0 \geq 0$ and $\omega_j - u_{j+1}^+ \geq 0$ for all $j \in \mathbb{N}$, which follows from the chosen rule of the perturbation direction (see (3.2.1)). Since the constraint set X is convex and $x_{ki} \in X$ for all i and k, we obtain $\bar{x}_k \in X$ and thus $\hat{x}_k \in X$ for all k. Hence, from the convexity of g, it follows that

$$\begin{aligned} g(\hat{x}_k) &\leq \frac{1}{k}\sum_{j=0}^{k-1} g(\bar{x}_j) \\ &\leq \frac{u_k}{kv}, \forall k \geq 1. \end{aligned}$$

Moreover, due to $u_k \geq 0$ for all k, we arrive at $g(\hat{x}_k)^+ \leq u_k/(kv)$, and hence

$$\|g(\hat{x}_k)^+\| \leq \frac{\|u_k\|}{kv}, \forall k \geq 1.$$

(ii) Since the objective function is convex and the primal vector x_{ki} is a minimizer of the Lagrangian function $f(x) + \langle u_{ki}, g(x)\rangle$ over X, we obtain

$$\begin{aligned} f(\bar{x}_k) &\leq \sum_{i=1}^{s} \lambda_i f(x_{ki}) \\ &= \Big[\sum_{i=1}^{s} \lambda_i \Big(f(x_{ki}) + \langle u_{ki}, g(x_{ki})\rangle\Big)\Big] - \sum_{i=1}^{s} \lambda_i \langle u_{ki}, g(x_{ki})\rangle \\ &= \Big[\sum_{i=1}^{s} \lambda_i q(u_{ki})\Big] - \sum_{i=1}^{s} \lambda_i \langle u_k + \delta_k \nu_{ki}, g(x_{ki})\rangle \\ &= \Big[\sum_{i=1}^{s} \lambda_i q(u_{ki})\Big] - \langle u_k, g_k\rangle - \delta_k \sum_{i=1}^{s} \lambda_i \langle \nu_{ki}, g(x_{ki})\rangle \\ &\leq q^* + \alpha M - \langle u_k, g_k\rangle, \end{aligned}$$

where the last inequality follows from $q(u_{ki}) \leq q^*$, the bound on gradients M and sampling points in the unit ball. Thus, by the convexity of f and (3.4.3), we have

$$\begin{aligned} f(\hat{x}_k) &\leq \tfrac{1}{k} \sum_{j=0}^{k-1} f(\bar{x}_j) \\ &\leq q^* + \alpha M - \tfrac{1}{k} \sum_{j=0}^{k-1} \langle u_j, g_j\rangle. \end{aligned} \tag{3.4.4}$$

Next, we estimate the term $\sum_{j=0}^{k-1} \langle u_j, g_j\rangle$ in the last inequality.

By using (3.3.5) with $u = 0$, we obtain

$$\|u_{j+1}\|^2 \leq \|u_j\|^2 + 2v\langle g_j, u_j\rangle + v^2 M^2 + \alpha^2, \forall j \geq 0,$$

that is,

$$-\langle g_j, u_j\rangle \leq \frac{\|u_j\|^2 - \|u_{j+1}\|^2 + v^2 M^2 + \alpha^2}{2v}, \forall j \geq 0. \tag{3.4.5}$$

Summing (3.4.5) over $j = 0, \cdots, k-1$, we obtain

$$-\sum_{j=0}^{k-1} \langle g_j, u_j\rangle \leq \frac{\|u_0\|^2 - \|u_k\|^2 + kv^2 M^2 + k\alpha^2}{2v}, \forall k \geq 1.$$

Thus, the relation (3.4.4) reduces to

$$\begin{aligned} f(\hat{x}_k) &\leq q^* + \alpha M + \frac{\|u_0\|^2 - \|u_k\|^2 + kv^2 M^2 + k\alpha^2}{2kv} \\ &\leq q^* + \frac{\|u_0\|^2}{2kv} + \frac{(vM + \alpha)^2}{2v}, \forall k \geq 1. \end{aligned}$$

(iii) Given a dual optimal solution $u^* \in U^*$, we have

$$\begin{aligned} f(\hat{x}_k) &= f(\hat{x}_k) + \langle u^*, g(\hat{x}_k)\rangle - \langle u^*, g(\hat{x}_k)\rangle \\ &\geq q^* - \langle u^*, g(\hat{x}_k)\rangle. \end{aligned} \tag{3.4.6}$$

Since $u^* \geq 0$ and $g(\hat{x}_k)^+ \geq g(\hat{x}_k)$, we obtain

$$\begin{aligned}\langle u^*, g(\hat{x}_k)\rangle &\leq \langle u^*, g(\hat{x}_k)^+\rangle \\ &\leq \|u^*\|\|g(\hat{x}_k)^+\|.\end{aligned}$$

Thus, the relation (3.4.6) reduces to

$$f(\hat{x}_k) \geq q^* - \|u^*\|\|g(\hat{x}_k)^+\|.$$

■

Theorem 3.4.1 demonstrates that the upper and lower bounds on the amount of constraint violation $\|g(\hat{x}_k)^+\|$ and the primal value $f(\hat{x}_k)$ are available provided if the upper bounds on the multiplier norms $\|u_k\|$ and dual optimal solution norms $\|u^*\|$ are given.

In the rest of this section, under the Slater condition, we provide an upper bound on the multiplier sequence $\{u_k\}$ and then derive the upper and lower bounds on the amount of constraint violation $\|g(\hat{x}_k)^+\|$ and the primal value $f(\hat{x}_k)$. The well-known Slater condition is described as follows (see e.g. [16, 17, 46]).

Assumption 3.4.1 (Slater condition) *There exists a vector $\bar{x} \in X$ such that*

$$g_i(\bar{x}) < 0, \forall i = 1, \cdots, m.$$

When it exists, the vector $\bar{x}$ is called a Slater point.

Under the assumption that f_* is finite, it is well-known that the Slater condition is a sufficient condition for a zero duality gap (i.e., $f_* = q^*$), as well as, for the existence of a dual optimal solution (see [16, 17, 46]). Furthermore, the following lemma, cited in [82], extends the result on the boundedness of the dual optimal solution set under the Slater condition (see [46]), and shows that the Slater condition guarantees the boundedness of the superlevel set $\bar{U}_q(\bar{u})$ of q, i.e., $\bar{U}_q(\bar{u}) = \{u \in \mathbb{R}^m_+ : q(u) \geq q(\bar{u})\}$.

Lemma 3.4.1 ([82, Lemma 1]) *Let the Slater condition hold (cf. Assumption 3.4.1). Then the superlevel set $\bar{U}_q(\bar{u})$ is bounded, in particular, we have*

$$\max_{u \in \bar{U}_q(\bar{u})} \|u\| \leq \frac{1}{\gamma}\big(f(\bar{x}) - q(\bar{u})\big),$$

where $\gamma := \min\limits_{i=1,\cdots,m}\{-g_i(\bar{x})\}$ and $\bar{x}$ is a Slater point.

Restricting $\bar{u}$ to $u^* \in U^*$, it follows from Lemma 3.4.1 that U^* is nonempty and bounded, i.e.,

$$\max_{u\in U^*} \|u\| \leq \frac{1}{\gamma}\big(f(\bar{x}) - q^*\big), \tag{3.4.7}$$

where $\gamma := \min\limits_{i=1,\cdots,m}\{-g_i(\bar{x})\}$ and $\bar{x}$ is a Slater point.

In the following lemma, we establish the boundedness of the multiplier sequence $\{u_k\}$ under the bounded constraint set assumption and the Slater condition.

Lemma 3.4.2 *Let Assumptions 3.3.1 and 3.4.1 hold. Suppose the multiplier sequence $\{u_k\}$ is generated by the GS-DSGM. Then, the sequence $\{u_k\}$ is bounded, in particular, we have*

$$\|u_k\| \leq \frac{2}{\gamma}\big(f(\bar{x}) - q^*\big) + \max\Big\{\|u_0\|, \frac{1}{\gamma}\big(f(\bar{x}) - q^*\big) + \frac{(vM+\alpha)^2}{2v\gamma} + vM + 2\alpha\Big\}, \tag{3.4.8}$$

where $\gamma = \min\limits_{i=1,\cdots,m}\{-g_i(\bar{x})\}$ and $\bar{x}$ is a Slater point.

Proof. Given an arbitrary $u^* \in U^*$, we first claim that for all $k \geq 0$ there holds

$$\|u_k - u^*\| \leq \max\Big\{\|u_0 - u^*\|, \|u^*\| + \frac{1}{\gamma}\big(f(\bar{x}) - q^*\big) + \frac{(vM+\alpha)^2}{2v\gamma} + vM + 2\alpha\Big\}. \tag{3.4.9}$$

We prove the relation (3.4.9) by induction. Note that the relation (3.4.9) holds for $k = 0$. Assuming that the relation (3.4.9) holds for some $k \geq 0$, we now consider the following two cases.

Case 1. If $\sum\limits_{i=1}^{s} \lambda_i q(u_{ki}) \geq q^* - \frac{(vM+\alpha)^2}{2v}$. According to the GS-DSGM, by using (3.2.1) and the nonexpansive property of the projection operator, we obtain

$$\begin{aligned}
\|u_{k+1} - u^*\| &= \|(1-\alpha)u_{k+1}^+ + \alpha\omega_k - u^*\| \\
&= \|u_{k+1}^+ - u^* - \alpha(u_{k+1}^+ - \omega_k)\| \\
&\leq \|u_{k+1}^+ - u^*\| + \alpha\|u_{k+1}^+ - \omega_k\| \\
&\leq \|u_k + vg_k - u^*\| + \alpha \\
&\leq \|u_k\| + vM + \|u^*\| + \alpha \\
&= \|\bar{u}_k - \alpha\bar{\nu}_k\| + \|u^*\| + vM + \alpha \\
&= \|\bar{u}_k\| + \|u^*\| + vM + 2\alpha,
\end{aligned} \tag{3.4.10}$$

where the third equality follows from $\bar{u}_k = u_k + \delta_k\bar{\nu}_k$. By the concavity of q, we have

$$\begin{aligned}
q(\bar{u}_k) &\geq \sum_{i=1}^{s} \lambda_i q(u_{ki}) \\
&\geq q^* - \tfrac{(vM+\alpha)^2}{2v}.
\end{aligned}$$

Thus, by using Lemma 3.4.1, we have

$$\|\bar{u}_k\| \leq \frac{1}{\gamma}\big(f(\bar{x}) - q^*\big) + \frac{(vM+\alpha)^2}{2v\gamma},$$

and hence, the relation (3.4.10) implies

$$\|u_{k+1} - u^*\| \leq \|u^*\| + \frac{1}{\gamma}\big(f(\bar{x}) - q^*\big) + \frac{(vM+\alpha)^2}{2v\gamma} + vM + 2\alpha.$$

Therefore, the relation (3.4.9) holds for $k+1$ in this case.

Case 2. If $\sum_{i=1}^{s} \lambda_i q(u_{ki}) < q^* - \frac{(vM+\alpha)^2}{2v}$. From the basic inequality (3.4.1), it follows that

$$\begin{aligned}\|u_{k+1} - u^*\|^2 &\leq \|u_k - u^*\|^2 + 2v\Big(\sum_{i=1}^{s} \lambda_i q(u_{ki}) - q^*\Big) + (vM+\alpha)^2 \\ &< \|u_k - u^*\|^2 - (vM+\alpha)^2 + (vM+\alpha)^2 \\ &= \|u_k - u^*\|^2.\end{aligned}$$

By induction, it follows that the relation (3.4.9) holds for $k+1$ as well. Therefore, the relation (3.4.9) holds for all $k \geq 0$.

From the relation (3.4.9), for all $k \geq 0$, we arrive at

$$\begin{aligned}\|u_{k+1}\| &\leq \|u_{k+1} - u^*\| + \|u^*\| \\ &\leq \max\Big\{\|u_0 - u^*\|, \|u^*\| + \tfrac{1}{\gamma}\big(f(\bar{x}) - q^*\big) + \tfrac{(vM+\alpha)^2}{2v\gamma} + vM + 2\alpha\Big\} + \|u^*\| \\ &\leq \max\Big\{\|u_0\| + \|u^*\|, \|u^*\| + \tfrac{1}{\gamma}\big(f(\bar{x}) - q^*\big) + \tfrac{(vM+\alpha)^2}{2v\gamma} + vM + 2\alpha\Big\} + \|u^*\| \\ &= 2\|u^*\| + \max\Big\{\|u_0\|, \tfrac{1}{\gamma}\big(f(\bar{x}) - q^*\big) + \tfrac{(vM+\alpha)^2}{2v\gamma} + vM + 2\alpha\Big\} \\ &\leq \tfrac{2}{\gamma}\big(f(\bar{x}) - q^*\big) + \max\Big\{\|u_0\|, \tfrac{1}{\gamma}\big(f(\bar{x}) - q^*\big) + \tfrac{(vM+\alpha)^2}{2v\gamma} + vM + 2\alpha\Big\},\end{aligned} \tag{3.4.11}$$

where the last inequality follows from (3.4.7). Thus, we obtain an upper bound on the multiplier sequence as (3.4.8). ■

It is worth mentioning that the upper bound of the multiplier sequence $\{u_k\}$ depends on parameters in the GS-DSGM and problem data only. In particular, this upper bound (see (3.4.8)) is given explicitly in terms of the norm of initial point u_0, the stepsize v, the sampling radius α (perturbation weight), the bound on gradients M, the Slater point $\bar{x}$ and the dual optimal value q^*.

Under the bounded constraint set assumption and the Slater condition, applying relations (3.4.7) and (3.4.8) to Theorem 3.4.1, we provide the upper and lower bounds

on the amount of constraint violation $\|g(\hat{x}_k)^+\|$ and the primal function value $f(\hat{x}_k)$. Thus, we strengthen Theorem 3.4.1 as follows.

Theorem 3.4.2 *Let Assumptions 3.3.1 and 3.4.1 hold. Suppose the multiplier sequence* $\{u_k\}$ *is generated by the GS-DSGM, and* $\{\hat{x}_k\}$ *is the sequence of averaged vectors given in* (3.4.3)*. Also, define*

$$Q^* := \frac{2}{\gamma}\big(f(\bar{x}) - q^*\big) + \max\Big\{\|u_0\|, \frac{1}{\gamma}\big(f(\bar{x}) - q^*\big) + \frac{(vM+\alpha)^2}{2v\gamma} + vM + 2\alpha\Big\}. \quad (3.4.12)$$

Then, for all $k \geq 1$*, the following statements are true:*

(i) *an upper bound on the amount of constraint violation at the averaged vector* $\hat{x}_k$ *is given by*

$$\|g(\hat{x}_k)^+\| \leq \frac{Q^*}{kv},$$

(ii) *an upper bound on the primal function value at the averaged vector* $\hat{x}_k$ *is given by*

$$f(\hat{x}_k) \leq f_* + \frac{\|u_0\|^2}{2kv} + \frac{(vM+\alpha)^2}{2v},$$

(iii) *a lower bound on the primal function value (a lower bound on the amount of constraint violation) at the averaged vector* $\hat{x}_k$ *is given by*

$$f(\hat{x}_k) \geq f_* - \frac{1}{\gamma}\big(f(\bar{x}) - q^*\big)\|g(\hat{x}_k)^+\|.$$

Proof.

(i) Under assumptions 3.3.1 and 3.4.1, it follows from Lemma 3.4.2 that

$$\|u_k\| \leq Q^*, \forall k \in \mathbb{N}.$$

Applying Theorem 3.4.1(i), for all $k \geq 1$, we arrive at

$$\|g(\hat{x}_k)^+\| \leq \frac{Q^*}{kv}.$$

(ii) It is well-known that the Slater condition is sufficient for a zero duality gap, i.e., $f_* = q^*$. Thus, from Theorem 3.4.1(ii), it follows that

$$f(\hat{x}_k) \leq f_* + \frac{\|u_0\|^2}{2kv} + \frac{(vM+\alpha)^2}{2v}.$$

(iii) It follows from the Slater condition that the dual optimal set is nonempty and there is a zero duality gap, i.e., $f_* = q^*$. Also, it follows from (3.4.7) that

$$\|u^*\| \leq \frac{1}{\gamma}\big(f(\bar{x}) - q^*\big).$$

Thus, by applying Theorem 3.4.1(iii), we arrive at

$$f(\hat{x}_k) \geq f_* - \frac{1}{\gamma}\big(f(\bar{x}) - q^*\big)\|g(\hat{x}_k)^+\|.$$

■

For the sake of simplicity, we choose the initial point in the dual space near the origin. Thus, the upper bound Q^* in Theorem 3.4.2(i) reduces to

$$Q^* = \frac{3}{\gamma}\big(f(\bar{x}) - f_*\big) + \frac{(vM+\alpha)^2}{2v\gamma} + vM + 2\alpha, \tag{3.4.13}$$

and the estimate in Theorem 3.4.2(ii) reduces to

$$f(\hat{x}_k) \leq f_* + \frac{(vM+\alpha)^2}{2v}. \tag{3.4.14}$$

Applying the preceding two relations, we can estimate the order of the number of iterations required to achieve an both ϵ-feasible and ϵ-optimal solution.

In particular, to achieve the ϵ-optimality, from (3.4.14), it is to satisfy

$$\begin{aligned}\epsilon &\geq (vM+\alpha)^2/(2v)\\ &\geq (4v\alpha M)/(2v)\\ &= 2\alpha M.\end{aligned}$$

Thus, the sampling radius is required to satisfy $\alpha \leq \epsilon/(2M)$. We do not need any restriction on the value of stepsize, which is a surprising result and different from the corresponding result in [82].

Suppose the sampling radius is proportional to the stepsize, i.e., $\alpha = lv$. In this case, to achieve the ϵ-optimality, the stepsize is required to satisfy $v \leq 2\epsilon/(M+l)^2$. Furthermore, to achieve the ϵ-feasibility, from (3.4.13), it is to satisfy

$$\begin{aligned}\epsilon &\geq \frac{Q^*}{kv}\\ &= \frac{3}{kv\gamma}\big(f(\bar{x}) - f_*\big) + \frac{(vM+\alpha)^2}{2kv^2\gamma} + \frac{M}{k} + \frac{2\alpha}{kv},\\ &= \frac{3}{kv\gamma}\big(f(\bar{x}) - f_*\big) + \frac{(M+l)^2}{2k\gamma} + \frac{M+2l}{k}.\end{aligned}$$

Thus, the number of iterations should satisfy

$$\begin{aligned} k &\geq \frac{3}{v\gamma\epsilon}\big(f(\bar{x}) - f_*\big) + \frac{(M+l)^2}{2\gamma\epsilon} + \frac{M+2l}{\epsilon} \\ &\geq \frac{3}{2\gamma\epsilon^2}(M+l)^2\big(f(\bar{x}) - f_*\big) + \frac{(M+l)^2}{2\gamma\epsilon} + \frac{M+2l}{\epsilon}, \end{aligned}$$

where the last inequality follows from $v \leq 2\epsilon/(M+l)^2$ (ϵ-optimality). Hence, to achieve an both ϵ-feasible and ϵ-optimal solution, the number of the iterations is of the order $1/\epsilon^2$, which is typical for subgradient methods.

The significance of the averaging scheme can be interpreted as follows. It follows from Theorem 3.4.2(i) that the amount of constraint violation $\|g(\hat{x}_k)^+\|$ diminishes to zero as the number of iterations k tends to infinity. From the results in Theorem 3.4.2, we have that the limit of the function value $f(\hat{x}_k)$ lies in the range $[f_*, f_* + (vM + \alpha)^2/(2v)]$.

Part II

Subgradient Methods for Quasi-Convex Programming

Chapter 4

Approximate Quasi-Subgradient Method

4.1 Introduction

Subgradient methods are popular and powerful techniques used to minimize the nondifferentiable convex function. Motivated by practical reasons, approximate subgradient methods (also called ϵ-subgradient methods) were widely studied in [2, 36, 46, 61, 68, 106]. Kiwiel [61] proposed a unified convergence framework for approximate subgradient methods. The author presented convergence in both objective values and iterates, and gave efficiency estimates, using both the diminishing and nonvanishing stepsize rules. Larsson et al. [68] proposed and analyzed conditional ϵ-subgradient methods for solving convex constrained optimization problems and convex-concave saddle-point problems. In order to improve conditional subgradient methods, D'Antonios and Frangioni [36] combined the deflection and the conditional subgradient technique into one iteration, and investigated the unified convergence analysis for the deflected conditional ϵ-subgradient method, using both the dynamic and diminishing stepsize rules. Applying the dual approach, Mijangos [78] studied the approximate dual subgradient method to solve constrained network flow problems. Furthermore, Auslender and Teboulle [2] proposed and developed an interior ϵ-subgradient method for convex constrained optimization problems over polyhedral sets, in particular over $\mathbb{R}^n_+$, via replacing the

Euclidean distance function by a logarithmic-quadratic distance-like function.

Besides errors in approximate subgradient, the issue of noise on subgradient methods has been studied for convex constrained optimization problems. Polyak [93, 94] first studied the effect of noise, which was assumed to be deterministic and bounded, on subgradient methods for convex programming. Polyak presented the convergence property of the subgradient method with noise, using both the diminishing and Shor-type (i.e., $v_k = \alpha v^k$, where $\alpha > 0$ and $0 < v < 1$) stepsize rules. A interesting conclusion is that the sequence exactly converges to the optimal solution when the objective function has a unique sharp minimum and satisfies a linear growth property, even if the noise is nonvanishing.

Recently, Nedić and Bertsekas [81] studied the influence of errors and noise on subgradient methods for convex constrained optimization problems. When the constraint set is compact or the objective function had a set of weak sharp minima (see [28, 91]), the authors established convergence to the optimal value within some tolerance, which is expressed in terms of errors and noise, under the bounded subgradient assumption.

There are numerous papers in the subgradient method literature focusing on convex optimization problems. To meet much broader class of problems, Gasimov [40] and Burachik et al. [23, 24] proposed and developed an exact/inexact modified dual subgradient algorithm for a nonconvex optimization problem with equality constraints by virtue of a sharp augmented Lagrangian. Moreover, Kiwiel [60] studied convergence properties and efficiency estimates of the exact quasi-subgradient method to minimize a upper semi-continuous and quasi-convex function, using the diminishing stepsize rule.

In this chapter, we focus on an inexact subgradient algorithm, which we also call the approximate quasi-subgradient method (in short, AQSGM), for the following quasi-convex optimization problem:

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in X, \end{aligned} \tag{4.1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a quasi-convex function, and the constraint set X is nonempty, closed and convex. We denote the optimal solution set and the optimal value of problem (4.1.1) respectively by X^* and f_*, anddd assume that X^* is nonempty and compact.

Inspired by the idea in [81, 93, 94] and references therein, we investigate the effect

of inexact terms, including both computation errors and noise, on the AQSGM. The computation errors, which give rise to the ϵ-subgradient, is inevitable in computing process. On the other hand, the noise stems from practical considerations and applications, and is manifested in inexact computation of subgradients. Considering a generic inexact subgradient algorithm for the quasi-convex optimization problem (4.1.1) and assuming the inexact terms are deterministic and bounded, we establish convergence properties in both objective values and iterates with the tolerance given explicitly in terms of errors and noise. We also give finite convergence to the approximate optimal value and efficiency estimates of iterates.

Our investigation is divided into two cases: (i) X is compact and (ii) X is noncompact. When X is compact, we assume the Hölder condition of order $p > 0$, instead of the upper semi-continuity of the objective function used in [60], to obtain convergence properties in objective values (see Theorems 4.3.1 and 4.3.2) and finite convergence (see Theorems 4.3.3 and 4.3.3). We give Examples 4.3.3-4.3.4 to illustrate that the Hölder condition and the upper semi-continuity are independent of each other, and show that the Hölder condition of order 1 is equivalent to the bounded subgradient assumption, used in [81], for a convex function (see Lemma 4.3.1). Moreover, we describe the convergence property in iterates (see Theorem 4.3.5), which are absent in [81], although we need the additional assumption of the upper semi-continuity of the objective function. Different from this theorem, the lower semi-continuity of the objective function is assumed in [61] to prove convergence in iterates of subgradient method for convex optimization. When X is noncompact, we need to assume an additional generalized weak sharp minima condition (see Assumption 4.4.1). This condition extends the concept of weak sharp minima in [81] and is presented by using $\text{dist}(x, X^*)$, the distance of the decision variable x from X^*.

We also investigate the quantification of the effect of errors and noise by using both the constant and diminishing stepsize rules, while only the diminishing stepsize rule is considered in studying convergence properties and efficiency estimates of an exact quasi-subgradient method in Kiwiel [60].

We also illustrate the AQSGM on two numerical experiments, including the fractional programming and large-scale minimax linear fractional programming. The numerical experiments show that the exact quasi-subgradient method arrives at a better solution

in fewer number of iterations comparing with the level function method proposed by Xu [109] on small-scale problems, and that the AQSGM is suitable for large-scale problems. We also indicate the sensitivity of inexact terms, which coincides with the convergence results in this Chapter.

This chapter is organized as follows. In Section 4.2, we present some preliminaries of the quasi-subdifferential theory and the AQSGM algorithm, which is to be investigated in this chapter. In Section 4.3, we demonstrate convergence properties in both objective values and iterates, and finite convergence behavior of the AQSGM when the constraint set X is compact. In Section 4.4, we describe convergence behavior in both objective values and iterates and finite convergence when the objective function satisfies a generalized weak sharp minima condition over a noncompact set X. Finally, Section 4.5 gives the efficiency estimates and Section 4.6 illustrates some numerical results.

4.2 AQSGM Algorithm

Preliminaries of quasi-subdifferential theory

There are many different types of subdifferentials, such as Clarke-Rockafellar subdifferential, Dini subdifferential, Fréchet subdifferential and so on (see [7] and references therein). They are the same for convex functions, but different for nonconvex functions. Here, we introduce the Greenberg-Pierskalla subdifferential, defined by Greenberg and Pierskalla [43], as follows.

Definition 4.2.1 (see [43]) *The z-quasi-conjugate of f is a function $f_z^* : \mathbb{R}^n \to \bar{\mathbb{R}}$, defined by*

$$f_z^*(x) = z - \inf\{f(y) : \langle x, y\rangle \geq z\}.$$

It is recalled in [43, Theorem 1] that the z-quasi-conjugate function provides a lower bound for the corresponding convex conjugate function, and indeed, the convex conjugate function is the supremum of the z-quasi-conjugates over z.

Definition 4.2.2 (see [43]) *A Greenberg-Pierskalla subgradient of f at x is a vector $g \in \mathbb{R}^n$ such that*

$$f(x) + f^*_{\langle g,x\rangle}(g) = \langle g, x\rangle. \tag{4.2.1}$$

The set of Greenberg-Pierskalla subgradients of f at x is called the Greenberg-Pierskalla subdifferential of f at x and is denoted by $\partial^* f(x)$.

The following proposition gives an equivalent formula and some important properties of the Greenberg-Pierskalla subdifferential.

Proposition 4.2.1 ([43, Theorem 6]) *The following statements are true:*

(i) $\partial^* f(x) = \{g : \langle g, y - x\rangle < 0, \forall y \in S_f(x)\}$, *where $S_f(x)$ is the strict sublevel set, defined by $S_f(x) = \{y \in \mathbb{R}^n : f(y) < f(x)\}$,*

(ii) $\partial^* f(x)$ *is a convex cone,*

(iii) $0 \in \partial^* f(x)$ *if and only if* $x \in \arg\min f$.

Unfortunately, different from traditional subdifferentials, the Greenberg-Pierskalla subdifferential of f is not a closed set. Thus, in this chapter, following [46], we define the following closed set, which contains the closure of $\partial^* f(x)$, instead as the quasi-subdifferential, and use it in the inexact subgradient method.

Definition 4.2.3 *The quasi-subdifferential of f at x is defined by*

$$\bar{\partial}^* f(x) = \{g : \langle g, y - x\rangle \leq 0, \forall y \in S_f(x)\}. \tag{4.2.2}$$

When f is convex, the quasi-subdifferential coincides with the convex cone hull of the convex subdifferential (i.e., $\bar{\partial}^* f(x) = \text{cone}(\partial f(x))$, see [46, Chapter VI, Theorem 1.3.5]), and thus the inexact subgradient method (4.2.5) reduces to a normalized version of inexact subgradient method in [81]. When f is quasi-convex, the existence and relationship between the Greenberg-Pierskalla subdifferential and the quasi-subdifferential are described in the following lemma.

Lemma 4.2.1 *If f is quasi-convex on $\mathbb{R}^n$, then $\bar{\partial}^* f(x) \setminus \{0\} \neq \varnothing$. In addition, if f is upper semi-continuous on $\mathbb{R}^n$, then $\partial^* f(x) \neq \varnothing$ and $\bar{\partial}^* f(x)$ coincides with the closure of $\partial^* f(x)$, i.e., $\bar{\partial}^* f(x) = \partial^* f(x) \cup \{0\}$.*

Proof. If $S_f(x) = \varnothing$, then $\bar{\partial}^* f(x) = \partial^* f(x) = \mathbb{R}^n$ and the conclusions hold automatically. Now suppose $S_f(x) \neq \varnothing$. Since the convex sets $\{x\}$ and $S_f(x)$ are disjoint, it follows from Lemma 1.3.2 that there exists a proper hyperplane separation, i.e., there exists a vector $g \neq 0$ such that

$$\sup_{y \in S_f(x)} \langle g, y \rangle \leq \langle g, x \rangle \quad \text{and} \quad \inf_{y \in S_f(x)} \langle g, y \rangle < \langle g, x \rangle.$$

Thus, the vector g is a nonzero vector in $\bar{\partial}^* f(x)$. For the second conclusion, see [60, Lemma 3]. ■

The above lemma shows that the existence of nonzero quasi-subgradient only requires the quasi-convexity. Hence, throughout the rest of this chapter, we assume that the objective function is quasi-convex. In particular, we do not assume the upper semi-continuity of the objective function, which is used in [60], unless otherwise specified.

Motivated by practical reasons, relaxing (4.2.1) by $f(x) + f^*_{\langle g,x \rangle}(g) \leq \langle g, x \rangle + \epsilon$, we define the ϵ-quasi-subdifferential ($\epsilon > 0$) as follows.

Definition 4.2.4 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a quasi-convex function. The ϵ-quasi-subdifferential of f at x is defined by*

$$\bar{\partial}^*_\epsilon f(x) \ = \{g : \langle g, y - x \rangle \leq 0, \forall y \in S_{f,f(x)-\epsilon}\}, \tag{4.2.3}$$

where $S_{f,f(x)-\epsilon} = \{y \in \mathbb{R}^n : f(y) < f(x) - \epsilon\}$.

It follows from (4.2.3) that the ϵ-quasi-subdifferential is a closed and convex cone. Furthermore, the following proposition shows some basic properties and calculus rules of the ϵ-quasi-subdifferential.

Proposition 4.2.2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a quasi-convex function. The following statements are true:*

(i) $\bar{\partial}^*_\epsilon f(x) \subset \bar{\partial}^*_{\epsilon'} f(x)$, *whenever* $\epsilon \le \epsilon'$,

(ii) $\bar{\partial}^* f(x) = \bigcap_{\epsilon>0} \bar{\partial}^*_\epsilon f(x)$ $[= \lim_{\epsilon \downarrow 0} \bar{\partial}^*_\epsilon f(x)]$,

(iii) *for the function* $h(x) = f(x) + r$, *then* $\bar{\partial}^*_\epsilon h(x) = \bar{\partial}^*_\epsilon f(x)$,

(iv) *for the function* $h(x) = \alpha f(x)$ *and* $\alpha > 0$, *then* $\bar{\partial}^*_\epsilon h(x) = \bar{\partial}^*_{\epsilon/\alpha} f(x)$,

(v) *for the function* $h(x) = f(\alpha x)$ *and* $\alpha \neq 0$, *then* $\bar{\partial}^*_\epsilon h(x) = \bar{\partial}^*_\epsilon f(\alpha x)$,

(vi) *more generally, if* A *is an invertible* $n \times n$ *matrix, then* $\bar{\partial}^*_\epsilon (f \circ A)(x) = A^T \bar{\partial}^*_\epsilon f(Ax)$,

(vii) *for the function* $h(x) = f(x + x_0)$, *then* $\bar{\partial}^*_\epsilon h(x) = \bar{\partial}^*_\epsilon f(x + x_0)$,

(viii) *if* $f_1 \le f_2$ *and* $f_1(x_0) = f_2(x_0)$, *then* $\bar{\partial}^*_\epsilon f_1(x_0) \subset \bar{\partial}^*_\epsilon f_2(x_0)$.

Proof. All statements follow easily from the definition (4.2.3) and elementary calculus rules. Below, we give the proofs for (ii) and (vi), which may need some explanations.

(ii) From (i) and [98, Exercise 4.3], we obtain $\lim_{\epsilon \downarrow 0} \bar{\partial}^*_\epsilon f(x) = \bigcap_{\epsilon>0} \mathrm{cl}(\bar{\partial}^*_\epsilon f(x)) = \bigcap_{\epsilon>0} \bar{\partial}^*_\epsilon f(x)$. Given $g \in \bigcap_{\epsilon>0} \bar{\partial}^*_\epsilon f(x)$, we have $\langle g, y - x\rangle \le 0$ for all $y \in S_{f,f(x)-\epsilon}$ and all $\epsilon > 0$. Hence, $\langle g, y - x\rangle \le 0$ for all $y \in S_f(x)$, that is, $g \in \bar{\partial}^* f(x)$. Therefore, $\bigcap_{\epsilon>0} \bar{\partial}^*_\epsilon f(x) \subset \bar{\partial}^* f(x)$. The reverse inclusion follows directly from (i), and thus the equality holds.

(vi) It follows from (4.2.3) and the invertibility of A that

$$\begin{aligned}
\bar{\partial}^*_\epsilon (f \circ A)(x) &= \{g : \langle g, y - x\rangle \le 0, \forall y \text{ satisfying } (f \circ A)(y) < (f \circ A)(x) - \epsilon\} \\
&= \{g : \langle (A^{-1})^T g, Ay - Ax\rangle \le 0, \forall Ay \text{ satisfying } f(Ay) < f(Ax) - \epsilon\} \\
&= A^T \bar{\partial}^*_\epsilon f(Ax).
\end{aligned}$$

■

The above proposition extends some properties of the ϵ-subdifferential of convex functions (see [46, Chapter XI, (1.1.3) and Proposition 1.3.1]). Besides, the ϵ-subdifferential of convex function (cf. (1.3.2)) is a convex mapping on variable ϵ (see [46, Chapter XI, (1.1.4)]), i.e., for all positive scalars ϵ, ϵ' and $\alpha \in [0, 1]$ there holds

$$\alpha \partial_\epsilon f(x) + (1 - \alpha) \partial_{\epsilon'} f(x) \subset \partial_{\alpha\epsilon + (1-\alpha)\epsilon'} f(x). \tag{4.2.4}$$

However, the ϵ-quasi-subdifferential $\bar{\partial}^*_\epsilon f$ does not satisfy this property. An example is given as follows.

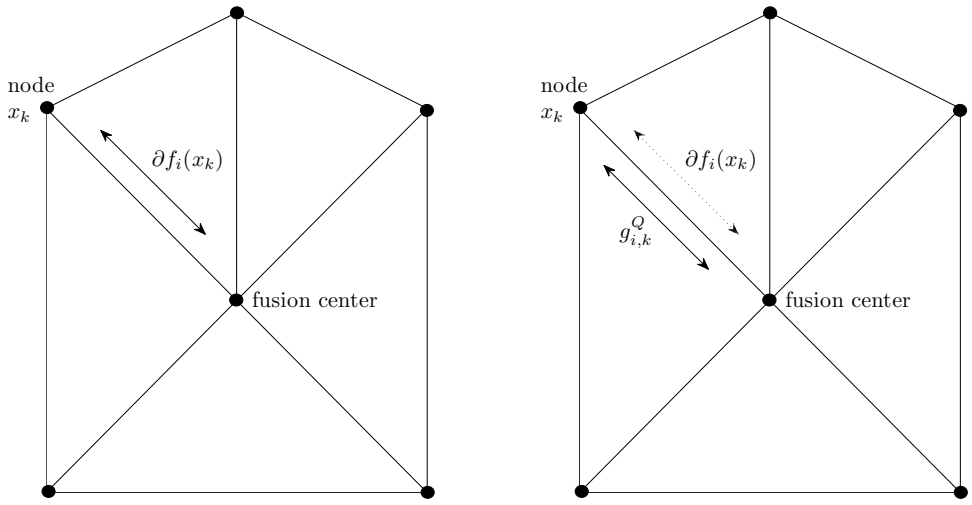
Example 4.2.1 *Consider the quasi-convex function*

$$f(x) := \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

It is easy to see $\bar{\partial}^*_\epsilon f(0) = \mathbb{R}_+$ *when* $\epsilon < 1$, *and* $\bar{\partial}^*_\epsilon f(0) = \mathbb{R}$ *when* $\epsilon \geq 1$. *However, when* $\epsilon = 2$, $\epsilon' = 1/2$ *and* $\alpha = 1/4$, *the relation* (4.2.4) *does not hold.*

AQSGM algorithm

We first describe the distributed optimization problem in networks consisting of m nodes and a fusion center (see e.g. [10, 54, 81, 95]), which is the motivating example of considering the noise in subgradient methods given by Nedić and Bertsekas [81].



(a) Unquantized model (b) Quantized model

Figure 4.1: The distributed optimization problem in networks.

Each node i has an objective function f_i known only at the node i, while the objective of the distributed optimization problem is to minimize $f(x) = \sum_{i=1}^{m} f_i(x)$ over a constraint set X. The fusion center is responsible for updating x_k and broadcasting

this information to the nodes in the network. In return, upon receiving x_k, each node i calculates a subgradient of its objective function f_i at x_k and sends the subgradient information to the fusion center (see Figure 4.1(a)).

However, in many applications, the links between the fusion center and the nodes can transmit only the quantized data (see e.g. [10, 54, 95]). For example, the quantization level is Q and every information x_k is quantized with respect to the level Q. Thus, the nodes receive the quantized data x_k^Q instead of the true information x_k. Then each node i calculates a subgradient $g_{i,k}$ of f_i at x_k^Q and sends it to the fusion center. Again, due to the quantization of the transmitted data, the fusion center only receives the quantized subgradient $g_{i,k}^Q$ and updates the information x_k using these quantized subgradients (see Figure 4.1(b)). Therefore, the approximate subgradient can be written as

$$g_k^Q = g_k + r_k \text{ with } g_k^Q = \sum_{i=1}^m g_{i,k}^Q,\ g_k = \sum_{i=1}^m g_{i,k},\ r_k = \sum_{i=1}^m (g_{i,k}^Q - g_{i,k}).$$

Since $g_{i,k} \in \partial f_i(x_k^Q)$, $g_k \in \partial f(x_k^Q)$ and thus g_k is an ϵ_k-subgradient of f at x_k with $\epsilon_k = f(x_k) - f(x_k^Q) - \langle g_k, x_k - x_k^Q \rangle$, which follows from Lemma 2.3.1. The noise r_k is deterministic and it is due to the quantization of the subgradients $g_{i,k}$. Both the error ϵ_k and noise norm $\|r_k\|$ are related to the quantization level Q.

Thus, in this chapter, we consider using a generic inexact subgradient method, which we also call the approximate quasi-subgradient method (AQSGM), to solve problem (4.1.1) as follows.

Approximate quasi-subgradient method (AQSGM)
Select a stepsize sequence $\{v_k\}$, an error sequence $\{\epsilon_k\}$ and a noise sequence $\{r_k\}$, start with an initial point $x_0 \in X$, and generate a sequence $\{x_k\} \in X$ via the iteration

$$x_{k+1} = P_X(x_k - v_k \tilde{g}_k), \tag{4.2.5}$$

where the iterative direction $\tilde{g}_k$ is an approximate quasi-subgradient of the following form

$$\tilde{g}_k := g_k / \|g_k\| + r_k, \tag{4.2.6}$$

where r_k is a noise vector and $g_k \in \bar{\partial}_{\epsilon_k}^* f(x_k)$ is an arbitrary nonzero ϵ_k-quasi-subgradient of f at x_k.

Let us first consider the following example, which says that the ϵ-quasi-subdifferential does not coincide with the quasi-subdifferential with noise.

Example 4.2.2 *Consider the quasi-convex function*

$$f(x,y) := \begin{cases} x^2+y^2, & x \geq 0, \\ y^2, & x < 0. \end{cases}$$

Its strict sublevel set $S_f(0,1) = S_{f,1}$ is illustrated in Figure 4.2, thus it is easy to see $\bar{\partial}^* f(0,1) = \text{cone}\{(0,1)\}$. Let the noise vector $r = (-\delta, 0)$ with $\delta > 0$. Then its quasi-subdifferential with noise and ϵ-quasi-subdifferential are respectively given by

$$\bar{\partial}^* f(0,1) + r = \{(-\delta, \lambda) : \lambda \in \mathbb{R}_+\},$$

and

$$\bar{\partial}^*_\epsilon f(0,1) = \begin{cases} \text{cone}\{(0,1), (\sqrt{\epsilon}, \sqrt{1-\epsilon})\}, & \epsilon < 1, \\ \mathbb{R}^2, & \epsilon \geq 1. \end{cases}$$

It is obvious that $(-\delta, 1) \notin \bar{\partial}^*_\epsilon f(0,1)$ for all $\delta > 0$ when $\epsilon < 1$. Thus, from this example, we see that the quasi-subdifferential with noise cannot be represented by the ϵ-quasi-subdifferential.

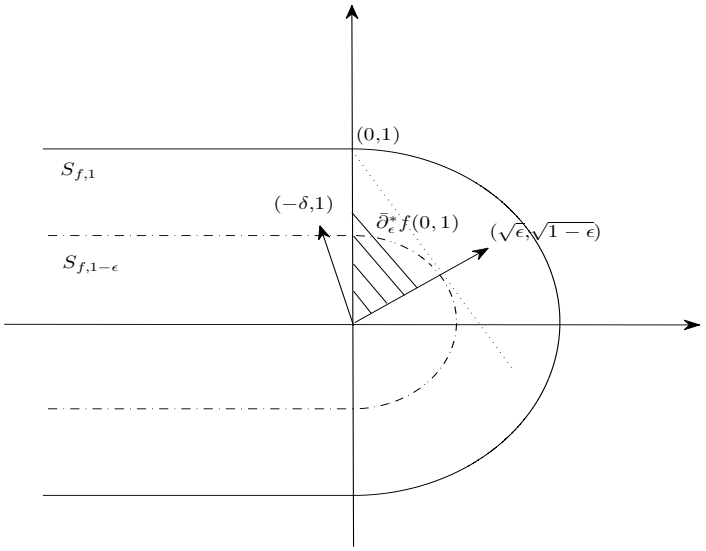


Figure 4.2: Illustration of Example 4.2.2.

It is well-known that the stepsize rule is critical in subgradient methods. In this chapter, assuming the noise and errors are deterministic and bounded, we investigate convergence properties of the AQSGM by using both the constant and diminishing stepsize rules.

4.3 Convergence Properties for a Compact X

In this section, we investigate convergence properties of the AQSGM when the constraint set X is compact. Throughout this section, the following three assumptions are made.

Assumption 4.3.1 *The constraint set X is compact.*

Assumption 4.3.2 *f satisfies the Hölder condition of order $p > 0$ with modulus $\mu > 0$ on $\mathbb{R}^n$, that is,*

$$f(x) - f_* \leq \mu\big(\mathrm{dist}(x, X^*)\big)^p, \forall x \in \mathbb{R}^n. \tag{4.3.1}$$

Assumption 4.3.3 *The noise and errors are bounded, i.e., there exist some scalars R and $\epsilon \geq 0$ such that*

$$\|r_k\| \leq R, \forall k \geq 0 \quad \text{and} \quad \overline{\lim_{k\to\infty}}\, \epsilon_k = \epsilon.$$

Since the constraint set X is compact, all iterates are bounded. Therefore, there exists some $d > 0$ (such as the diameter of X) such that $\|x_k - x\| \leq d$ for all $x \in X$ and $k \geq 0$. Moreover, under the bounded noise assumption, it follows from (4.2.6) that approximate quasi-subgradients are uniformly bounded, i.e., $\|\tilde{g}_k\| \leq 1 + R$ for all $k \geq 0$.

Konnov [64] used the Hölder condition of order p to describe the property of the quasi-subgradient. Here, we use this condition to investigate convergence properties of the AQSGM.

Nedić and Bertsekas [81] used the bounded subgradient assumption to study the influence of errors and noise on subgradient methods for convex optimization. The bounded subgradient assumption is quite natural in the subgradient method literature

(see [17, 61, 80, 106] and references therein). The following lemma shows that the bounded subgradient assumption is equivalent to the Hölder condition of order 1 when f is convex.

Lemma 4.3.1 *If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then the following two statements are equivalent:*

(i) *there is a positive scalar M such that $\|g\| \leq M$ for all $g \in \partial f(x)$ and $x \in R^n$,*

(ii) *f satisfies the Hölder condition of order 1 with modulus M on $\mathbb{R}^n$.*

Proof.

[(i) $\Rightarrow$ (ii)]: It follows from [32, Theorem 7.3] that (i) (uniformly bounded subgradient) is equivalent to the Lipschitz continuity of rank M, which implies (ii).

[(ii) $\Rightarrow$ (i)]: By contradiction, suppose there exist some $\tilde{x} \in \mathbb{R}^n$ and $\tilde{g} \in \partial f(\tilde{x})$ such that $\|\tilde{g}\| > M$. Then, by the Hölder condition, for any scalar $\lambda > \max\{\frac{\text{dist}(\tilde{x}, X^*)}{\|\tilde{g}\| - M}, \frac{f_* - f(\tilde{x})}{(\|\tilde{g}\| - M)^2}\}$, we have

$$\begin{aligned} f(\tilde{x} + \lambda\tilde{g}) - f_* &\leq M\text{dist}(\tilde{x} + \lambda\tilde{g}, X^*) \\ &\leq M(\text{dist}(\tilde{x}, X^*) + \lambda\|\tilde{g}\|) \\ &< M(\lambda(\|\tilde{g}\| - M) + \lambda\|\tilde{g}\|) \\ &= \lambda(2M\|\tilde{g}\| - M^2) \\ &= \lambda\|\tilde{g}\|^2 - \lambda(\|\tilde{g}\| - M)^2, \end{aligned} \tag{4.3.2}$$

where the second inequality follows from the triangular inequality, and the third inequality follows from $\lambda > \text{dist}(\tilde{x}, X^*)/(\|\tilde{g}\| - M)$. Moreover, since f is convex and $\tilde{g} \in \partial f(\tilde{x})$, we have

$$\begin{aligned} f(\tilde{x} + \lambda\tilde{g}) - f(\tilde{x}) &\geq \langle \tilde{g}, \lambda\tilde{g} \rangle \\ &= \lambda\|\tilde{g}\|^2 \\ &> f(\tilde{x} + \lambda\tilde{g}) - f_* + \lambda(\|\tilde{g}\| - M)^2 \\ &> f(\tilde{x} + \lambda\tilde{g}) - f_* + (f_* - f(\tilde{x})) \\ &= f(\tilde{x} + \lambda\tilde{g}) - f(\tilde{x}), \end{aligned}$$

where the second inequality follows from (4.3.2) and the third inequality follows from $\lambda > (f_* - f(\tilde{x}))/(\|\tilde{g}\| - M)^2$. Hence, we arrive at a contradiction. ■

However, the relationship between the Hölder condition of order 1 and the Lipschitz continuity does not hold for quasi-convex functions. The following Figure 4.3 shows an example of the quasi-conve function which satisfies the Hölder condition but is not Lipschitz continuous.

Example 4.3.1

$$f(x) = \begin{cases} 1, & x \leq 0, \\ x+1, & 0 < x < 1, \\ 3, & x \geq 1. \end{cases}$$

It is easy to check that the function satisfies the Hölder condition of order 1 with modulus 2, but it is not Lipschitz continuous at $x = 1$.

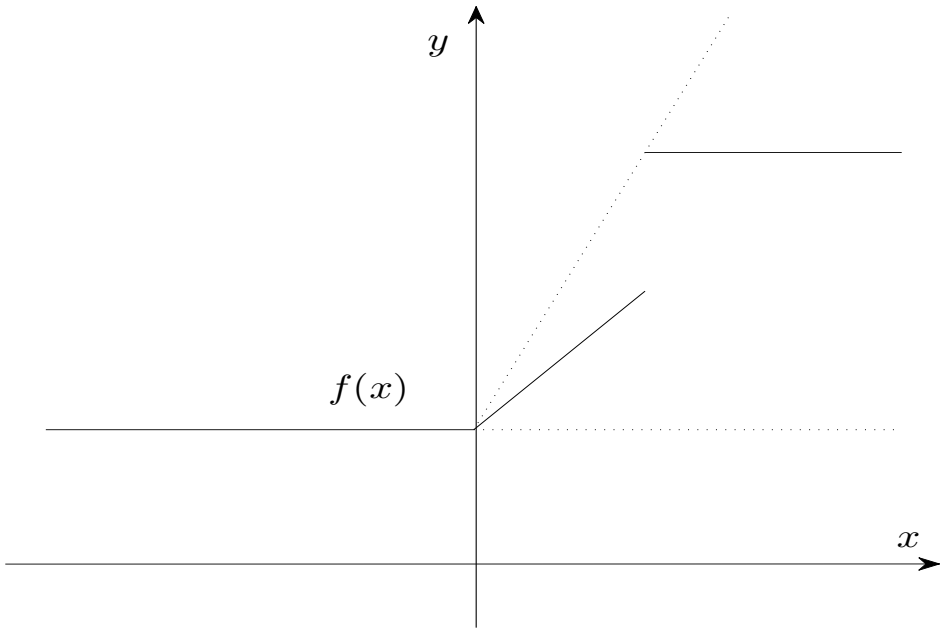


Figure 4.3: A function satisfies the Hölder condition but is not Lipschitz continuous.

Convergence in objective values

We now give the basic inequality and convergence properties in objective values using both the constant and diminishing stepsize rules. We start with the basic inequality, which shows a significant property of a subgradient iteration.

Lemma 4.3.2 *Suppose Assumptions 4.3.1 and 4.3.3 hold and the sequence* $\{x_k\}$ *is generated by the AQSGM. Then for all* $x \in X$*, we have*

$$\|x_{k+1} - x\|^2 \leq \|x_k - x\|^2 - 2v_k\Big(\langle g_k/\|g_k\|, x_k - x\rangle - Rd - \frac{1}{2}v_k(1+R)^2\Big), \forall k. \quad (4.3.3)$$

Proof. By relations (4.2.5)-(4.2.6) and the nonexpansive property of the projection operator, for all $x \in X$, we have the following basic inequality

$$\begin{aligned}
\|x_{k+1} - x\|^2 &= \|P_X(x_k - v_k\tilde{g}_k) - x\|^2 \\
&\leq \|x_k - v_k\tilde{g}_k - x\|^2 \\
&= \|x_k - x\|^2 - 2v_k\langle g_k/\|g_k\| + r_k, x_k - x\rangle + v_k^2\|g_k/\|g_k\| + r_k\|^2 \\
&\leq \|x_k - x\|^2 - 2v_k\Big(\langle g_k/\|g_k\|, x_k - x\rangle - Rd - \tfrac{1}{2}v_k(1+R)^2\Big),
\end{aligned}$$

where the last inequality follows from the compactness of X and the boundedness of noise. ■

The basic inequality (4.3.3) is quite different from that of convex subgradient method (see (1.1.4)). As shown in (1.1.4), the basic inequality describes the relationship between the function value and the distance from any point of the current iterate, while (4.3.3) does not describe any connection with function values. This difference originates from different definitions of subgradients. The convex subgradient is directly connected with function values, however, the quasi-convex subgradient coincides with the normal direction to its current sublevel set. Hence, we need to assume the Hölder condition to connect the quasi-convex subgradient with function values.

To make the thesis more self-contained, we cite and describe the following lemmas which are repeatedly used in the convergence analysis.

Lemma 4.3.3 ([60, Lemma 6]) *If* $B(\bar{x}, \bar{r}) \subset \mathrm{cl}S_{f,f(x_k)-\epsilon_k}$ *for some* $\bar{x} \in \mathbb{R}^n$ *and* $\bar{r} \geq 0$*, then* $\langle g_k/\|g_k\|, x_k - \bar{x}\rangle \geq \bar{r}$*.*

Lemma 4.3.4 *If Assumption 4.3.2 holds and* $f(x_k) > f_* + \mu\bar{r}^p + \epsilon_k$ *holds for some* $\bar{r} \geq 0$*, then* $\langle g_k/\|g_k\|, x_k - x^*\rangle \geq \bar{r}$ *for all* $x^* \in X^*$*.*

Proof. Given $x^* \in X^*$, by the Hölder condition of order p and the assumption given in the lemma, for all $x \in B(x^*, \bar{r})$, we have

$$\begin{aligned} f(x) - f_* &\leq \mu\big(\mathrm{dist}(x, X^*)\big)^p \\ &\leq \mu \bar{r}^p \\ &< f(x_k) - f_* - \epsilon_k, \end{aligned}$$

which implies $B(x^*, \bar{r}) \subset S_{f, f(x_k)-\epsilon_k}$. Hence, the conclusion follows from Lemma 4.3.3. ■

We first describe the convergence property of the AQSGM by using the constant stepsize rule.

Theorem 4.3.1 *Let Assumptions 4.3.1-4.3.3 hold. Then, for a sequence $\{x_k\}$ generated by the AQSGM with the constant stepsize rule, we have*

$$\underline{\lim}_{k\to\infty} f(x_k) \leq f_* + \mu(Rd + \frac{v}{2}(1+R)^2)^p + \epsilon.$$

Proof. We prove by contradiction, assuming that

$$\underline{\lim}_{k\to\infty} f(x_k) > f_* + \mu(Rd + \frac{v}{2}(1+R)^2)^p + \epsilon,$$

that is, there exist some $\delta > 0$ and positive integer k_0 such that

$$f(x_k) > f_* + \mu(Rd + \frac{v}{2}(1+R)^2 + \delta)^p + \epsilon_k, \forall k \geq k_0.$$

It follows from Lemma 4.3.4 that for all $x^* \in X^*$ and $k \geq k_0$ there holds

$$\langle g_k/\|g_k\|, x_k - x^* \rangle \geq Rd + \frac{v}{2}(1+R)^2 + \delta.$$

Therefore, by using the basic inequality (4.3.3) with $v_k \equiv v$ and $x = x^*$, we obtain

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 - 2v\Big(Rd + \tfrac{v}{2}(1+R)^2 + \delta - Rd - \tfrac{v}{2}(1+R)^2\Big) \\ &= \|x_k - x^*\|^2 - 2v\delta \\ &\leq \cdots \leq \|x_{k_0} - x^*\|^2 - 2(k - k_0 + 1)v\delta, \end{aligned}$$

which yields a contradiction for sufficiently large k. ■

In Assumption 4.3.2, we assume that f satisfies the Hölder condition on the whole space $\mathbb{R}^n$. Actually, this assumption is essential for the convergence result in Theorem 4.3.1. Relaxing it by the assumption that f satisfies the Hölder condition on the constraint set X cannot ensure the validity of Theorem 4.3.1, even if f is continuous on $\mathbb{R}^n$ as shown by the following example.

Example 4.3.2 *Consider the objective function*

$$f(u,v) := \begin{cases} M|v|, & u \le 0, \\ u + M|v|, & u > 0, \end{cases}$$

with $M = 100$ *and the constraint set* $X = \{(u,v) : -1 \le u \le 1, v = 0\}$. *Obviously, the optimal value of problem* (4.1.1) *is* $f_* = 0$ *and the optimal solution set is* $X^* = \{(u,v) : -1 \le u \le 0, v = 0\}$. *It is easy to check that* f *is continuous and quasi-convex on* $\mathbb{R}^2$ *and satisfies the Hölder condition (cf.* (4.3.1)*) on* X *with* $\mu = p = 1$.

Starting from $x_0 = (1,0)$, *we use the AQSGM to solve this problem. Specially, we choose the quasi-subgradient* $g = (1/\sqrt{1+M^2}, M/\sqrt{1+M^2}) \in \bar{\partial}^* f(x_0)$, *the noise vector* $r = (-1/\sqrt{1+M^2}, 0)$ *and the constant stepsize rule* $v = 1/2$, *then we have*

$$\begin{aligned} x_1 &= P_X\big(x_0 - v(g+r)\big) \\ &= P_X\big((1,0) - v(0, M/\sqrt{1+M^2})\big) \\ &= (1,0) = x_0. \end{aligned}$$

Hence, a fixed sequence is generated and $\lim_{k\to\infty} f(x_k) = f(x_0) = 1$. *However, when* $R = 0.01$, $\epsilon = 0$, $d = 2$ *and* $v = 1/2$, *the total error* $\mu(Rd + \frac{v}{2}(1+R)^2)^p + \epsilon < 1/2 < 1$. *Therefore, Theorem 4.3.1 fails for this problem.*

Using the diminishing stepsize rule, the error term involving the stepsize v in Theorem 4.3.1 vanishes and the following theorem is obtained.

Theorem 4.3.2 *Let Assumptions 4.3.1-4.3.3 hold. Then, for a sequence* $\{x_k\}$ *generated by the AQSGM with the diminishing stepsize rule, we have*

$$\underline{\lim}_{k\to\infty} f(x_k) \le f_* + \mu(Rd)^p + \epsilon.$$

Proof. By contradiction, we assume that

$$\underline{\lim}_{k\to\infty} f(x_k) > f_* + \mu(Rd)^p + \epsilon,$$

that is, there exist some $\delta > 0$ and positive integer k_0 such that

$$f(x_k) > f_* + \mu(Rd + \delta)^p + \epsilon_k, \forall k \ge k_0.$$

It follows from Lemma 4.3.4 that for all $x^* \in X^*$ and $k \geq k_0$,

$$\langle g_k/\|g_k\|, x_k - x^* \rangle \geq Rd + \delta.$$

Since the stepsize v_k diminishes, there exists some positive integer $k_\delta > k_0$ such that

$$v_k \leq \delta/(1+R)^2, \forall k \geq k_\delta. \tag{4.3.4}$$

Therefore, by using the basic inequality (4.3.3) with $v_k \equiv v$ and $x = x^*$, we obtain

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 - 2v_k\Big(Rd + \delta - Rd - \tfrac{v_k}{2}(1+R)^2\Big) \\ &\leq \|x_k - x^*\|^2 - v_k\delta \\ &\leq \cdots \leq \|x_{k_\delta} - x^*\|^2 - \delta \sum_{i=k_\delta}^{k} v_i, \end{aligned}$$

where the second inequality follows from (4.3.4). Thus, it yields a contradiction for sufficiently large k, since $\sum_{i=k_\delta}^{\infty} v_i = +\infty$. ■

Theorems 4.3.1-4.3.2 show convergence to the optimal value within some tolerance given in terms of errors and noise by using the constant and diminishing stepsize rules respectively. In Theorem 4.3.2, the total error $c := \mu(Rd)^p + \epsilon$, which is a similar formula as in [81], has an additive form, including the noise level R and the error level ϵ. By contrast, in Theorem 4.3.1, the total error additionally includes a term related to the constant stepsize v. In the presence of persistent noise ($R > 0$), the total error is not zero even if computation errors vanish ($\epsilon_k \equiv 0$).

Corollary 4.3.1 *Let Assumptions 4.3.1-4.3.3 hold. Then, for a sequence $\{x_k\}$ generated by the AQSGM with the diminishing stepsize rule and $\epsilon_k \equiv 0$, we have*

$$\underline{\lim}_{k\to\infty} f(x_k) \leq f_* + \mu(Rd)^p.$$

Considering another special case when the noise vanishes ($R = 0$), the AQSGM reduces to the ϵ-quasi-subgradient method. In this case, the term $\langle r_k, x_k - x\rangle$ vanishes in the corresponding basic inequality, and we obtain Lemma 4.3.2 (where $R = 0$) without the need for X to be compact. Therefore, when the noise vanishes, the convergence property holds regardless of compactness of X.

Corollary 4.3.2 *Let Assumptions 4.3.2-4.3.3 hold. Then, for a sequence $\{x_k\}$ generated by the AQSGM with the diminishing stepsize rule and $r_k \equiv 0$, we have*

$$\underline{\lim}_{k\to\infty} f(x_k) \leq f_* + \epsilon.$$

Corollary 4.3.3 *Let Assumption 4.3.2 hold. Then, for a sequence $\{x_k\}$ generated by the AQSGM with the diminishing stepsize rule, $r_k \equiv 0$ and $\epsilon_k \equiv 0$, we have*

$$\underline{\lim}_{k\to\infty} f(x_k) = f_*.$$

Note again that Corollaries 4.3.2-4.3.3 are obtained without the compactness assumption on X. The latter is the main result in [60], where the upper semi-continuity of f is assumed. Here, we have obtained the convergence result as in [60] without the upper semi-continuity assumption, but using the Hölder condition of order p instead. The following two examples show that the Hölder condition and upper semi-continuity are independent of each other.

Example 4.3.3 (The function satisfies the Hölder condition but is not upper semi-continuous.) *Consider the objective function*

$$f(x) := \begin{cases} 0, & x \leq 0, \\ x^2, & 0 < x \leq 1, \\ 2, & x > 1, \end{cases}$$

and the constraint set $X = \{x \in \mathbb{R} : 0 \leq x \leq 10\}$. Obviously, the optimal value of problem (4.1.1) is $f_* = 0$ and the optimal solution set is $X^* = \{0\}$. It is easy to verify that f is quasi-convex (since its sublevel sets are all convex) and satisfies the Hölder condition of order 2 with modulus 2 on $\mathbb{R}$. However, f is not upper semi-continuous at $x = 1$. Thus, this example shows that the Hölder condition does not imply the upper semi-continuity.

Thus, from [60], we cannot obtain the convergence property of the exact quasi-subgradient method (cf. (14)-(15) in [60]) for this example. However, the sequence generated by the exact quasi-subgradient method converges to X^*. Indeed, for any $x \in X \setminus X^*$, the strict sublevel set $S_f(x)$ is the line segment $[0, \min\{1, x\})$ and the

quasi-subdifferential $\bar{\partial}^* f(x) = \mathbb{R}_+$. *Therefore,*

$$\begin{aligned} x_{k+1} &= P_X(x_k - v_k g_k/\|g_k\|) \\ &= \max\{x_k - v_k, 0\}, \end{aligned}$$

and thus the sequence $\{x_k\}$ *converges to the origin, which is the optimal solution, due to the properties of the diminishing stepsize rule. This iterative result coincides with the result in Corollary 4.3.3.*

Example 4.3.4 (The function is upper semi-continuous but does not satisfy the Hölder condition.) *Consider the objective function*

$$f(x) = e^x,$$

and the constraint set R_+. *Obviously, the optimal value of problem* (4.1.1) *is* $f_* = 1$ *and the optimal solution set is* $X^* = \{0\}$. *It is easy to check that* f *is continuous and quasi-convex (since it is monotone) on* $\mathbb{R}$. *However, by the Taylor expansion* $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$, *we claim that* f *does not satisfy the Hölder condition on* $\mathbb{R}$ *for any positive scalars* p *and* μ. *Indeed, given positive scalars* p *and* μ, *when* $x \geq \exp(\frac{\log(\mu\lceil p+1\rceil!)}{\lceil p+1\rceil - p})$, *where* $\exp(\cdot)$ *and* $\lceil p\rceil$ *denote the exponential function and the largest integer not greater than* p *respectively, we have*

$$\begin{aligned} f(x) - f_* &= e^x - 1 \\ &> \frac{x^{\lceil p+1\rceil}}{\lceil p+1\rceil!} + \frac{x^0}{0!} - 1 \\ &= \frac{x^{\lceil p+1\rceil}}{\lceil p+1\rceil!} \\ &\geq \mu x^p, \end{aligned}$$

which contradicts with (4.3.1). *Thus, this example shows that upper semi-continuity does not imply the Hölder condition.*

Although, from [60], we obtain the convergence property of the exact quasi-subgradient method for this example. However, the convergence result of the AQSGM (see Theorem 4.3.2) fails for this example. Indeed, given positive scalars p *and* μ, *we consider the constraint set* $X = \{x \in \mathbb{R} : 0 \leq x \leq \exp(\frac{\log(\mu\lceil p+1\rceil!)}{\lceil p+1\rceil - p})\}$, *noise* $r_k \equiv -1$ *and errors* $\epsilon_k \equiv 0$. *For any* $x \in X \setminus X^*$, *the strict sublevel set* $S_f(x)$ *is the line segment* $[0, x)$ *and the quasi-subdifferential* $\bar{\partial}^* f(x) = \mathbb{R}_+$. *Thus, starting from* $x_0 = \exp(\frac{\log(\mu\lceil p+1\rceil!)}{\lceil p+1\rceil - p})$, *we have*

$$\begin{aligned} x_1 &= P_X(x_0 - v_0(g_0/\|g_0\| + r_0)) \\ &= x_0. \end{aligned}$$

Hence, the AQSGM generates a fixed sequence and $\lim_{k\to\infty} f(x_k) = f(x_0) = e^{x_0}$. *However, when* $R = 1$, $\epsilon = 0$ *and* $d = \exp(\frac{\log(\mu\lceil p+1\rceil!)}{\lceil p+1\rceil - p})$, *the total error* $\mu(Rd)^p + \epsilon = \mu d^p = d^{\lceil p+1\rceil}/\lceil p+1\rceil! < e^d = e^{x_0}$, *where the inequality follows from the Taylor expansion. Therefore, Theorem 4.3.2 fails for this example.*

From the above two examples, we observe that the Hölder condition of order p describes some property of the objective function, which is essentially different from the upper semi-continuity, and it can be used to investigate convergence properties of the AQSGM. Hence, using the mild assumptions, we have established convergence properties of the AQSGM from a new perspective, which is different from that in [60].

Finite convergence

The optimal solution set X^* has a nonempty interior in many applications, such as surrogate relaxation of discrete programming problems (see [37]). Here, we demonstrate the finite convergence to the approximate optimal value of problem (4.1.1) under the assumption that the optimal solution set X^* has a nonempty interior.

Theorem 4.3.3 *Let Assumptions 4.3.1-4.3.3 hold,* $\text{int}X^* \neq \emptyset$ *and the diminishing stepsize rule be chosen. Then* $f(x_k) \leq f_* + \mu(Rd)^p + \epsilon$ *for some* k.

Proof. By contradiction, we assume that $f(x_k) > f_* + \mu(Rd)^p + \epsilon, \forall k \in \mathbb{N}$. Since $\text{int}X^* \neq \emptyset$, we set $B(\bar{x}, \bar{\delta}) \subset X^*$ with $\bar{\delta} > 0$. For all $x \in B(\bar{x}, Rd + \frac{2}{3}\bar{\delta})$, we have

$$\begin{aligned} f(x) - f_* &\leq \mu\big(\text{dist}(x, X^*)\big)^p \\ &\leq \mu(Rd - \tfrac{1}{3}\bar{\delta})^p \\ &= \mu(Rd)^p - \delta' \\ &< f(x_k) - f_* - \epsilon - \delta', \end{aligned} \tag{4.3.5}$$

where δ' is a scalar in $[\frac{1}{3}\mu p\bar{\delta}(Rd - \frac{1}{3}\bar{\delta})^{p-1}, \frac{1}{3}\mu p\bar{\delta}(Rd)^{p-1}]$ satisfying the mean value theorem. In addition, since $\overline{\lim_{k\to\infty}} \epsilon_k = \epsilon$, there exists some k_0 such that $\epsilon_k \leq \epsilon + \delta'$ for all $k \geq k_0$. Therefore, (4.3.5) implies $f(x) < f(x_k) - \epsilon_k$ and thus $B(\bar{x}, Rd + \frac{2}{3}\bar{\delta}) \subset S_{f, f(x_k)-\epsilon_k}$ for all $k \geq k_0$. Hence, it follows from Lemma 4.3.3 that

$$\langle g_k/\|g_k\|, x_k - \bar{x}\rangle \geq Rd + \frac{2}{3}\bar{\delta}. \tag{4.3.6}$$

However, summing the basic inequality (4.3.3) with $x = \bar{x}$ over $i = k_0, \cdots, k$, we obtain

$$\begin{aligned} \min_{i=k_0,\ldots,k} \langle g_i/\|g_i\|, x_i - \bar{x}\rangle &\leq \frac{\sum_{i=k_o}^{k} v_i \langle g_i/\|g_i\|, x_i - \bar{x}\rangle}{\sum_{i=k_o}^{k} v_i} \\ &\leq \frac{\|x_{k_0} - \bar{x}\|^2}{2\sum_{i=k_0}^{k} v_i} + Rd + \frac{\sum_{i=k_0}^{k} v_i^2}{2\sum_{i=k_0}^{k} v_i}(1+R)^2. \end{aligned} \tag{4.3.7}$$

By the properties of the diminishing stepsize rule (cf. (1.1.5)), it follows from Lemma 1.3.5 that $\lim_{k\to\infty}(\sum_{i=k_0}^{k} v_i^2 / \sum_{i=k_0}^{k} v_i) = 0$, and hence the right hand side of (4.3.7) tends to Rd as k tends to infinity. Therefore, we arrive at a contradiction with (4.3.6). ■

Under the same assumptions as in Theorem 4.3.3, we now describe a related result for the nonvanishing stepsize rule.

Theorem 4.3.4 *Let Assumptions 4.3.1-4.3.3 hold. If $B(\bar{x}, \bar{\delta}) \subset X^*$ with $\bar{\delta} > 0$ and there exist some $0 < \kappa < 1$ and $k_0 \in \mathbb{N}$ such that $v_k \in [\frac{\kappa^2\bar{\delta}}{(1+R)^2}, \frac{\kappa\bar{\delta}}{(1+R)^2}]$ for all $k \geq k_0$. Then $f(x_k) \leq f_* + \mu(Rd)^p + \epsilon$ for some k.*

Proof. By contradiction, suppose $f(x_k) > f_* + \mu(Rd)^p + \epsilon, \forall k \in \mathbb{N}$. As in the proof of Theorem 4.3.3 and (4.3.7), we have

$$\begin{aligned} Rd + \frac{2}{3}\bar{\delta} &\leq \min_{i=k_0,\ldots,k} \langle g_i/\|g_i\|, x_i - \bar{x}\rangle \\ &\leq \frac{\|x_{k_0} - \bar{x}\|^2}{2\sum_{k_0}^{k} v_i} + Rd + \frac{\sum_{k_0}^{k} v_i^2}{2\sum_{k_0}^{k} v_i}(1+R)^2 \\ &\leq \frac{\|x_{k_0} - \bar{x}\|^2}{2\kappa^2\bar{\delta}(k - k_0 + 1)}(1+R)^2 + Rd + \bar{\delta}/2, \end{aligned}$$

whose last right hand side tends to $Rd + \bar{\delta}/2$ as k tends to infinity. The contradiction happens. ■

Convergence in iterates

We have shown convergence properties in objective values in preceding theorems, and here we consider the convergence property in iterates. In [81], where noise in subgradient methods for convex optimization was considered, Nedić and Bertsekas did not give the convergence property in iterates. In fact, convergence of $\{x_k\}$ is quite difficult to obtain.

Kiwiel [61] has described convergence of $\{x_k\}$ that was generated by ϵ-subgradient method for convex optimization. Although Kiwiel [61] did not consider the effect of noise, his work is really helpful for our research. Following the framework of [61], we will give convergence of $\{x_k\}$ by using the diminishing stepsize rule. Besides the extension to the AQSGM, another improvement of our work is to maintain the convergence property without the lower semi-continuity and coercive condition assumptions used in [61], although we assume the upper semi-continuity instead.

Since we are considering the inexact subgradient method, we can only expect that $\{x_k\}$ converges to some approximate optimal solution set. First, let us show a useful property of a convergent sequence, which also converges in objective values. This result requires the additional upper semi-continuity assumption.

Lemma 4.3.5 *Suppose f is upper semi-continuous on $\mathbb{R}^n$, $\alpha > 0$, and the sequence $\{x_k\}$ converges to $\bar{x}$ with $\lim_{k\to\infty} f(x_k) \leq f_* + \alpha$. Then $\operatorname{dist}(\bar{x}, \bar{S}_{f,f_*+\alpha}) = 0$, where $\bar{S}_{f,f_*+\alpha} = \{x \in \mathbb{R}^n : f(x) \leq f_* + \alpha\}$.*

Proof. For all $\beta > f_* + \alpha$, since $\lim_{k\to\infty} f(x_k) \leq f_* + \alpha < \beta$, we have $\operatorname{dist}(\bar{x}, S_{f,\beta}) = 0$. Observe that $S_{f,f_*+\alpha}$ is a nonempty and open set (since α is positive and f is upper semi-continuous) and $S_{f,f_*+\alpha} \subset \bigcap_{\beta > f_*+\alpha} S_{f,\beta}$. Furthermore, since the sequence of sublevel sets $\{S_{f,\beta}\}$ is decreasing as $\beta \downarrow f_* + \alpha$, by using [98, Exercise 4.3(b)], we have

$$\begin{aligned} \lim_{\beta \downarrow f_*+\alpha} S_{f,\beta} &= \bigcap_{\beta > f_*+\alpha} \operatorname{cl} S_{f,\beta} \\ &= \operatorname{cl} \bigcap_{\beta > f_*+\alpha} S_{f,\beta} \\ &= \operatorname{cl} \bar{S}_{f,f_*+\alpha}, \end{aligned} \tag{4.3.8}$$

where the second equality follows from [97, Theorem 6.5] and the fact that $S_{f,\beta}$ are all convex due to the quasi-convexity. Finally, by using [98, Corollary 4.7] and (4.3.8), we arrive at

$$\begin{aligned} \operatorname{dist}(\bar{x}, \bar{S}_{f,f_*+\alpha}) &= \operatorname{dist}(\bar{x}, \operatorname{cl}(\bar{S}_{f,f_*+\alpha})) \\ &= \lim_{\beta \downarrow f_*+\alpha} \operatorname{dist}(\bar{x}, S_{f,\beta}) \\ &= 0. \end{aligned}$$

■

Another proof of Lemma 4.3.5. By contradiction, we assume $\operatorname{dist}(\bar{x}, \bar{S}_{f,f_*+\alpha}) =$

$\rho > 0$. For all $x \notin \bar{S}_{f,f_*+\alpha}$, we define $X_\alpha(x) := \mathrm{conv}\big(\bar{S}_{f,f_*+\alpha} \cup \{x\}\big) \setminus \bar{S}_{f,f_*+\alpha}$. By the quasi-convexity of f, we have $X_\alpha(x) \subset \bar{S}_f(x)$.

We first claim that $\bigcap_{x \in B(\bar{x},\theta)} X_\alpha(x) \neq \emptyset$ for sufficiently small $\theta > 0$. Indeed, since $\alpha > 0$ and f is quasi-convex and upper semi-continuous, $\bar{S}_{f,f_*+\alpha}$ is a closed and convex set, whose interior is nonempty. Thus, we set $B(z,\delta) \subset \bar{S}_{f,f_*+\alpha}$ with some $\delta > 0$ and thus $\rho_z := \|\bar{x} - z\| > \rho$ due to $\mathrm{dist}(\bar{x}, \bar{S}_{f,f_*+\alpha}) = \rho$. Let $y := \bar{x}(1 - \rho/(2\rho_z)) + z\rho/(2\rho_z)$, which is on the segment between $\bar{x}$ and z but not in $\bar{S}_{f,f_*+\alpha}$, due to $\|y - \bar{x}\| = \rho/2 < \rho$. When $\theta \leq \delta\rho/(2\rho_z - \rho)$, for any point $\bar{x} + \theta w \in B(\bar{x},\theta)$ with $\|w\| \leq 1$, there exists $z - w\theta(2\rho_z - \rho)/\rho \in B(z,\delta)$ such that y is on the segment between $\bar{x} + \theta w$ and $z - w\theta(2\rho_z - \rho)/\rho$. That is $y \in X_\alpha(x)$ for all $x \in B(\bar{x},\theta)$. Therefore, $\bigcap_{x \in B(\bar{x},\theta)} X_\alpha(x) \neq \emptyset$ when $\theta \leq \delta\rho/(2\rho_z - \rho)$.

Since $y \in \bigcap_{x \in B(\bar{x},\theta)} X_\alpha(x)$, the quasi-convexity of f implies $f(y) \leq f(x)$ for all $x \in B(\bar{x},\theta)$. Furthermore, by the assumption that $\{x_k\}$ converges to $\bar{x}$ with $\lim_{k\to\infty} f(x_k) \leq f_* + \alpha$, we obtain $f(y) \leq f_* + \alpha$. However, y does not meet $\bar{S}_{f,f_*+\alpha}$. Hence, we arrive at the contradiction. ■

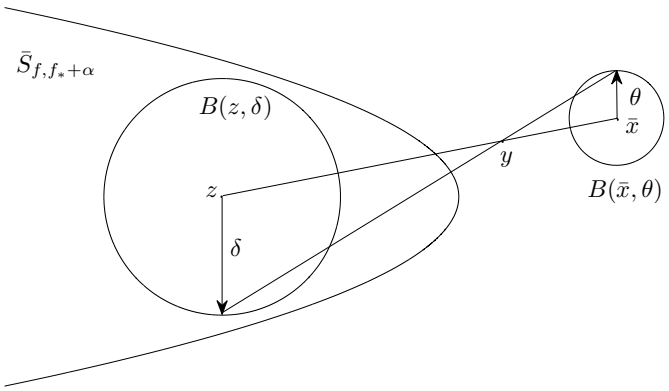


Figure 4.4: Illustration of the second proof of Lemma 4.3.5: $y \in X_\alpha(x)$, $\forall x \in B(\bar{x},\theta)$.

Next, we describe convergence of $\{x_k\}$ to some approximate optimal solution set by

using the diminishing stepsize rule.

Theorem 4.3.5 *Let Assumptions 4.3.1-4.3.3 hold, the total error $c := \mu(Rd)^p + \epsilon > 0$, f be upper semi-continuous on $\mathbb{R}^n$ and the diminishing stepsize rule be chosen. Then the following statements are true:*

(i) $\underline{\lim}_{k\to\infty} \operatorname{dist}(x_k, \bar{S}_{f,f_*+c} \cap X) = 0$,

(ii) $\lim_{k\to\infty} \operatorname{dist}(x_k, X^* + \rho(c)B) = 0$, *where $\rho(c)$ is defined by*

$$\rho(c) := \max\{\operatorname{dist}(x, X^*) : x \in \bar{S}_{f,f_*+c} \cap X\}.$$

Proof. First, observe that $X^* \subset \bar{S}_{f,f_*+c} \cap X \subset X^* + \rho(c)B$. Furthermore, the nonemptiness of X^* and the compactness of X imply that $\bar{S}_{f,f_*+c} \cap X$ is nonempty and bounded.

(i) Theorem 4.3.2 gives that $\underline{\lim}_{k\to\infty} f(x_k) \leq f_* + c$. The compactness of X then implies that there exists some subsequence $\{x_{k_i}\}$ that converges to some $\bar{x} \in X$ with $\lim_{i\to\infty} f(x_{k_i}) \leq f_* + c$. Thus, the conclusion follows from Lemma 4.3.5.

(ii) Given $\sigma > 0$, define

$$V_{2\sigma} := X^* + \rho(c)B + 2\sigma B,$$

and

$$e_\sigma := \inf\{f(x) : x \in X, \operatorname{dist}(x, \bar{S}_{f,f_*+c} \cap X) \geq \sigma\} - (f_* + c). \tag{4.3.9}$$

We first claim that $e_\sigma > 0$. Indeed, if $e_\sigma = 0$, there exists a sequence $\{z_i\}$, in $\{x : x \in X, \operatorname{dist}(x, \bar{S}_{f,f_*+c} \cap X) \geq \sigma\}$, converging to some $\bar{z} \in X$ with $\lim_{i\to\infty} f(z_i) = f_* + c$. It follows from Lemma 4.3.5 that $\operatorname{dist}(\bar{z}, \bar{S}_{f,f_*+c}) = 0$. Moreover, since $\bar{z} \in X$, $\operatorname{dist}(\bar{z}, \bar{S}_{f,f_*+c} \cap X) = 0$, which is impossible as $\sigma > 0$.

For such positive e_σ, there exists some $\delta > 0$ such that

$$\mu(Rd + \delta)^p \leq \mu(Rd)^p + e_\sigma/2. \tag{4.3.10}$$

Since the stepsize v_k diminishes, there exists some $k_\delta \in \mathbb{N}$ such that

$$v_k \leq \delta/(1+R)^2, \forall k \geq k_\delta. \tag{4.3.11}$$

Since $\overline{\lim_{k\to\infty}}\, \epsilon_k = \epsilon$ and $\lim_{k\to\infty} \|x_{k+1} - x_k\| = 0$ (since v_k diminishes), there exists some $k_\sigma \geq k_\delta$ such that

$$\epsilon_k < \epsilon + e_\sigma/2, \tag{4.3.12}$$

and

$$\|x_{k+1} - x_k\| \leq \sigma, \tag{4.3.13}$$

for all $k \geq k_\sigma$. Since $\underline{\lim}_{k\to\infty} \operatorname{dist}(x_k, \bar{S}_{f,f_*+c} \cap X) = 0$ (cf. (i)), there exists some $k'_\sigma \geq k_\sigma \geq k_\delta$ such that

$$\begin{aligned} x_{k'_\sigma} &\in (\bar{S}_{f,f_*+c} \cap X) + \sigma B \\ &\subset X^* + \rho(c)B + \sigma B \\ &\subset V_{2\sigma}, \end{aligned}$$

that is $x_{k'_\sigma} \in V_{2\sigma}$. Next, we claim that $x_k \in V_{2\sigma}$ for all $k \geq k'_\sigma$. Proving by induction, we assume that $x_k \in V_{2\sigma}$ for some $k \geq k'_\sigma$ and consider the following two cases.

Case 1. If $\operatorname{dist}(x_k, \bar{S}_{f,f_*+c} \cap X) \leq \sigma$, from (4.3.13), we have

$$\begin{aligned} x_{k+1} &\in \{x_k\} + \sigma B \\ &\subset (\bar{S}_{f,f_*+c} \cap X + \sigma B) + \sigma B \\ &\subset X^* + \rho(c)B + 2\sigma B \\ &= V_{2\sigma}. \end{aligned}$$

Case 2. Suppose $\operatorname{dist}(x_k, \bar{S}_{f,f_*+c} \cap X) > \sigma$. From (4.3.9), we have

$$\begin{aligned} f(x_k) &\geq e_\sigma + f_* + c \\ &= f_* + (\mu(Rd)^p + e_\sigma/2) + (\epsilon + e_\sigma/2) \\ &> f_* + \mu(Rd + \delta)^p + \epsilon_k, \forall k \geq k'_\sigma, \end{aligned}$$

where the second inequality follows from relations (4.3.10) and (4.3.12). Hence, from Lemmas 4.3.2 and 4.3.4, we have

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 - 2v_k(\delta - \tfrac{v_k}{2}(1+R)^2) \\ &\leq \|x_k - x^*\|^2, \end{aligned}$$

where the second inequality follows from (4.3.11). Thus, $x_k \in V_{2\sigma}$ implies $x_{k+1} \in V_{2\sigma}$.

Therefore, by induction, $x_k \in V_{2\sigma}$ and hence $\operatorname{dist}(x_k, X^* + \rho(c)B) \leq 2\sigma$ for all $k \geq k'_\sigma$. Since $\sigma > 0$ is arbitrary, $\operatorname{dist}(x_k, X^* + \rho(c)B)$ vanishes as k tends to infinity.

■

4.4 Convergence Properties for f with Generalized Weak Sharp Minima

In this section, we consider the other case when X is noncompact. Considering the similar case, Nedić and Bertsekas [81] assumed that the objective function f had a set of weak sharp minima and the ϵ-subgradients were uniformly bounded on X (see [81, Assumptions 3.1-3.2]). The function f is said to have a set of weak sharp minima over X (see [28]) if for some scalar $\eta > 0$ there holds

$$f(x) - f_* \geq \eta \operatorname{dist}(x, X^*), \forall x \in X. \tag{4.4.1}$$

A natural extension to generalize the weak sharp minima is the weak sharp minima of order q (see [19, 107]), that is, there exist some scalars $\eta, q > 0$ such that

$$f(x) - f_* \geq \eta\big(\operatorname{dist}(x, X^*)\big)^q, \forall x \in X. \tag{4.4.2}$$

However, if $p > q$, contradiction between (4.3.1) and (4.4.2) arises as $\operatorname{dist}(x, X^*)$ tends to zero. Also, if $p < q$, contradiction arises again as $\operatorname{dist}(x, X^*)$ tends to infinity. In order to avoid the contradiction, we weaken the assumption (4.4.2) as the generalized weak sharp minima, in which the constant q is replaced by a real-valued function $g(t)$.

Furthermore, in what follows we consider a noise sequence $\{r_k\}$ whose bound R is lower than $(\eta/\mu)^{1/p}$, which we refer to as a low level noise sequence (see [81]). In particular, we introduce the following two assumptions.

Assumption 4.4.1 *The function f satisfies the generalized weak sharp minima condition over X, that is, there exist some scalars $\eta > 0$, $q \geq p$ and a function $g : \mathbb{R}_+ \to \mathbb{R}_+$, satisfying $g(\cdot) \geq p$, $\sup\limits_{t \geq 0} g(t) = q$ and $\lim\limits_{t \to \infty} g(t) = p$, such that*

$$f(x) - f_* \geq \eta\big(\operatorname{dist}(x, X^*)\big)^{g(\operatorname{dist}(x, X^*))}, \forall x \in X, \tag{4.4.3}$$

where p is the order used in Assumption 4.3.2.

Assumption 4.4.2 *$\{r_k\}$ is a low level noise sequence, i.e., $R < (\eta/\mu)^{1/p}$.*

When $g(t) \equiv p$, Assumption 4.4.1 reduces to the weak sharp minima of order p, whose sufficient and necessary conditions have been described by Studniarski and Ward [107] and Bonnans and Ioffe [19] for specified $p = 2$. Furthermore, if $p = 1$, it reduces to the well-known weak sharp minima, which was introduced by Burke and Ferris [28] and widely studied by Burke and Deng [25, 26, 27] and Zheng and Ng [110, 111]. Note that, to arrive at the corresponding convergence results, Assumptions 4.3.2 and 4.4.1 with specified $g(t) \equiv p$ and $p = 1$ are used in [81].

When

$$g(t) := \begin{cases} q, & 0 \le t \le 1, \\ p, & t > 1, \end{cases}$$

where $q > p$, Assumption 4.4.1 reduces to

$$f(x) - f_* \ge \min\{\eta\big(\mathrm{dist}(x, X^*)\big)^{g(0)}, \eta\big(\mathrm{dist}(x, X^*)\big)^p\},$$

which is equivalent to that f has Höldrian level sets over X (see [91]).

Before we go on, for each $v \ge 0$, $\theta \ge 0$ and $x \in X$, we define a new function $H^x_{v,\theta} : \mathbb{R}_+ \to \mathbb{R}$ by

$$H^x_{v,\theta}(z) := \mu\Big(\frac{v}{2}(1+R)^2 + R(\frac{z}{\eta})^{1/g(\mathrm{dist}(x,X^*))}\Big)^p + \epsilon + \theta - z, \tag{4.4.4}$$

where μ and p are scalars given in Assumption 4.3.2, and R and ϵ are scalars given in Assumption 4.3.3. The maximum solution $z^*_{v,\theta}$ of the inequality $H^x_{v,\theta}(z) \ge 0$ over X is defined by

$$z^*_{v,\theta} := \sup\{z : H^x_{v,\theta}(z) \ge 0 \text{ for some } x \in X\}. \tag{4.4.5}$$

Assumption 4.4.1 says that $p \le g(\mathrm{dist}(x, X^*)) \le q$ for all $x \in X$. Hence, from (4.4.4), for given $v \ge 0$ and $\theta \ge 0$, we have

$$H^x_{v,\theta}(z) \le \max\{H^p_{v,\theta}(z), H^q_{v,\theta}(z)\}, \forall z \ge 0, x \in X,$$

where $H^p_{v,\theta}(z) := \mu\Big(\frac{v}{2}(1+R)^2 + R(\frac{z}{\eta})^{1/p}\Big)^p + \epsilon + \theta - z$ and $H^q_{v,\theta}(z) := \mu\Big(\frac{v}{2}(1+R)^2 + R(\frac{z}{\eta})^{1/q}\Big)^p + \epsilon + \theta - z$. Thus, applying (4.4.5) and Assumption 4.4.1, $z^*_{v,\theta}$ can be rewritten as

$$z^*_{v,\theta} = \max\{\sup\{z : H^p_{v,\theta}(z) \ge 0\}, \sup\{z : H^q_{v,\theta}(z) \ge 0\}\}.$$

For the sake of simplicity, denote

$$z^p_{v,\theta} := \sup\{z : H^p_{v,\theta}(z) \geq 0\} \quad \text{and} \quad z^q_{v,\theta} := \sup\{z : H^q_{v,\theta}(z) \geq 0\}, \tag{4.4.6}$$

and hence

$$z^*_{v,\theta} = \max\{z^p_{v,\theta}, z^q_{v,\theta}\}. \tag{4.4.7}$$

Since $H^x_{v,\theta}(0) > 0$ and $H^x_{v,\theta}(z)$ is continuous on variable z for all $x \in X$, $z^*_{v,\theta}$ is positive. However, it might be $+\infty$. The following lemma shows that $z^*_{v,\theta}$ is finite and continuous on parameters v and θ under Assumptions 4.4.1-4.4.2.

Lemma 4.4.1 *If Assumptions 4.4.1-4.4.2 hold, then the following statements hold:*

(i) $z^*_{v,\theta}$ *is finite for all* $v \geq 0$ *and* $\theta \geq 0$,

(ii) $\lim\limits_{\theta \to 0_+} z^*_{v,\theta} = z^*_{v,0}$ *for all* $v \geq 0$,

(iii) $\lim\limits_{v \to 0_+} z^*_{v,\theta} = z^*_{0,\theta}$ *for all* $\theta \geq 0$.

Proof.

(i) By the assumption, since $R < (\eta/\mu)^{1/p}$ and $q \geq p$, we have

$$\lim_{z\to\infty} \mu\Big(\frac{R}{\eta^{1/q}} z^{1/q-1/p}\Big)^p < 1,$$

which is equivalent to

$$\lim_{z\to\infty} \Big[\frac{\mu}{z}\Big(\frac{v}{2}(1+R)^2 + R(\frac{z}{\eta})^{1/q}\Big)^p + \frac{\epsilon+\theta}{z}\Big] < 1, \forall v \geq 0,\ \theta \geq 0.$$

This implies $\lim\limits_{z\to\infty} H^q_{v,\theta}(z) < 0$. Hence, $z^q_{v,\theta} < +\infty$ for all $v \geq 0$ and $\theta \geq 0$ since $H^q_{v,\theta}(\cdot)$ is continuous. Similarly, we can prove that $z^p_{v,\theta} < +\infty$ for all $v \geq 0$ and $\theta \geq 0$. Thus, by using (4.4.7), we arrive at that $z^*_{v,\theta}$ is finite for all $v \geq 0$ and $\theta \geq 0$.

(ii) Since $H^q_{v_1,\theta_1}(\cdot) \leq H^q_{v_2,\theta_2}(\cdot)$ for all $v_1 \leq v_2$ and $\theta_1 \leq \theta_2$, then $z^q_{v_1,\theta_1} \leq z^q_{v_2,\theta_2}$. This monotonicity immediately implies $\lim\limits_{\theta\to 0} z^q_{v,\theta} \geq z^q_{v,0}$.

Next, we prove the reverse inequality. By the definition of $z^q_{v,\theta}$, for given $v \geq 0$ and each positive integer n, there exists some z_n satisfying $z_n > z^q_{v,1/n} - 1/n$ and

$H^q_{v,1/n}(z_n) \geq 0$. Together with the monotonicity of $z^q_{v,\theta}$, we have $-1 < z_n \leq z^q_{v,1/n} \leq z^q_{v,1}$, where the last term is finite by (i). So the sequence $\{z_n\}$ is bounded and has cluster points. Thus, for each of its cluster points $\bar{z}$, taking a subsequence of $\{z_n\}$ if necessary, we have

$$\begin{aligned}\lim_{n\to\infty} H^q_{v,1/n}(z_n) &= \lim_{n\to\infty} \mu\Big(\tfrac{v}{2}(1+R)^2 + R(\tfrac{z_n}{\eta})^{1/q}\Big)^p + \epsilon + \tfrac{1}{n} - z_n \\ &= \mu\Big(\tfrac{v}{2}(1+R)^2 + R(\tfrac{\bar{z}}{\eta})^{1/q}\Big)^p + \epsilon - \bar{z} \\ &= H^q_{v,0}(\bar{z}),\end{aligned}$$

which is nonnegative, since $\{H^q_{v,1/n}(z_n)\}$ are all nonnegative. Then, by the definition of $z^q_{v,\theta}$, we have $z^q_{v,0} \geq \bar{z} \geq \lim\limits_{\theta\to 0_+} z^q_{v,\theta}$, where the second inequality holds due to $z_n > z^q_{v,1/n} - 1/n$. Therefore, we arrive at $\lim\limits_{\theta\to 0_+} z^q_{v,\theta} = z^q_{v,0}$.

Similarly, we can prove that $\lim\limits_{\theta\to 0_+} z^p_{v,\theta} = z^p_{v,0}$. Thus, from (4.4.7), we arrive at $\lim\limits_{\theta\to 0_+} z^*_{v,\theta} = z^*_{v,0}$ for all $v \geq 0$.

(iii) The proof is similar to that of (ii).

■

These properties of $z^*_{v,\theta}$ will be used in the study of convergence in both objective values and iterates when X is noncompact in what follows.

Convergence in objective values

Similar to Section 4.3, we start with the following basic inequality.

Lemma 4.4.2 *Let Assumption 4.3.3 hold and $\{x_k\}$ be the sequence generated by the AQSGM. Then for all $x \in X$, we have*

$$\|x_{k+1} - x\|^2 \leq \|x_k - x\|^2 - 2v_k\Big(\langle g_k/\|g_k\|, x_k - x\rangle - R\|x_k - x\| - \frac{1}{2}v_k(1+R)^2\Big), \forall k.$$

Proof. By relations (4.2.5)-(4.2.6) and the nonexpansive property of the projection operator, for all $x \in X$, we have the following basic inequality

$$\begin{aligned}\|x_{k+1} - x\|^2 &= \|P_X(x_k - v_k\tilde{g}_k) - x\|^2 \\ &\leq \|x_k - v_k\tilde{g}_k - x\|^2 \\ &= \|x_k - x\|^2 - 2v_k\langle g_k/\|g_k\| + r_k, x_k - x\rangle + v_k^2\|g_k/\|g_k\| + r_k\|^2 \\ &\leq \|x_k - x\|^2 - 2v_k\Big(\langle g_k/\|g_k\|, x_k - x\rangle - R\|x_k - x\| - \tfrac{1}{2}v_k(1+R)^2\Big),\end{aligned}$$

where the last inequality follows from the boundedness of noise. ■

Before we discuss convergence in objective values which is the main result in this subsection, we consider the following two lemmas which show the boundedness of the sequence $\{x_k\}$, generated by the AQSGM using both the constant and diminishing stepsize rules.

Lemma 4.4.3 *Suppose Assumptions 4.3.2-4.4.2 hold and $\{x_k\}$ is generated by the AQSGM with the constant stepsize rule. Then, $\{x_k\}$ is bounded.*

Proof. Since $\overline{\lim_{k\to\infty}}\, \epsilon_k = \epsilon$, for any $\theta > 0$, there exists some positive integer k_0 such that

$$\epsilon_k < \epsilon + \theta, \forall k \geq k_0. \tag{4.4.8}$$

Define the maximum solution of $t^{g(t)} \leq z^*_{v,\theta}/\eta$ by

$$T := \sup\{t \in \mathbb{R}_+ : t^{g(t)} \leq z^*_{v,\theta}/\eta\}, \tag{4.4.9}$$

which is finite, since $z^*_{v,\theta}$ is finite (cf. Lemma 4.4.1(i)) and $\lim_{t\to\infty} t^{g(t)} = +\infty$ (cf. Assumption 4.4.1). Next, we claim that the following inequality holds for all $i \geq k_0$:

$$\operatorname{dist}(x_i, X^*) \leq \max\{\operatorname{dist}(x_{k_0}, X^*), T + v(1+R)\}. \tag{4.4.10}$$

It is obvious that the relation (4.4.10) holds if $i = k_0$. Proving by induction, we assume the relation (4.4.10) holds for some $i = k$ ($\geq k_0$) and consider the following two cases.

Case 1. If $f(x_k) \leq f_* + \mu\Big(\frac{v}{2}(1+R)^2 + R(\frac{f(x_k)-f_*}{\eta})^{1/g(\operatorname{dist}(x_k,X^*))}\Big)^p + \epsilon_k$, together with (4.4.8), we have

$$\mu\Big(\frac{v}{2}(1+R)^2 + R(\frac{f(x_k) - f_*}{\eta})^{1/g(\operatorname{dist}(x_k,X^*))}\Big)^p + \epsilon + \theta - (f(x_k) - f_*) \geq 0,$$

that is, $H^{x_k}_{v,\theta}(f(x_k)-f_*) \geq 0$. Hence, from (4.4.5), we obtain $f(x_k)-f_* \leq z^*_{v,\theta}$, and thus

$$\operatorname{dist}(x_k, X^*)^{g(\operatorname{dist}(x_k,X^*))} \leq z^*_{v,\theta}/\eta,$$

which follows from (4.4.3). Thus, from (4.4.9), we arrive at $\operatorname{dist}(x_k, X^*) \leq T$, and thus relations (4.2.5)-(4.2.6) imply

$$\begin{aligned} \operatorname{dist}(x_{k+1}, X^*) &\leq \operatorname{dist}(x_k, X^*) + v_k\|g_k/\|g_k\| + r_k\| \\ &\leq T + v(1+R), \end{aligned}$$

that is, the relation (4.4.10) holds for $i = k+1$.

Case 2. Suppose $f(x_k) > f_* + \mu\Big(\frac{v}{2}(1+R)^2 + R(\frac{f(x_k)-f_*}{\eta})^{1/g(\operatorname{dist}(x_k,X^*))}\Big)^p + \epsilon_k$. Then, it follows from Lemma 4.3.4 that

$$\begin{aligned} \langle g_k/\|g_k\|, x_k - x^*\rangle &\geq \tfrac{v}{2}(1+R)^2 + R(\tfrac{f(x_k)-f_*}{\eta})^{1/g(\operatorname{dist}(x_k,X^*))} \\ &\geq \tfrac{v}{2}(1+R)^2 + R\operatorname{dist}(x_k, X^*), \end{aligned}$$

where the second inequality follows from (4.4.3). Hence, applying Lemma 4.4.2 with $v_k = v$ and $x^* = P_{X^*}(x_k)$, we obtain

$$\begin{aligned} &\big(\operatorname{dist}(x_{k+1}, X^*)\big)^2 \\ &\leq \|x_{k+1} - x^*\|^2 \\ &\leq \|x_k - x^*\|^2 - 2v\Big(\tfrac{v}{2}(1+R)^2 + R\operatorname{dist}(x_k, X^*) - R\|x^k - x^*\| - \tfrac{v}{2}(1+R)^2\Big) \\ &= \big(\operatorname{dist}(x_k, X^*)\big)^2. \end{aligned}$$

Hence, the relation (4.4.10) holds for $i = k+1$.

Therefore, by induction, the relation (4.4.10) holds for all $i \geq k_0$. Since the right hand side of (4.4.10) is finite and X^* is compact, $\{x_k\}$ is bounded. ■

Lemma 4.4.4 *If Assumptions 4.3.2-4.4.2 hold and $\{x_k\}$ is generated by the AQSGM with the diminishing stepsize rule. Then, $\{x_k\}$ is bounded.*

Proof. The proof uses the properties of the diminishing stepsize rule and a line of analysis similar to that of Theorem 4.4.1. The details are omitted. ■

From Lemmas 4.4.3-4.4.4, $\{x_k\}$ is bounded and hence $\{f(x_k)\}$ is bounded from above due to the Hölder condition (cf. (4.3.1)), using both types of stepsize rules. We denote the upper bound on $\{f(x_k)\}$ by M in what follows.

Theorem 4.4.1 *Let Assumptions 4.3.2-4.4.2 hold. Then, for a sequence $\{x_k\}$ be generated by the AQSGM with the constant stepsize rule, we have*

$$\underline{\lim}_{k\to\infty} f(x_k) \leq f_* + z^*_{v,0},$$

where $z^*_{v,0}$ is finite.

Proof. The finiteness of $z^*_{v,0}$ has been proved in Lemma 4.4.1(i). To prove the convergence property, we first show that

$$\underline{\lim}_{k\to\infty} f(x_k) < f_* + z^*_{v,\theta}$$

for all $\theta > 0$ by contradiction, that is, assume that the following inequality holds for some $\theta > 0$,

$$\underline{\lim}_{k\to\infty} f(x_k) \geq f_* + z^*_{v,\theta}.$$

Thus, there exist some $\delta \in (0, \min\{\theta/2, z^*_{v,\theta}\})$ and positive integer k_0 such that

$$f(x_k) > f_* + z^*_{v,\theta} - \delta, \tag{4.4.11}$$

and

$$\epsilon_k < \epsilon + \theta/2, \tag{4.4.12}$$

for all $k \geq k_0$, where (4.4.12) holds due to $\overline{\lim}_{k\to\infty} \epsilon_k = \epsilon$.

From (4.4.5) and (4.4.11), we obtain $f(x_k) - f_* + \delta > z^*_{v,\theta}$ and thus $H^{x_k}_{v,\theta}(f(x_k) - f_* + \delta) < 0$, that is,

$$\begin{aligned} f(x_k) &> f_* + \mu\left(\tfrac{v}{2}(1+R)^2 + R(\tfrac{f(x_k)-f_*+\delta}{\eta})^{1/g(\mathrm{dist}(x_k,X^*))}\right)^p + \epsilon + \theta - \delta \\ &> f_* + \mu\left(\tfrac{v}{2}(1+R)^2 + R(\tfrac{f(x_k)-f_*+\delta}{\eta})^{1/g(\mathrm{dist}(x_k,X^*))}\right)^p + \epsilon_k \\ &\geq f_* + \mu\left(\tfrac{v}{2}(1+R)^2 + R(\tfrac{f(x_k)-f_*}{\eta})^{1/g(\mathrm{dist}(x_k,X^*))} + \delta'\right)^p + \epsilon_k, \forall k \geq k_0, \end{aligned}$$

where the second inequality follows from (4.4.12) and $0 < \delta < \theta/2$, and the third inequality follows from the Taylor expansion with $\delta' = \min\{\frac{\delta}{\eta q}(\frac{z^*_{v,\theta}}{\eta})^{1/q-1}, \frac{\delta}{\eta p}(\frac{M-f_*}{\eta})^{1/p-1}\} > 0$ (recall that M is an upper bound on $\{f(x_k)\}$). Therefore, by using Lemmas 4.3.4 and 4.4.2, we obtain

$$\begin{aligned} \langle g_k/\|g_k\|, x_k - x^*\rangle &\geq \tfrac{v}{2}(1+R)^2 + R(\tfrac{f(x_k)-f_*}{\eta})^{1/g(\mathrm{dist}(x_k,X^*))} + \delta' \\ &\geq \tfrac{v}{2}(1+R)^2 + R\,\mathrm{dist}(x_k, X^*) + \delta', \forall k \geq k_0, \end{aligned}$$

and hence

$$\begin{aligned}\big(\mathrm{dist}(x_{k+1},X^*)\big)^2 &\le \big(\mathrm{dist}(x_k,X^*)\big)^2 - 2v\delta'\\ &\le \cdots \le \big(\mathrm{dist}(x_{k_0},X^*)\big)^2 - 2(k-k_0+1)v\delta',\end{aligned}$$

which yields a contradiction for sufficiently large k. Therefore, we have

$$\underline{\lim}_{k\to\infty} f(x_k) \le f_* + z^*_{v,\theta}, \forall\theta > 0.$$

Taking the limit as $\theta \to 0$, by using Lemma 4.4.1(ii), we arrive at the conclusion. ■

We now give explicit expressions for the tolerance in approaching f_* in Theorem 4.4.1 in some specific cases of p and $g(t)$. By solving relations (4.4.6)-(4.4.7), we have the following corollaries where the total errors are given in explicit expressions.

Corollary 4.4.1 *Let Assumptions 4.3.2-4.4.2 hold with $g(t) \equiv p$ and $p = 1$. Then, for a sequence $\{x_k\}$ generated by the AQSGM with the constant stepsize rule, we have*

$$\underline{\lim}_{k\to\infty} f(x_k) \le f_* + \Big(\frac{1}{2}\mu v(1+R)^2 + \epsilon\Big)\frac{\eta}{\eta - R\mu}.$$

Proof. By the assumptions, since $g(t) \equiv p$ and $p = q = 1$, we have

$$H^p_{v,0}(z) = H^q_{v,0}(z) = \mu\Big(\frac{v}{2}(1+R)^2 + R\frac{z}{\eta}\Big) + \epsilon - z \ \text{ and } \ z^p_{v,0} = z^q_{v,0}.$$

It is clear that $H^p_{v,0}(z)$ is linear and decreasing due to $R < \eta/\mu$. Thus, by (4.4.6), $z^p_{v,0}$ is just the solution of $H^p_{v,0}(z) = 0$. Then, from (4.4.7), we have $z^*_{v,0} = z^p_{v,0} = \Big(\frac{1}{2}\mu v(1+R)^2 + \epsilon\Big)\frac{\eta}{\eta - R\mu}$. Hence, by Theorem 4.4.1, we arrive at the conclusion. ■

Corollary 4.4.2 *Let Assumptions 4.3.2-4.4.2 hold with $g(t) \equiv p$ and $p = 2$. Then, for a sequence $\{x_k\}$ generated by the AQSGM with the constant stepsize rule, we have*

$$\underline{\lim}_{k\to\infty} f(x_k) \le f_* + \eta\Big(\frac{\mu v R(1+R)^2 + \sqrt{\eta\mu v^2(1+R)^4 + 4\epsilon(\eta - \mu R^2)}}{2(\eta - \mu R^2)}\Big)^2.$$

Proof. By the assumptions, $g(t) \equiv p$ and $p = q = 2$, we have

$$H^p_{v,0}(z) = H^q_{v,0}(z) = \mu\Big(\frac{v}{2}(1+R)^2 + R(\frac{z}{\eta})^{1/2}\Big)^2 + \epsilon - z \ \text{ and } \ z^p_{v,0} = z^q_{v,0}.$$

Taking $y = (\frac{z}{\eta})^{1/2}$, we define $H^p_{v,0}(z(y)) := \mu\Big(\frac{v}{2}(1+R)^2 + Ry\Big)^2 + \epsilon - \eta y^2$. By calculation, the solution of $H^p_{v,0}(z(y)) = 0$ is $\bar{y} = \frac{\mu v R(1+R)^2 + \sqrt{\eta\mu v^2(1+R)^4 + 4\epsilon(\eta - \mu R^2)}}{2(\eta - \mu R^2)}$ (the negative solution should be ignored since $y \geq 0$). Moreover, the derivative $\big(H^p_{v,0}(z(y))\big)' = \mu v R(1+R)^2 + 2y(\mu R^2 - \eta)$, which is negative if $y > \frac{\mu v R(1+R)^2}{2(\eta - \mu R^2)}$. Hence, $H^p_{v,0}(z(y))$ is decreasing and thus negative if $y > \bar{y}$. Then, from (4.4.7), we have $z^*_{v,0} = z^p_{v,0} = z(\bar{y}) = \eta\bar{y}^2$. Thus, by Theorem 4.4.1, we arrive at the conclusion. ■

Using the diminishing stepsize rule, the total error tends to $z^*_{0,0}$ as v_k diminishes and the following theorem is obtained.

Theorem 4.4.2 *Let Assumptions 4.3.2-4.4.2 hold and the sequence $\{x_k\}$ be generated by the AQSGM with the diminishing stepsize rule. Then, $z^*_{0,0}$ is finite and*

$$\varliminf_{k\to\infty} f(x_k) \leq f_* + z^*_{0,0}.$$

Proof. The proof uses the properties of the diminishing stepsize rule and a line of analysis similar to that of Theorem 4.4.1. The details are omitted. ■

Corollary 4.4.3 *Let Assumptions 4.3.2-4.4.2 hold with $g(t) \equiv p$ and $p = 1$. Then, for a sequence $\{x_k\}$ generated by the AQSGM with the diminishing stepsize rule, we have*

$$\varliminf_{k\to\infty} f(x_k) \leq f_* + \frac{\eta\epsilon}{\eta - R\mu}.$$

Proof. Solving relations (4.4.6)-(4.4.7) with $v = \theta = 0$, $g(t) \equiv p$ and $p = 1$, by Theorem 4.4.2, we arrive at this corollary. ■

So far, we have established convergence properties in objective values of the AQSGM and extended the corresponding results [81] in the presence of the generalized weak sharp minima condition (see Theorems 4.4.1-4.4.2). Specified $g(t) \equiv p$ and $p = 1$, the generalized weak sharp minima reduces to the weak sharp minima as in [81] and the total error in Corollaries 4.4.1 and 4.4.3 have similar formulae to that in [81, Propositions 3.1-3.2].

The exact quasi-subgradient method with the noise and the ϵ-quasi-subgradient method can be considered as two special cases of the AQSGM with $\epsilon_k \equiv 0$ and $R =$

0 respectively. We now demonstrate how the noise r_k and the errors ϵ_k affect the established results under the assumption of the weak sharp minima (see (4.4.1)).

When exact subgradients are used ($\epsilon_k \equiv 0$) and the low level noise is persistent ($R > 0$), the total error $e := \eta\epsilon/(\eta - R\mu)$ in Corollary 4.4.3 vanishes and the objective value converges to f_*.

Corollary 4.4.4 *Let Assumptions 4.3.2-4.4.2 hold with $g(t) \equiv p$ and $p = 1$. Then, for a sequence $\{x_k\}$ generated by the AQSGM with the diminishing stepsize rule and $\epsilon_k \equiv 0$, we have*

$$\varliminf_{k\to\infty} f(x_k) = f_*.$$

By contrast, exact convergence cannot be guaranteed in the corresponding result of Corollary 4.3.1. When ϵ_k-subgradients are used ($\epsilon > 0$) and noise vanishes ($R = 0$), the total error in the estimate of Corollary 4.4.3 does not vanish. In particular, the total error $e = \eta\epsilon/(\eta - R\mu)$ is proportional to the error level ϵ. This demonstrates the different effect of the noise r_k and the errors ϵ_k on the AQSGM under the assumption of the weak sharp minima.

Finite convergence

In this subsection, we describe the finite convergence to the approximate optimal value of problem (4.1.1) under the assumption that the optimal solution set X^* has a nonempty interior.

Theorem 4.4.3 *Let Assumptions 4.3.2-4.4.2 hold,* $\mathrm{int}X^* \neq \emptyset$ *and the diminishing stepsize rule be chosen. Then, $f(x_k) \leq f_* + z^*_{0,0}$ for some k.*

Proof. By contradiction, we assume that $f(x_k) > f_* + z^*_{0,0}$ for all $k \in \mathbb{N}$. Then, from (4.4.5), we have $H^{x_k}_{0,0}(f(x_k) - f_*) < 0$, that is,

$$f(x_k) > f_* + \mu\Big(R(\frac{f(x_k) - f_*}{\eta})^{1/g(\mathrm{dist}(x_k, X^*))}\Big)^p + \epsilon. \tag{4.4.13}$$

Due to Lemma 4.4.4, we have $z^*_{0,0} + f_* < f(x_k) \leq M$. Since $\text{int}X^* \neq \emptyset$, we set $B(\bar{x}, \bar{\delta}) \subset X^*$ with $\bar{\delta} > 0$. Hence, for all $x \in B(\bar{x}, R(\frac{f(x_k)-f_*}{\eta})^{1/g(\text{dist}(x_k,X^*))} + \frac{2}{3}\bar{\delta})$ there holds

$$\begin{aligned} f(x) - f_* &\leq \mu\big(\text{dist}(x, X^*)\big)^p \\ &\leq \mu\Big(R(\tfrac{f(x_k)-f_*}{\eta})^{1/g(\text{dist}(x_k,X^*))} - \tfrac{1}{3}\bar{\delta}\Big)^p \\ &\leq \mu\Big(R(\tfrac{f(x_k)-f_*}{\eta})^{1/g(\text{dist}(x_k,X^*))}\Big)^p - \delta' \\ &< f(x_k) - f_* - \delta' - \epsilon, \end{aligned} \tag{4.4.14}$$

where $\delta' = \min\{\frac{\delta}{3\eta q}(\frac{z^*_{0,0}}{\eta})^{1/q-1}, \frac{\delta}{3\eta p}(\frac{M-f_*}{\eta})^{1/p-1}\} > 0$, the third inequality follows from the Taylor expansion, and the fourth inequality follows from (4.4.13). In addition, since $\overline{\lim_{k\to\infty}} \epsilon_k = \epsilon$, there exists some k_0 such that $\epsilon_k \leq \epsilon + \delta'$ for all $k \geq k_0$. Therefore, (4.4.14) implies $f(x) < f(x_k) - \epsilon_k$ and thus $B(\bar{x}, R(\frac{f(x_k)-f_*}{\eta})^{1/g(\text{dist}(x_k,X^*))} + \frac{2}{3}\bar{\delta}) \subset S_{f,f(x_k)-\epsilon_k}$ for all $k \geq k_0$. Thus, it follows from Lemma 4.3.3 and (4.4.3) that

$$\begin{aligned} \langle g_k/\|g_k\|, x_k - \bar{x}\rangle &\geq R(\tfrac{f(x_k)-f_*}{\eta})^{1/g(\text{dist}(x_k,X^*))} + \tfrac{2}{3}\bar{\delta} \\ &\geq R\,\text{dist}(x_k, X^*) + \tfrac{2}{3}\bar{\delta}. \end{aligned}$$

Hence, by using Lemma 4.4.2, we obtain

$$\big(\text{dist}(x_{k+1}, X^*)\big)^2 \leq \big(\text{dist}(x_k, X^*)\big)^2 - \frac{4}{3}v_k\bar{\delta} + v_k^2(1+R)^2.$$

It gives the upper bound on $\bar{\delta}$ as follows

$$\bar{\delta} \leq \frac{3\big(\text{dist}(x_{k_0}, X^*)\big)^2}{4\sum_{i=k_0}^k v_i} + \frac{3\sum_{i=k_0}^k v_i^2}{4\sum_{i=k_0}^k v_i}(1+R)^2, \tag{4.4.15}$$

whose right hand side tends to zero as k tends to infinity. Hence, we arrive at a contradiction with $\bar{\delta} > 0$. ■

Under the same assumptions, we now describe a related result for nonvanishing stepsize rule.

Theorem 4.4.4 *Let Assumptions 4.3.2-4.4.2 hold. If $B(\bar{x}, \bar{\delta}) \subset X^*$ with $\bar{\delta} > 0$ and there exist some $0 < \kappa < 1$ and $k_0 \in \mathbb{N}$ such that $v_k \in [\frac{\kappa^2\bar{\delta}}{(1+R)^2}, \frac{\kappa\bar{\delta}}{(1+R)^2}]$ for all $k \geq k_0$. Then, $f(x_k) \leq f_* + z^*_{0,0}$ for some k.*

Proof. Suppose $f(x_k) > f_* + z^*_{0,0}$ for all $k \in \mathbb{N}$. As in the proof of Theorem 4.4.3 and (4.4.15), we have

$$\begin{aligned} \bar{\delta} &\leq \frac{3\big(\mathrm{dist}(x_{k_0}, X^*)\big)^2}{4\sum_{i=k_0}^{k} v_i} + \frac{3\sum_{i=k_0}^{k} v_i^2}{4\sum_{i=k_0}^{k} v_i}(1+R)^2 \\ &\leq \frac{3\big(\mathrm{dist}(x_{k_0}, X^*)\big)^2}{4\kappa^2\bar{\delta}(k-k_0+1)}(1+R)^2 + \frac{3}{4}\bar{\delta}, \end{aligned}$$

whose last right hand side tends to $\frac{3}{4}\bar{\delta}$ as k tends to infinity. The contradiction happens. ■

Convergence in iterates

Similar to Section 4.3, we prove the following theorem which describes convergence of $\{x_k\}$ to some approximate optimal solution set in the presence of a generalized weak sharp minima condition (see (4.4.3)).

Theorem 4.4.5 *Let Assumptions 4.3.2-4.4.2 hold with $z^*_{0,0} > 0$ (see (4.4.5)), f be upper semi-continuous on $\mathbb{R}^n$ and the diminishing stepsize rule be chosen. Then the following statements are true:*

(i) $\underline{\lim}_{k\to\infty} \mathrm{dist}(x_k, \bar{S}_{f,f_*+z^*_{0,0}} \cap X) = 0.$

(ii) $\lim_{k\to\infty} \mathrm{dist}(x_k, X^* + \rho(z^*_{0,0})B) = 0$, *where $\rho(z^*_{0,0})$ is defined by*

$$\rho(z^*_{0,0}) := \max\{\mathrm{dist}(x, X^*) : x \in \bar{S}_{f,f_*+z^*_{0,0}} \cap X\}.$$

Proof. The proof uses a line of analysis similar to that of Theorem 4.3.5.

(i) The boundedness of $\{x_k\}$ is given by Lemma 4.4.4, and it is given by Theorem 4.4.2 that $\underline{\lim}_{k\to\infty} f(x_k) \leq f_* + z^*_{0,0}$. Hence, there exists some subsequence $\{x_{k_i}\}$ that converges to some $\bar{x} \in X$ with $\lim_{i\to\infty} f(x_{k_i}) \leq f_* + z^*_{0,0}$. Thus, the conclusion follows from Lemma 4.3.5.

(ii) Given $\sigma > 0$, define

$$V_{2\sigma} := X^* + \rho(z^*_{0,0})B + 2\sigma B,$$

and

$$e_\sigma := \inf\{f(x) : x \in X, \operatorname{dist}(x, \bar{S}_{f,f_*+z^*_{0,0}} \cap X) \geq \sigma\} - (f_* + z^*_{0,0}). \quad (4.4.16)$$

Since e_σ is positive (by the proof of Theorem 4.3.5) and $z^*_{v,\theta}$ is continuous on parameters v and θ (cf. Lemma 4.4.1), there exist some positive v and θ such that

$$z^*_{v,\theta} < z^*_{0,0} + e_\sigma. \quad (4.4.17)$$

Since the stepsize v_k diminishes, there exists $k_v \in \mathbb{N}$ such that

$$v_k \leq v, \forall k \geq k_v. \quad (4.4.18)$$

Since $\overline{\lim_{k\to\infty}} \epsilon_k = \epsilon$ and $\lim_{k\to\infty} \|x_{k+1} - x_k\| = 0$ (since v_k diminishes), there exists some $k_\sigma \geq k_v$ such that

$$\epsilon_k < \epsilon + \theta, \quad (4.4.19)$$

and

$$\|x_{k+1} - x_k\| \leq \sigma, \quad (4.4.20)$$

for all $k \geq k_\sigma$. Since $\underline{\lim_{k\to\infty}} \operatorname{dist}(x_k, \bar{S}_{f,f_*+z^*_{0,0}} \cap X) = 0$ (cf. (i)), there exists some $k'_\sigma \geq k_\sigma \geq k_v$ such that

$$\begin{aligned} x_{k'_\sigma} &\in (\bar{S}_{f,f_*+z^*_{0,0}} \cap X) + \sigma B \\ &\subset X^* + \rho(z^*_{0,0})B + \sigma B \\ &\subset V_{2\sigma}, \end{aligned}$$

that is $x_{k'_\sigma} \in V_{2\sigma}$.

Next, we claim that $x_k \in V_{2\sigma}$ for all $k \geq k'_\sigma$. Proving by induction, we assume that $x_k \in V_{2\sigma}$ for some $k \geq k'_\sigma$ and consider the following two cases.

Case 1. If $\operatorname{dist}(x_k, \bar{S}_{f,f_*+z^*_{0,0}} \cap X) \leq \sigma$, from (4.4.20), we have

$$\begin{aligned} x_{k+1} &\in \{x_k\} + \sigma B \\ &\subset (\bar{S}_{f,f_*+z^*_{0,0}} \cap X + \sigma B) + \sigma B \\ &\subset X^* + \rho(z^*_{0,0})B + 2\sigma B \\ &= V_{2\sigma}. \end{aligned}$$

Case 2. Suppose $\operatorname{dist}(x_k, \bar{S}_{f,f_*+z^*_{0,0}} \cap X) > \sigma$. From (4.4.16), we have

$$\begin{aligned} f(x_k) &\geq e_\sigma + f_* + z^*_{0,0} \\ &> f_* + z^*_{v,\theta}, \end{aligned}$$

where the second inequality follows from (4.4.17). Thus, from (4.4.5), we obtain $H^{x_k}_{v,\theta}(f(x_k) - f_*) < 0$, that is,

$$\begin{aligned} f(x_k) &> f_* + \mu\Big(\tfrac{v}{2}(1+R)^2 + R(\tfrac{f(x_k)-f_*}{\eta})^{1/g(\mathrm{dist}(x_k,X^*))}\Big)^p + \epsilon + \theta \\ &> f_* + \mu\Big(\tfrac{v}{2}(1+R)^2 + R(\tfrac{f(x_k)-f_*}{\eta})^{1/g(\mathrm{dist}(x_k,X^*))}\Big)^p + \epsilon_k, \forall k \geq k_\sigma, \end{aligned}$$

where the second inequality follows from (4.4.19). Hence, from Lemmas 4.3.4 and 4.4.2, we have

$$\begin{aligned} \big(\mathrm{dist}(x_{k+1}, X^*)\big)^2 &\leq \big(\mathrm{dist}(x_k, X^*)\big)^2 - 2v_k\big(\tfrac{v}{2}(1+R)^2 - \tfrac{v_k}{2}(1+R)^2\big) \\ &\leq \big(\mathrm{dist}(x_k, X^*)\big)^2, \end{aligned}$$

where the second inequality follows from (4.4.18). Thus, $x_k \in V_{2\sigma}$ implies $x_{k+1} \in V_{2\sigma}$.

Therefore, by induction, $x_k \in V_{2\sigma}$ and hence $\mathrm{dist}(x_k, X^* + \rho(z^*_{0,0})B) \leq 2\sigma$ for all $k \geq k'_\sigma$. Since $\sigma > 0$ is arbitrary, $\mathrm{dist}(x_k, X^* + \rho(z^*_{0,0})B)$ vanishes as k tends to infinity.

■

4.5 Efficiency

In this section, under the bounded assumption (i.e., Assumptions 4.3.1 and 4.3.3), we discuss efficiency estimates of the AQSGM. In order to quantify the efficiency of the AQSGM, we introduce some concepts as in [60].

The inradius of a set Z is the radius of the largest ball contained in Z, denoted by

$$\dot{r}(Z) := \sup\{r > 0 : B(x, r) \subset Z \text{ for some } x \in Z\}. \tag{4.5.1}$$

For any $\gamma \in (0, 1)$, the γ-solution set of problem (4.1.1) is defined by

$$X^*_\gamma := \{x \in X : \dot{r}(S_f(x)) < \gamma\dot{r}(X)\}. \tag{4.5.2}$$

It follows from (4.5.2) that x is an γ-solution of problem (4.1.1) if $x \in X$ and $S_f(x)$ does not contain a ball with radius $\gamma\dot{r}(X)$. Thus, the significance of the inradius is to estimate the efficiency of algorithms, inasmuch as x is an γ-solution if $\dot{r}(S_f(x)) < \gamma\dot{r}(X)$.

The criterion is that the quality of iterates improves if the inradius of its strict sublevel set decreases.

At iteration $k \geq 1$, the record value $f_{\epsilon,k}^{rec}$ denotes the best approximate value found so far, and is defined by

$$f_{\epsilon,k}^{rec} := \min_{j=1,\ldots,k} \{f(x_j) - \epsilon_j\}. \tag{4.5.3}$$

Let $\dot{r}_k$ denote the inradius of the record strict sublevel set, defined by

$$\dot{r}_k := \dot{r}(S_{f,f_{\epsilon,k}^{rec}}),$$

which is nonincreasing in k.

In view of application considerations, we would like our algorithm to reach the γ-solution set as fast as possible. Since the quality of the record value/point improves if the inradius $\dot{r}_k$ decreases (cf. [60, Lemma 13]), we would like $\dot{r}_k$ to decrease as fast as possible. For this purpose, we now give an upper bound on $\dot{r}_k$ that depends on the stepsize rule.

Lemma 4.5.1 *Let Assumptions 4.3.1 and 4.3.3 hold. For a sequence $\{x_k\}$ generated by the AQSGM, we have*

$$\dot{r}_k \leq Rd + \frac{d^2 + (1+R)^2 \sum_{j=i}^k v_j^2}{2\sum_{j=i}^k v_j}, \ for\ i = 1, \cdots, k. \tag{4.5.4}$$

Proof. Suppose $\dot{r}_k > 0$. For any $\delta < \dot{r}_k$, it follows from (4.5.1) that there exists some $\bar{x}$ such that $B(\bar{x}, \delta) \subset S_{f,f_{\epsilon,k}^{rec}}$. Then for each $j = 1, \cdots, k$, from (4.5.3), we have $B(\bar{x}, \delta) \subset S_{f,f(x_j)-\epsilon_j}$. Hence, it follows from Lemma 4.3.3 that

$$\langle g_j/\|g_j\|, x_j - \bar{x}\rangle \geq \delta, \ for\ j = 1, \cdots, k.$$

Therefore, from Lemma 4.3.2, we have

$$\|x_{j+1} - \bar{x}\|^2 \leq \|x_j - \bar{x}\|^2 - 2v_j\delta + 2v_jRd + v_j^2(1+R)^2.$$

Summing these inequalities over $j = i, \cdots, k$, we arrive at

$$\delta \leq Rd + \frac{d^2 + (1+R)^2 \sum_{j=i}^k v_j^2}{2\sum_{j=i}^k v_j}, \ for\ i = 1, \cdots, k.$$

Since $\delta < \dot{r}_k$ is arbitrary, we arrive at the conclusion. ■

In the sense of guaranteeing that the record values/points become γ-solutions as fast as possible, the best stepsize may be found by minimizing the upper bound of $\dot{r}_k$ in (4.5.4). In the following, we offer the best choice on the constant stepsize rule and estimate the rate of efficiency by using the diminishing stepsize rule.

Theorem 4.5.1 *Let Assumptions 4.3.1 and 4.3.3 hold. For a sequence $\{x_k\}$ generated by the AQSGM, the following statements hold:*

(i) *if a constant stepsize v is chosen, then $\dot{r}_k \leq \frac{d^2}{2kv} + Rd + \frac{v}{2}(1+R)^2$,*

(ii) *the best constant stepsize is $v = \frac{d}{(1+R)\sqrt{k}}$ and then $\dot{r}_k \leq \frac{d(1+R)}{\sqrt{k}} + Rd$,*

(iii) *if the diminishing stepsize is chosen as $v_i = a/\sqrt{i}$, then*

$$\dot{r}_k \leq Rd + ck^{-1/2} \quad with \quad c = \frac{d^2 + a^2(1+\log 2)(1+R)^2}{a(4-2\sqrt{2})}.$$

More general, if v_k is chosen as the diminishing stepsize rule, then $\lim_{k\to\infty} \dot{r}_k \leq Rd$.

Proof.

(i) It is (4.5.4) specifying $i = 1$ and $v_i \equiv v$.

(ii) Minimizing the upper bound on $\dot{r}_k$ in (i) with respect to v, we obtain the best constant stepsize $v = \frac{d}{(1+R)\sqrt{k}}$ and the corresponding upper bound on the inradius.

(iii) It follows from [86, p.157] that

$$\sum_{j=i}^{k} j^{-1} \leq 1 + \log 2 \quad \text{and} \quad \sum_{j=i}^{k} j^{-1/2} \geq (2-\sqrt{2})k^{1/2}, \ for\ i = \lceil \frac{k}{2} \rceil.$$

Using the relation (4.5.4), we obtain

$$\dot{r}_k \leq Rd + \frac{d^2 + a^2(1+\log 2)(1+R)^2}{a(4-2\sqrt{2})k^{1/2}} = Rd + ck^{-1/2}.$$

Furthermore, from the properties of the diminishing stepsize rule (see (1.1.5)), we have $\lim_{k\to\infty} (\sum_{j=i}^{k} v_j^2 / \sum_{j=i}^{k} v_j) = 0$ (cf. [61, Lemma 2.1]), the relation (4.5.4) implies $\lim_{k\to\infty} \dot{r}_k \leq Rd$.

■

4.6 Numerical Experiments

In this section, we show two numerical experiments to illustrate the performance of the AQSGM on the fractional programming, which is widely applied in applications and has been extensively studied by many researchers (see e.g. [9, 34, 39, 44, 50]). In the first experiment, we compare the exact quasi-subgradient method (in short, QSGM) with the projected level function method (in short, PLFM) proposed by Xu [109]. The computation result shows that the exact quasi-subgradient method arrives at a better solution in fewer number of iterations on this small-scale problem. In the second experiment, we show the performance of the AQSGM and illustrate the sensitivity of inexact terms on the large-scale fractional programming. The computation result coincides with the obtained convergence results and shows that the AQSGM is suitable for large-scale problems.

Fractional programming

Consider the following fractional programming problems (see [109])

$$\begin{aligned} \min \quad & p(x)/q(x) \\ \text{s.t.} \quad & x \in C, \end{aligned} \tag{4.6.1}$$

where

$$p(x) = \max\{x_1^2 + x_2^4; (2-x_1)^2 + (2-x_1)^2(2-x_2)^2; 2e^{x_2-x_1}\},$$

$$q(x) = c_1x_1 + c_2x_2 + 1,$$

with c_1 and c_2 specified below, and

$$C = \{x \in \mathbb{R}^2 : x \geq 0; x_1 + x_2 \leq 3\}.$$

Since $p(x)$ is convex and $q(x)$ is affine and positive when (c_1, c_2) is restricted to be nonnegative, it can be shown that the function $p(x)/q(x)$ is nonsmooth and quasi-convex on C for specified c in Table 4.1 (cf. criterion (K) in [9, page 209]). We compare the exact quasi-subgradient method (QSGM) with projected level function method (PLFM) in [109] for this example. We illustrate the numerical results in Table 4.1. In this table, $c = (c_1, c_2)^T$ is a vector representing the parameters in the function $q(x)$. At the column of parameters, it denotes stepsize and parameter λ respectively in QSGM

and PLFM. NIT denotes the number of iterations needed to reach the approximate optimal value, f_{opt} denotes the objective function value as the algorithms terminate and the last column presents the approximate solution as the algorithms terminate. It is illustrated in Table 4.1 that the QSGM arrives at a better solution in fewer number of iterations on this small-scale problem.

Table 4.1: Computation results for fractional programming

Algorithm	c	Parameters	NIT	f_{opt}	Solution
PLFM	(0,0)	1	36	1.9552	(1.1669,0.8770)
QSGM	(0,0)	$0.1/(1+0.1k)$	14	1.9530	(1.1473,0.8879)
PLFM	(2,1)	1	33	0.4615	(1.0001,1.0005)
QSGM	(2,1)	$0.1/(1+0.1k)$	23	0.4614	(1.3839,0.8402)
PLFM	(20,10)	1	31	0.0583	(1.2609,0.7967)
QSGM	(20,10)	$0.1/(1+0.1k)$	29	0.0583	(1.2635,0.8129)

Minimax linear fractional programming

Consider the following minimax linear fractional programming

$$\begin{aligned} \min \quad & \max_{1\le k\le p} \frac{c_k^T x+\alpha_k}{d_k^T x+\beta_k} \\ \text{s.t.} \quad & Ax \le b, \\ & x \ge 0, \end{aligned}$$

which is a nonsmooth and quasi-convex maximization problem when the denominators are positive (cf. criterion (K) in [9, page 209]). All elements of the matrix A, vectors b, c_k, d_k and scalars α_k, β_k are randomly generated from the uniform distribution on certain intervals: $A \in [0,1]$, $b \in [n,2n]$, $c_k \in [0,50]$, $d_k \in [0,5]$, $\alpha_k \in [-50,50]$, $\beta_k \in [0,5]$.

The minimax linear fractional programming has been well studied in many articles (see [34, 39, 44] and references therein), so the quasi-convex programming formulation does not lead to improve solutions for this problem, but it provides an illuminating example of how to find such a formulation more generally, and we use the large-scale minimax linear fractional programming example to illustrate the AQSGM.

We use the AQSGM to solve this fractional programming in medium-scale and large-scale. In this experiment, we add a deterministic noise into each quasi-subgradient. In Table 4.2, we illustrate the numerical results of the AQSGM when the noise level $R = 0.01$. The stopping criterion is that the error in approaching the optimal value is less than 0.05, i.e., $\|f_k - f_*\| < 0.05$. NIT and time denote the corresponding number and time of iterations needed to reach the specified level respectively. f_{opt} denotes the objective function value as the AQSGM terminates. It is illustrated that NIT lies on a stable level as the dimension increases. Hence, the AQSGM is suitable for large-scale problems.

Table 4.2: Computation results of the AQSGM

Dimension	Number	Stepsize	NIT	Time	f_{opt}
10	100	$1/(1+0.1k)$	93	1.3sec	6.6344
10	1000	$1/(1+0.1k)$	57	6.5sec	4.8343
50	100	$1/(1+0.1k)$	113	8.0sec	9.0081
50	1000	$1/(1+0.1k)$	68	47sec	7.8315
100	100	$3/(1+0.1k)$	106	16sec	9.7454
100	1000	$3/(1+0.1k)$	70	1.8min	8.4501
200	100	$3/(1+0.1k)$	109	35sec	9.9213
200	2000	$3/(1+0.1k)$	76	6min	8.9462

We also do some tests on the effect of noise on the AQSGM. In Figure 4.5, we show the convergence behavior of the AQSGM and the sensitivity of inexact terms as $R = 0.1$, 0.2 and 1. The gaps between curves describe the total errors in approaching the optimal value (cf. Theorem 4.3.2). Recall that the approximate quasi-subgradient has the form (see (4.2.6))

$$\tilde{g}_k := g_k/\|g_k\| + r_k,$$

where the first term $g_k/\|g_k\|$ has a norm of 1. It is worth mentioning in Figure 4.5 that the sequence does not converge when $R \geq 1$. This is because R is too big that the quasi-subgradient is not the primary direction in the iteration (cf. (4.2.6)).

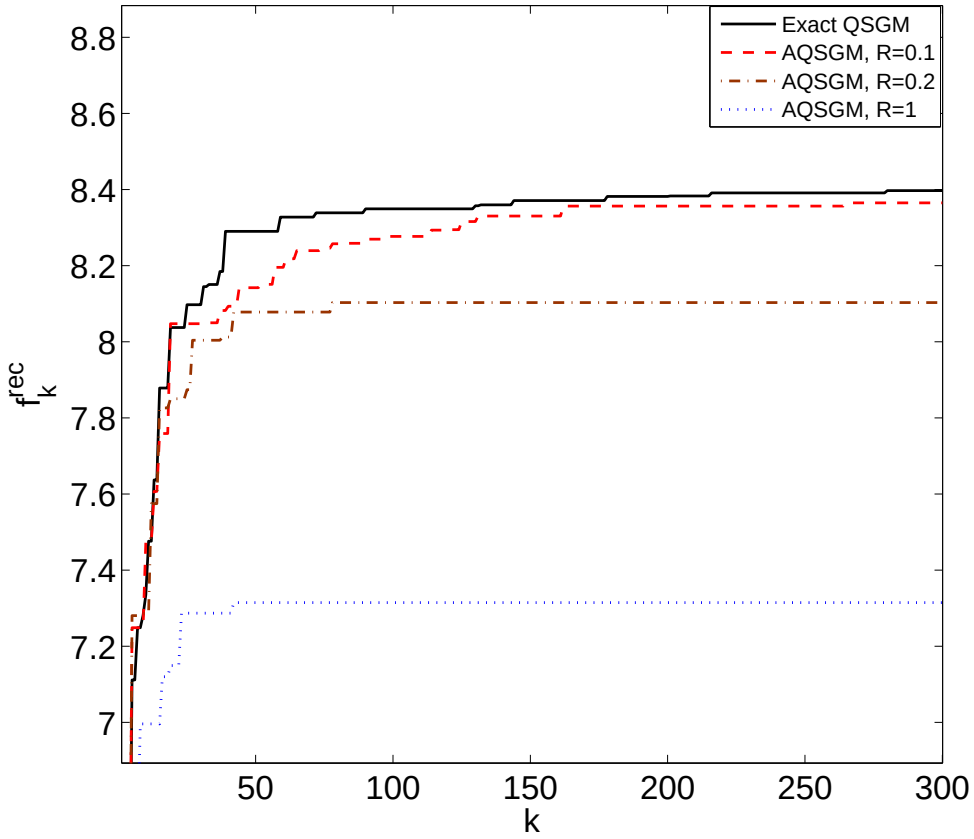


Figure 4.5: The convergence behavior of the AQSGM.

Chapter 5

Quantized Approximate Quasi-Subgradient Method

5.1 Introduction

Polyak [93, 94] and Nedić and Bertsekas [81] studied the effect of noise on subgradient methods for convex constrained optimization problems. In Chapter 4, to meet much broader class of problems, we have proposed an approximate quasi-subgradient method (AQSGM) and investigated the effect of inexact terms on the AQSGM for quasi-convex constrained optimization problems.

However, as the motivating example given in Section 4.2, the distributed optimization problem in networks usually requires the data at each node and transmitted data to reach a quantization level (see e.g. [10, 54, 95]). As in Chapter 4, only considering noise on the transmitted data is not enough. Therefore, in this chapter, we investigate the influence of inexact items and convergence behavior on the quantized approximate quasi-subgradient method (in short, QAQSGM), which applies a quantization operator after the subgradient iteration along the approximate quasi-subgradient, for quantized quasi-convex constrained optimization problems.

In this chapter, we consider the following nondifferentiable quantized quasi-convex

constrained optimization problem:

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in X \\ & x \in \Lambda, \end{array} \tag{5.1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a quasi-convex function, X is a closed and convex set, and Λ is a quantization lattice defined by

$$\Lambda = \{(\lambda_1\Delta, \lambda_2\Delta, \cdots, \lambda_n\Delta) : \lambda_i \in \mathbb{Z}\}, \tag{5.1.2}$$

where $\Delta > 0$ is the given quantization scalar. We denote the optimal solution set and the optimal value of problem (5.1.1) respectively by X^* and f_*, and we assume that X^* is nonempty and compact.

Inspired by the idea in [81] and references therein, we propose a quantized approximate quasi-subgradient method (QAQSGM), and investigate the effect of inexact terms and convergence behavior on the QAQSGM. Considering the generic inexact subgradient algorithm for problem (5.1.1) and assuming the inexact terms are deterministic and bounded, we establish convergence results in two cases: (i) X is compact and (ii) X is noncompact. Throughout this chapter, we only consider the constant stepsize rule and obtain the best constant stepsize by minimizing the tolerance in approaching the optimal value.

This chapter is organized as follows. In Section 5.2, we present the QAQSGM algorithm. In Section 5.3, we demonstrate convergence properties of the QAQSGM when the constraint set X is compact or when f satisfies a generalized weak sharp minima condition over a noncompact set X.

5.2 QAQSGM Algorithm

For a given scalar $\Delta > 0$, the quantization lattice Λ in (5.1.2) consists of points regularly spaced by Δ along each coordinate axis. We define the quantization operator $Q : \mathbb{R}^n \to X \cap \Lambda$ by

$$Q(\cdot) = P_{X\cap\Lambda}(P_X(\cdot)), \tag{5.2.1}$$

which projects its argument first onto the constraint set X and then onto the nearest lattice point in $X \cap \Lambda$. Note that applying the quantization operator $Q(\cdot)$ to $x \in \mathbb{R}^n$ is not equivalent to directly projecting x to the nearest point in $X \cap \Lambda$. In particular, when $x \notin X$, the nearest point to x in $X \cap \Lambda$ can be different from $Q(x)$, and may result in a large error (see [95]).

In this section, we propose a quantized approximate quasi-subgradient method (QAQSGM) to solve problem (5.1.1) as follows.

Quantized approximate quasi-subgradient method (QAQSGM)
Select the stepsize v, an error sequence $\{\epsilon_k\}$ and a noise sequence $\{r_k\}$, start with an initial point $x_0 \in X$, and generate a sequence $\{x_k\} \in X$ via the iteration

$$x_{k+1} = Q(x_k - v\tilde{g}_k), \tag{5.2.2}$$

where the direction $\tilde{g}_k$ is an approximate quasi-subgradient of the following form

$$\tilde{g}_k := g_k/\|g_k\| + r_k, \tag{5.2.3}$$

where r_k is a noise vector and $g_k \in \bar{\partial}^*_{\epsilon_k} f(x_k)$ is an arbitrary nonzero ϵ_k-quasi-subgradient of f at x_k (cf. (4.2.3)).

If the diminishing stepsize rule is utilized, the QAQSGM may terminate at some point that is far away from the optimal value/solution. An example is given as follows.

Example 5.2.1 *Consider the following quantized quasi-convex constrained optimization problem*

$$\begin{array}{ll} min & \|x\| \\ s.t. & x \in \mathbb{R}_+ \\ & x \in \mathbb{N}. \end{array}$$

Obviously, its optimal value is $f_* = 0$ and its optimal solution set is $X^* = \{0\}$. In particular, we choose $\epsilon_k \equiv 0$ and $r_k \equiv 0$, and the QAQSGM reduces to the exact quantized quasi-subgradient method. It is easy to verify that $\bar{\partial}^* f(x) = \mathbb{R}_+$ for all $x > 0$, and the algorithm generates the iterates via

$$x_{k+1} \ = Q_X(x_k - v_k).$$

Suppose we choose the diminishing stepsize rule, for instance, $v_k = v/k$ ($v \gg 1$). Since the quantization scalar $\Delta = 1$, the algorithm is sure to terminate after $k = 2v$ iterations. Starting from the initial point $x_0 = 4v^2 + 10v$, we have

$$\begin{aligned} x_{2v} &\geq x_{2v-1} - \frac{v}{2v} - 1 \\ &\geq \cdots \\ &\geq x_0 - \sum_{i=1}^{2v} \frac{v}{i} - 2v \\ &\geq x_0 - v - \sum_{i=2}^{2^2} \frac{v}{i} - \cdots - \sum_{i=2^{\lceil \log 2v \rceil}}^{2^{(\lceil \log 2v \rceil + 1)}} \frac{v}{i} - 2v \\ &\geq x_0 - v - v(1 + \log 2)\lceil \log 2v \rceil - 2v \\ &> x_0 - v - 4v^2 - 2v \\ &= 7v, \end{aligned}$$

where the first inequality holds due to the quantization lattice, and the fifth inequality follows from the fact that $\sum_{j=i}^{k} j^{-1} \leq 1 + \log 2$ for all $i = k/2$ and $k \in \mathbb{N}$ (see [86, p.157]). Thus, the algorithm terminates at a point that is far away from ($> 7v$) the optimal solution.

Hence, in this chapter, assuming the noise and errors are deterministic and bounded, we investigate convergence properties of the QAQSGM only using the constant stepsize rule.

5.3 Convergence Analysis

The convergence analysis is divided into two cases: (i) X is compact and (ii) X is noncompact while the objective function satisfies the generalized weak sharp minima condition.

5.3.1 Convergence for a Compact X

In this subsection, we investigate the convergence property of the QAQSGM when the constraint set X is compact. Same as Section 4.3, throughout this subsection, the

following three assumptions are made.

Assumption 5.3.1 *The constraint set X is compact.*

Assumption 5.3.2 *f satisfies the Hölder condition of order $p > 0$ with modulus $\mu > 0$ on $\mathbb{R}^n$, that is,*

$$f(x) - f_* \leq \mu\big(\mathrm{dist}(x, X^*)\big)^p, \forall x \in \mathbb{R}^n. \tag{5.3.1}$$

Assumption 5.3.3 *The noise and errors are bounded, i.e., there exist some scalars R and $\epsilon \geq 0$ such that*

$$\|r_k\| \leq R, \forall k \geq 0 \quad \text{and} \quad \overline{\lim_{k\to\infty}}\, \epsilon_k = \epsilon.$$

Since the constraint set X is compact, all iterates are bounded. Therefore, there exists some $d > 0$ (such as the diameter of X) such that $\|x_k - x\| \leq d$ for all $x \in X$ and $k \geq 0$. Moreover, under the bounded noise assumption, it follows from (5.2.3) that approximate quasi-subgradients are uniformly bounded, i.e., $\|\tilde{g}_k\| \leq 1 + R$ for all $k \geq 0$.

We start with the following lemmas that describe a very important property of the quantization operator and show the basic inequality of the subgradient iteration respectively.

Lemma 5.3.1 *For all $x \in R^n$ and $y \in X$, we have*

$$\|Q(x) - y\| \leq \|x - y\| + \sqrt{n}\Delta. \tag{5.3.2}$$

Proof. Due to the structure of quantization lattice Λ (cf. (5.1.2)), for all $x \in R^n$ and $y \in X$, we obtain

$$\begin{aligned} \|Q(x) - y\| &= \|Q(x) - P_X(x) + P_X(x) - y\| \\ &\leq \|Q(x) - P_X(x)\| + \|P_X(x) - y\| \\ &\leq \sqrt{n}\Delta + \|P_X(x) - y\| \\ &\leq \sqrt{n}\Delta + \|x - y\|, \end{aligned}$$

where the third inequality follows from the nonexpansive property of the projection operator. ■

Lemma 5.3.2 *Suppose Assumptions 5.3.1 and 5.3.3 hold, and the sequence $\{x_k\}$ is generated by the QAQSGM. Then for all $x \in X$ and $k \in \mathbb{N}$, we have*

$$\|x_{k+1}-x\|^2 \le \|x_k-x\|^2 - 2v\Big(\langle \frac{g_k}{\|g_k\|}, x_k - x\rangle - d(\frac{\sqrt{n}\Delta}{v}+R) - \frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2\Big).$$

Proof. By relations (5.2.2)-(5.2.3) and Lemma 5.3.1, for all $x \in X$, we have the following basic inequality

$$\begin{aligned}
\|x_{k+1}-x\|^2 &= \|Q(x_k - v\tilde{g}_k) - x\|^2 \\
&\le (\|x_k - v\tilde{g}_k - x\| + \sqrt{n}\Delta)^2 \\
&= \|x_k - v\tilde{g}_k - x\|^2 + 2\sqrt{n}\Delta\|x_k - v\tilde{g}_k - x\| + n\Delta^2 \\
&= \|x_k - x\|^2 - 2v\langle g_k/\|g_k\| + r_k, x_k - x\rangle + v^2\|g_k/\|g_k\| + r_k\|^2 \\
&\quad +2\sqrt{n}\Delta\|x_k - v\tilde{g}_k - x\| + n\Delta^2 \\
&\le \|x_k - x\|^2 - 2v\Big(\langle \frac{g_k}{\|g_k\|}, x_k - x\rangle - Rd - \frac{1}{2}v(1+R)^2 \\
&\quad -\frac{\sqrt{n}\Delta(d+v(1+R))}{v} - \frac{n\Delta^2}{2v}\Big) \\
&\le \|x_k - x\|^2 - 2v\Big(\langle \frac{g_k}{\|g_k\|}, x_k - x\rangle - d(R + \frac{\sqrt{n}\Delta}{v}) - \frac{\big(v(1+R)+\sqrt{n}\Delta\big)^2}{2v}\Big),
\end{aligned}$$

where the second inequality follows from the compactness of X and boundedness of noise. ■

The convergence result of the QAQSGM is demonstrated as follows.

Theorem 5.3.1 *Let Assumptions 5.3.1-5.3.3 hold. Then, for a sequence $\{x_k\}$ generated by the QAQSGM, we have*

$$\underline{\lim}_{k\to\infty} f(x_k) \le f_* + \mu\Big(d(\frac{\sqrt{n}\Delta}{v}+R) + \frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2\Big)^p + \epsilon. \tag{5.3.3}$$

Proof. By contradiction, we assume that

$$\underline{\lim}_{k\to\infty} f(x_k) > f_* + \mu\Big(d(\frac{\sqrt{n}\Delta}{v}+R) + \frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2\Big)^p + \epsilon,$$

that is, there exist some $\delta > 0$ and positive integer k_0 such that

$$f(x_k) > f_* + \mu\Big(d(\frac{\sqrt{n}\Delta}{v}+R) + \frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + \delta\Big)^p + \epsilon_k, \forall k \ge k_0.$$

Thus, it follows from Lemma 4.3.4 (only require Assumption 5.3.2) that for all $x^* \in X^*$ and $k \geq k_0$ there holds

$$\langle g_k/\|g_k\|, x_k - x^* \rangle \geq d(\frac{\sqrt{n}\Delta}{v} + R) + \frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + \delta.$$

Therefore, from Lemma 5.3.2 with $x^* \in X^*$, we obtain

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 - 2v\Big(d(\frac{\sqrt{n}\Delta}{v} + R) + \frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + \delta \\ &\quad -d(\frac{\sqrt{n}\Delta}{v} + R) - \frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2\Big) \\ &= \|x_k - x^*\|^2 - 2v\delta \\ &< \cdots < \|x_{k_0} - x^*\|^2 - 2(k - k_0 + 1)v\delta, \end{aligned}$$

which yields a contradiction for sufficiently large k. ■

Thus, for a given quantization scalar Δ, the best constant stepsize can be obtained by minimizing the tolerance estimated in (5.3.3), i.e.,

$$\min_{v \geq 0} \Big\{ d(\frac{\sqrt{n}\Delta}{v} + R) + \frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 \Big\}.$$

It is trivial to verify that the best constant stepsize, which is the optimal solution of the preceding optimization problem, is given by

$$v^* = \frac{\sqrt{n\Delta^2 + 2d\sqrt{n}\Delta}}{1+R}.$$

It is also observed that the tolerance, given in (5.3.3), has the same expression as that of the AQSGM (cf. Theorem 4.3.1) if the quantization operator is infinitely precise (i.e., Δ is sufficiently small).

5.3.2 Convergence for f with Generalized Weak Sharp Minima

In this subsection, we consider the case when X is noncompact. In this case, we assume that the objective function f satisfies the generalized weak sharp minima condition over X, as in Chapter 4. In particular, we introduce the following two assumptions.

Assumption 5.3.4 *The function f satisfies the generalized weak sharp minima condition over X, that is, there exist some scalars $\eta > 0$, $q \geq p$ and a function $g : \mathbb{R}_+ \to \mathbb{R}_+$,*

satisfying $g(\cdot) \geq p$, $\sup_{t\geq 0} g(t) = q$ *and* $\lim_{t\to\infty} g(t) = p$, *such that*

$$f(x) - f_* \geq \eta\big(\mathrm{dist}(x, X^*)\big)^{g(\mathrm{dist}(x,X^*))}, \forall x \in X, \tag{5.3.4}$$

where p *is the order used in Assumption 5.3.2.*

Assumption 5.3.5 $\{r_k\}$ *is a low level noise sequence, i.e.,* $R + \sqrt{n}\Delta/v < (\eta/\mu)^{1/p}$.

Before we go on, for each $\theta \geq 0$ and $x \in X$, we define a new function $K_\theta^x : \mathbb{R}_+ \to \mathbb{R}$ by

$$K_\theta^x(y) := \mu\Big(\frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\frac{\sqrt{n}\Delta}{v} + R)(\frac{y}{\eta})^{1/g(\mathrm{dist}(x,X^*))}\Big)^p + \epsilon + \theta - y, \tag{5.3.5}$$

where μ and p are scalars given in Assumption 4.3.2 and R and ϵ are scalars given in Assumption 4.3.3. The maximum solution y_θ^* of the inequality $K_\theta^x(y) \geq 0$ over X is defined by

$$y_\theta^* := \sup\{y : K_\theta^x(y) \geq 0 \text{ for some } x \in X\}. \tag{5.3.6}$$

Assumption 5.3.4 says that $p \leq g(\mathrm{dist}(x, X^*)) \leq q$ for all $x \in X$. Hence, from (5.3.5), for given $\theta \geq 0$, we have

$$K_\theta^x(y) \leq \max\{K_\theta^p(y), K_\theta^q(y)\}, \forall y \geq 0, x \in X,$$

where $K_\theta^p(y) := \mu\Big(\frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\frac{\sqrt{n}\Delta}{v} + R)(\frac{y}{\eta})^{1/p}\Big)^p + \epsilon + \theta - y$ and $K_\theta^q(y) := \mu\Big(\frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\frac{\sqrt{n}\Delta}{v} + R)(\frac{y}{\eta})^{1/q}\Big)^p + \epsilon + \theta - y$. Thus, applying (5.3.6) and Assumption 5.3.4, y_θ^* can be rewritten as

$$y_\theta^* = \max\{\sup\{y : K_\theta^p(y) \geq 0\}, \sup\{y : K_\theta^q(y) \geq 0\}\}.$$

For the sake of simplicity, denote

$$y_\theta^p := \sup\{y : K_\theta^p(y) \geq 0\} \quad \text{and} \quad y_\theta^q := \sup\{y : K_\theta^q(y) \geq 0\}, \tag{5.3.7}$$

and hence

$$y_\theta^* = \max\{y_\theta^p, y_\theta^q\}. \tag{5.3.8}$$

Since $K_\theta^x(0) > 0$ and $K_\theta^x(y)$ is continuous on variable y for all $x \in X$, then y_θ^* is positive. However, it might be $+\infty$. The following lemma shows that y_θ^* is finite and continuous on parameter θ under Assumptions 5.3.4-5.3.5.

Lemma 5.3.3 *Let Assumptions 5.3.4-5.3.5 hold. Then the following statements hold:*

(i) y_θ^* *is finite for all* $\theta \geq 0$,

(ii) $\lim_{\theta \to 0_+} y_\theta^* = y_0^*$.

Proof.

(i) By the assumptions, since $R + \sqrt{n}\Delta/v < (\eta/\mu)^{1/p}$ and $q \geq p$, we have

$$\lim_{y \to \infty} \mu\Big(\frac{R + \sqrt{n}\Delta/v}{\eta^{1/q}} y^{1/q - 1/p}\Big)^p < 1,$$

which is equivalent to

$$\lim_{y \to \infty} \Big[\frac{\mu}{y}\Big(\frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\frac{\sqrt{n}\Delta}{v} + R)(\frac{y}{\eta})^{1/q}\Big)^p + \frac{\epsilon + \theta}{y}\Big] < 1, \forall \theta \geq 0.$$

This implies $\lim_{y \to \infty} K_\theta^q(y) < 0$. Hence, $y_\theta^q < +\infty$ for all $\theta \geq 0$ since $K_\theta^q(\cdot)$ is continuous. Similarly, we can prove that $y_\theta^p < +\infty$ for all $\theta \geq 0$. Thus, by using (5.3.8), we arrive at that y_θ^* is finite for all $\theta \geq 0$.

(ii) Since $K_{\theta_1}^q(\cdot) \leq K_{\theta_2}^q(\cdot)$ for all $\theta_1 \leq \theta_2$, then $y_{\theta_1}^q \leq y_{\theta_2}^q$. This monotonicity immediately implies $\lim_{\theta \to 0} y_\theta^q \geq y_0^q$.

Next, we prove the reverse inequality. By definition of y_θ^q, for each positive integer i, there exists some y_i satisfying $y_i > y_{1/i}^q - 1/i$ and $K_{1/i}^q(y_i) \geq 0$. Together with the monotonicity of y_θ^q, we have $-1 < y_i \leq y_{1/i}^q \leq y_1^q$, where the last term is finite by (i). So the sequence $\{y_i\}$ is bounded and has cluster points. Thus, for each of its cluster points $\bar{y}$, taking a subsequence of $\{y_i\}$ if necessary, we have

$$\begin{aligned}
\lim_{i \to \infty} K_{1/i}^q(y_i) &= \lim_{i \to \infty} \mu\Big(\tfrac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\tfrac{\sqrt{n}\Delta}{v} + R)(\tfrac{y_i}{\eta})^{1/q}\Big)^p + \epsilon + \tfrac{1}{i} - y_i \\
&= \mu\Big(\tfrac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\tfrac{\sqrt{n}\Delta}{v} + R)(\tfrac{\bar{y}}{\eta})^{1/q}\Big)^p + \epsilon - \bar{y} \\
&= K_0^q(\bar{y}),
\end{aligned}$$

which is nonnegative, since $\{K_{1/i}^q(y_i)\}$ are all nonnegative. Then, by the definition of y_θ^q, we have $y_0^q \geq \bar{y} \geq \lim_{\theta \to 0_+} y_\theta^q$, where the second inequality follows from $y_i > y_{1/i}^q - 1/i$. Therefore, we arrive at $\lim_{\theta \to 0_+} y_\theta^q = y_0^q$.

Similarly, we can prove that $\lim_{\theta \to 0_+} y_\theta^p = y_0^p$. Thus, from (5.3.8), we arrive at $\lim_{\theta \to 0_+} y_\theta^* = y_0^*$.

■

These properties of y_θ^* will be used in the study of the convergence property of the QAQSGM when X is noncompact in what follows. Also, we start with the following basic inequality.

Lemma 5.3.4 *Let Assumption 5.3.3 hold and $\{x_k\}$ be the sequence generated by the QAQSGM. Then, for all $x \in X$ and $k \in \mathbb{N}$, we have*

$$\|x_{k+1}-x\|^2 \leq \|x_k-x\|^2-2v\Big(\langle\frac{g_k}{\|g_k\|}, x_k-x\rangle-\|x_k-x\|(\frac{\sqrt{n}\Delta}{v}+R)-\frac{1}{2v}\big(\sqrt{n}\Delta+v(1+R)\big)^2\Big).$$

Proof. By (5.2.2)-(5.2.3) and Lemma 5.3.1, for all $x \in X$, we have the following basic inequality

$$\begin{aligned}
\|x_{k+1}-x\|^2 &= \|Q(x_k-v\tilde{g}_k)-x\|^2 \\
&\leq (\|x_k-v\tilde{g}_k-x\|+\sqrt{n}\Delta)^2 \\
&= \|x_k-v\tilde{g}_k-x\|^2+2\sqrt{n}\Delta\|x_k-v\tilde{g}_k-x\|+n\Delta^2 \\
&= \|x_k-x\|^2-2v\langle g_k/\|g_k\|+r_k, x_k-x\rangle+v^2\|g_k/\|g_k\|+r_k\|^2 \\
&\quad +2\sqrt{n}\Delta\|x_k-v\tilde{g}_k-x\|+n\Delta^2 \\
&\leq \|x_k-x\|^2-2v\Big(\langle\frac{g_k}{\|g_k\|}, x_k-x\rangle-\|x_k-x\|(R+\frac{\sqrt{n}\Delta}{v}) \\
&\quad -\frac{\big(v(1+R)+\sqrt{n}\Delta\big)^2}{2v}\Big),
\end{aligned} \tag{5.3.9}$$

where the second inequality follows from the boundedness of noise. ■

Before we discuss the convergence property of the QAQSGM which is the main result in this subsection, we consider the following lemma which shows the boundedness of the sequence $\{x_k\}$.

Lemma 5.3.5 *Suppose Assumptions 5.3.2-5.3.5 hold and $\{x_k\}$ is generated by the QAQSGM. Then, $\{x_k\}$ is bounded.*

Proof. Since $\overline{\lim_{k\to\infty}}\,\epsilon_k = \epsilon$, for any $\theta > 0$, there exists some positive integer k_0 such that

$$\epsilon_k < \epsilon+\theta, \forall k \geq k_0. \tag{5.3.10}$$

Define the maximum solution of $t^{g(t)} \le y_\theta^*/\eta$ by

$$T := \sup\{t \in \mathbb{R}_+ : t^{g(t)} \le y_\theta^*/\eta\}, \tag{5.3.11}$$

which is finite, since y_θ^* is finite (cf. Lemma 5.3.3(i)) and $\lim_{t\to\infty} t^{g(t)} = +\infty$. Next, we claim that the following inequality holds for all $i \ge k_0$:

$$\operatorname{dist}(x_i, X^*) \le \max\{\operatorname{dist}(x_{k_0}, X^*), T + v(1+R) + \sqrt{n}\Delta\}. \tag{5.3.12}$$

It is obvious that the relation (5.3.12) holds if $i = k_0$. Proving by induction, we assume the relation (5.3.12) holds for some $i = k$ ($\ge k_0$) and consider the following two cases.

Case 1. If $f(x_k) \le f_* + \mu\Big(\frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\frac{\sqrt{n}\Delta}{v} + R)(\frac{f(x_k)-f_*}{\eta})^{1/g(\operatorname{dist}(x_k,X^*))}\Big)^p + \epsilon_k$, together with (5.3.10), we have

$$\mu\Big(\frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\frac{\sqrt{n}\Delta}{v} + R)(\frac{f(x_k)-f_*}{\eta})^{1/g(\operatorname{dist}(x_k,X^*))}\Big)^p + \epsilon + \theta - (f(x_k) - f_*) \ge 0,$$

that is, $K_\theta^{x_k}(f(x_k) - f_*) \ge 0$. Hence, from (5.3.6), we obtain $f(x_k) - f_* \le y_\theta^*$, and thus

$$\operatorname{dist}(x_k, X^*)^{g(\operatorname{dist}(x_k,X^*))} \le y_\theta^*/\eta,$$

which follows from (5.3.4). Thus, we arrive at $\operatorname{dist}(x_k, X^*) < T$, which follows from (5.3.11), and thus the relation (5.3.2) implies

$$\begin{aligned}
\operatorname{dist}(x_{k+1}, X^*) &\le \operatorname{dist}(x_k - v\tilde{g}_k, X^*) + \sqrt{n}\Delta \\
&\le \operatorname{dist}(x_k, X^*) + v\|g_k/\|g_k\| + r_k\| + \sqrt{n}\Delta \\
&< T + v(1+R) + \sqrt{n}\Delta,
\end{aligned}$$

that is, the relation (5.3.12) holds for $i = k+1$.

Case 2. If $f(x_k) > f_* + \mu\Big(\frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\frac{\sqrt{n}\Delta}{v} + R)(\frac{f(x_k)-f_*}{\eta})^{1/g(\operatorname{dist}(x_k,X^*))}\Big)^p + \epsilon_k$, then it follows from Lemma 4.3.4 that

$$\begin{aligned}
\langle g_k/\|g_k\|, x_k - x^*\rangle &\ge \tfrac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\tfrac{\sqrt{n}\Delta}{v} + R)(\tfrac{f(x_k)-f_*}{\eta})^{1/g(\operatorname{dist}(x_k,X^*))} \\
&\ge \tfrac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\tfrac{\sqrt{n}\Delta}{v} + R)\operatorname{dist}(x_k, X^*),
\end{aligned}$$

where the second inequality follows from (5.3.4). Hence, applying Lemma 5.3.4 with $x^* = P_{X^*}(x_k)$, we obtain

$$\begin{aligned}
\big(\operatorname{dist}(x_{k+1}, X^*)\big)^2 &\le \|x_{k+1} - x^*\|^2 \\
&\le \|x_k - x^*\|^2 - 2v\Big(\tfrac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\tfrac{\sqrt{n}\Delta}{v} + R)\operatorname{dist}(x_k, X^*) \\
&\quad -\|x_k - x\|(\tfrac{\sqrt{n}\Delta}{v} + R) - \tfrac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2\Big) \\
&= \big(\operatorname{dist}(x_k, X^*)\big)^2.
\end{aligned}$$

Hence, the relation (5.3.12) holds for $i = k + 1$.

Therefore, by induction, the relation (5.3.12) holds for all $i \geq k_0$. Since the right hand side of (5.3.12) is finite and X^* is compact, $\{x_k\}$ is bounded. ■

From Lemma 5.3.5, $\{x_k\}$ is bounded and hence $\{f(x_k)\}$ is bounded from above by the Hölder condition (cf. (5.3.1)). We denote the upper bound on $\{f(x_k)\}$ by M in what follows.

Theorem 5.3.2 *Let Assumptions 5.3.2-5.3.5 hold. Then, for a sequence $\{x_k\}$ generated by the QAQSGM, we have*

$$\underline{\lim}_{k\to\infty} f(x_k) \leq f_* + y_0^*,$$

where y_0^* is finite.

Proof. The finiteness of y_0^* has been proved in Lemma 5.3.3(i). To prove the convergence property, we first show that

$$\underline{\lim}_{k\to\infty} f(x_k) < f_* + y_\theta^*$$

for all $\theta > 0$ by contradiction, that is, assume that the following inequality holds for some $\theta > 0$,

$$\underline{\lim}_{k\to\infty} f(x_k) \geq f_* + y_\theta^*.$$

Thus, there exists some $\delta \in (0, \min\{\theta/2, y_\theta^*\})$ and positive integer k_0 such that

$$f(x_k) > f_* + y_\theta^* - \delta, \tag{5.3.13}$$

and

$$\epsilon_k < \epsilon + \theta/2, \tag{5.3.14}$$

for all $k \geq k_0$, where (5.3.14) holds due to $\overline{\lim}_{k\to\infty} \epsilon_k = \epsilon$.

From (5.3.6) and (5.3.13), we obtain $f(x_k) - f_* + \delta > \sup\{y : K_\theta^{x_k}(y) \geq 0\}$ and thus $K_\theta^{x_k}(f(x_k) - f_* + \delta) < 0$, that is, for all $k \geq k_0$ there holds

$$\begin{aligned}
&f(x_k)\\
&> f_* + \mu\Big(\tfrac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\tfrac{\sqrt{n}\Delta}{v} + R)(\tfrac{f(x_k)-f_*+\delta}{\eta})^{1/g(\mathrm{dist}(x_k,X^*))}\Big)^p + \epsilon + \theta - \delta\\
&> f_* + \mu\Big(\tfrac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\tfrac{\sqrt{n}\Delta}{v} + R)(\tfrac{f(x_k)-f_*+\delta}{\eta})^{1/g(\mathrm{dist}(x_k,X^*))}\Big)^p + \epsilon_k\\
&\geq f_* + \mu\Big(\tfrac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\tfrac{\sqrt{n}\Delta}{v} + R)(\tfrac{f(x_k)-f_*}{\eta})^{1/g(\mathrm{dist}(x_k,X^*))} + \delta'\Big)^p + \epsilon_k,
\end{aligned}$$

where the second inequality follows from (5.3.14) and $0 < \delta < \theta/2$, and the third inequality follows from the Taylor expansion with some positive scalar $\delta' = (\frac{\sqrt{n}\Delta}{v} + R)\min\{\frac{\delta}{\eta q}(\frac{y_\theta^*}{\eta})^{1/q-1}, \frac{\delta}{\eta p}(\frac{M-f_*}{\eta})^{1/p-1}\}$ (recall that M is an upper bound on $\{f(x_k)\}$). Therefore, by using Lemmas 4.3.4 and 5.3.4, we obtain

$$\begin{aligned}\langle g_k/\|g_k\|, x_k - x^*\rangle &\geq \tfrac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\tfrac{\sqrt{n}\Delta}{v} + R)(\tfrac{f(x_k)-f_*}{\eta})^{1/g(\mathrm{dist}(x_k,X^*))} + \delta' \\ &\geq \tfrac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\tfrac{\sqrt{n}\Delta}{v} + R)\,\mathrm{dist}(x_k, X^*) + \delta', \forall k \geq k_0,\end{aligned}$$

and hence

$$\begin{aligned}\big(\mathrm{dist}(x_{k+1}, X^*)\big)^2 &\leq \big(\mathrm{dist}(x_k, X^*)\big)^2 - 2v\delta' \\ &\leq \cdots \leq \big(\mathrm{dist}(x_0, X^*)\big)^2 - 2(k - k_0 + 1)v\delta',\end{aligned}$$

which yields a contradiction for sufficiently large k. Thus, we have

$$\underline{\lim}_{k\to\infty} f(x_k) \leq f_* + y_\theta^*, \forall \theta > 0.$$

Taking the limit as $\theta \to 0$, applying Lemma 5.3.3(ii), we arrive at the conclusion. ■

We now give an explicit expression for the tolerance in approaching f_* in Theorem 5.3.2 in a specific case of p and $g(t)$. By solving relations (5.3.7)-(5.3.8), we obtain the following corollary where the total error is given in an explicit expression.

Corollary 5.3.1 *Let Assumptions 5.3.2-5.3.5 hold with $g(t) \equiv p$ and $p = 1$. Then, for a sequence $\{x_k\}$ generated by the QAQSGM, we have*

$$\underline{\lim}_{k\to\infty} f(x_k) \leq f_* + \eta\frac{\mu\big(\sqrt{n}\Delta + v(1+R)\big)^2 + 2v\epsilon}{2(v\eta - v\mu R - \mu\sqrt{n}\Delta)}.$$

Proof. By the assumptions, since $g(t) \equiv p$ and $p = q = 1$, we have

$$K_0^p(y) = K_0^q(y) = \mu\Big(\frac{1}{2v}\big(\sqrt{n}\Delta + v(1+R)\big)^2 + (\frac{\sqrt{n}\Delta}{v} + R)\frac{y}{\eta}\Big) + \epsilon - y,$$

and

$$y_0^p = y_0^q.$$

It is clear that $K_0^p(y)$ is linear and decreasing due to $R + \sqrt{n}\Delta/v < \eta/\mu$. Thus, by (5.3.7), y_0^p is just the solution of $K_0^p(y) = 0$. Thus, from (5.3.8), we have

$$y_0^* = y_0^p = \eta\frac{\mu\big(\sqrt{n}\Delta + v(1+R)\big)^2 + 2v\epsilon}{2(v\eta - v\mu R - \mu\sqrt{n}\Delta)}.$$

Hence, by Theorem 5.3.2, we arrive at the conclusion. ■

It is also observed that the total error, given in Corollary 5.3.1, has the same expression as that of the AQSGM (cf. Corollary 4.4.1) if the quantization operator is infinitely precise (i.e., Δ is sufficiently small).

Chapter 6

Conclusion and Future Work

In this thesis, we proposed several new types of subgradient methods, investigated convergence properties of the proposed algorithms, and illustrated the high efficiency and wide applicability by numerical experiments for both convex and quasi-convex optimization problems.

Based on the GS technique, we proposed the GS-SGM and GS-DSGM to solve nondifferentiable convex (constrained) optimization problems. Using both the constant and divergent stepsize rules, we proved that our proposed algorithms converge to an (approximate) optimal value/solution with probability 1. Numerical results demonstrate that the GS technique improves the convergence behavior of subgradient methods, especially for the low-rank recovery problems.

To meet much broader class of problems, we considered using a generic inexact subgradient method (AQSGM) to solve nondifferentiable quasi-convex constrained optimization problems. Assuming that the computational errors and noise are deterministic and bounded, we studied the effect of the inexact terms on subgradient methods when the constraint set is compact or when the objective function satisfies the generalized weak sharp minima condition. In both cases, using both the constant and diminishing stepsize rules, we described convergence results in both objective values and iterates, where the tolerances are given explicitly in terms of errors and noise, finite convergence to the approximate optimal value and efficiency estimates of iterates. We also proposed and analyzed the QAQSGM for a quantized quasi-convex constrained

optimization problem. Several numerical experiments illustrate that the AQSGM is comparable with some existing algorithms, and suitable for large-scale problems.

Based on these results and contributions, there are many other issues that are needed to deal with in the future work. We summarize three directions for my future work as follows. (i) We will extend the quasi-subgradient method by using other types of stepsize rules and other types of quasi-subgradients. (ii) Nedić and Bertsekas [80] investigated an increment subgradient method to solve the convex optimization problem, where the objective function is a summation of a number of component convex functions, and the high efficiency of the incremental approach was illustrated in [14, 18, 80, 108]. However, the convergence property of the incremental quasi-subgradient method still remains an open question. The difficulty stems from the limitation of the definition of the quasi-subdifferential and the fact that the summation of some quasi-convex functions may not be quasi-convex. We will try to define some new type of quasi-subgradient and investigate the convergence property of the incremental quasi-subgradient method. (iii) Auslender and Teboulle [2, 4, 5] designed the interior subgradient method for convex optimization problems, and Langenberg and Tichatschke [65] proposed an interior proximal point method to solve quasi-convex optimization problems. Since there is a close link between the proximal point method and the subgradient method, we will propose and investigate an interior quasi-subgradient method for quasi-convex optimization problems in the future work. This could be an interesting research topic in the future.

Bibliography

[1] A. Auslender and M. Teboulle. Entropic proximal decomposition methods for convex programs and variational inequalities. *Mathematical Programming*, 91:33–47, 2001.

[2] A. Auslender and M. Teboulle. Interior gradient and epsilon-subgradient descent methods for constrained convex minimization. *Mathematics of Operations Research*, 29(1):1–26, 2004.

[3] A. Auslender and M. Teboulle. Interior projection-like methods for monotone variational inequalities. *Mathematical Programming*, 104:39–68, 2005.

[4] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16(3):697–725, 2006.

[5] A. Auslender and M. Teboulle. Projected subgradient methods with non-euclidean distances for non-differentiable convex minimization and variational inequalities. *Mathematical Programming*, 120:27–48, 2009.

[6] A. Auslender, M. Teboulle, and S. Ben-Tiba. A logarithmic-quadratic proximal method for variational inequalities. *Computational Optimization and Applications*, 12:31–40, 1999.

[7] D. Aussel, J.-N. Corvellec, and M. Lassonde. Mean value property and subdifferential criteria for lower semicontinuous functions. *Transactions of the American Mathematical Society*, 347(10):4147–4161, 1995.

[8] D. Aussel and A. Daniilidis. Normal characterization of the main classes of quasiconvex functions. *Set-Valued Analysis*, 8:219–236, 2000.

[9] M. Avriel, W. E. Diewert, S. Schaible, and I. Zang. *Generalized Concavity*. Plenum Press, New York, 1988.

[10] T. C. Aysal, M. J. Coates, and M. G. Rabbat. Distributed average consensus with dithered quantization. *IEEE Transactions on Signal Processing*, 56:4905–4918, 2008.

[11] M. S. Bazaraa and H. D. Sherali. On the choice of step size in subgradient optimization. *European Journal of Operational Research*, 7(4):380–388, 1981.

[12] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[13] A. Ben-tal, T. Margalit, and A. Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12:79–108, 2001.

[14] D. P. Bertsekas. A new class of incremental gradient methods for least squares problems. *SIAM Journal on Optimization*, 7(4):913–926, 1997.

[15] D. P. Bertsekas. *Network Optimization: Continuous and Discrete Models*. Athena Scientific, Massachusetts, 1998.

[16] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Cambridge, 1999.

[17] D. P. Bertsekas, A. Nedić, , and A. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, Cambridge, 2003.

[18] D. P. Bertsekas and J. N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

[19] J. F. Bonnans and A. Ioffe. Second-order sufficiency and quadratic growth for nonisolated minima. *Mathematics of Operations Research*, 20(4):801–817, 1995.

[20] J. Borde and J.-P. Crouzeix. Continuity properties of the normal cone to the level sets of a quasiconvex function. *Journal of Optimization Theory and Applications*, 66:415–429, 1990.

[21] S. Boyd and L. Vandenbergh. *Convex Optimization*. Cambridge University Press, New York, 2004.

[22] P. S. Bradley, U. M. Fayyad, and O. L. Mangasarian. Mathematical programming for data mining: Formulations and challenges. *INFORMS Journal on Computing*, 11:217–238, 1998.

[23] R. S. Burachik, R. N. Gasimov, N. A. Ismayilova, and C. Y. Kaya. On a modified subgradient algorithm for dual problems via sharp augmented Lagrangian. *Journal of Global Optimization*, 34:55–78, 2006.

[24] R. S. Burachik, C. Y. Kaya, and M. Mammadov. An inexact modified subgradient algorithm for nonconvex optimization. *Computational Optimization and Applications*, 45(1):1–24, 2010.

[25] J. V. Burke and S. Deng. Weak sharp minima revisited, part I: Basic theory. *Control and Cybernetics*, 31:439–469, 2002.

[26] J. V. Burke and S. Deng. Weak sharp minima revisited, part II: Application to linear regularity and error bounds. *Mathematical Programming*, 104:235–261, 2005.

[27] J. V. Burke and S. Deng. Weak sharp minima revisited, part III: Error bounds for differentiable convex inclusions. *Mathematical Programming*, 116:37–56, 2009.

[28] J. V. Burke and M. C. Ferris. Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, 31(5):1340–1359, 1993.

[29] J. V. Burke, A. S. Lewis, and M. L. Overton. Approximating subdifferentials by random sampling of gradients. *Mathematics of Operations Research*, 27:567–584, 2002.

[30] J. V. Burke, A. S. Lewis, and M. L. Overton. Two numerical methods for optimizing matrix stability. *Linear Algebra and its Applications*, 351-352:117–145, 2002.

[31] J. V. Burke, A. S. Lewis, and M. L. Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2005.

[32] F. Clarke, Y. Ledyaev, R. Stern, and P. Wolenski. *Nonsmooth Analysis and Control Theory*. Springer-Verlag, New York, 1998.

[33] J.-P. Crouzeix. Conditions for convexity of quasiconvex functions. *Mathematics of Operations Research*, 5(1):120–125, 1980.

[34] J.-P. Crouzeix and J. A. Ferland. Algorithms for generalized fractional programming. *Mathematical Programming*, 52:191–207, 1991.

[35] J.-P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle. *Generalized Convexity, Generalized Monotonicity*. Kluwer Academic Publishers, Dordrecht, 1998.

[36] G. D'Antonio and A. Frangioni. Convergence analysis of deflected conditional approximate subgradient methods. *SIAM Journal on Optimization*, 20(1):357–386, 2009.

[37] M. E. Dyer. Calculating surrogate constraints. *Mathematical Programming*, 19:255–278, 1980.

[38] Y. M. Ermoliev. Methods of solution of nonlinear extremal problems. *Cybernetics and Systems Analysis*, 2:1–14, 1966.

[39] J. A. Ferland and J.-Y. Potvin. Generalized fractional programming: Algorithms and numerical experimentation. *European Journal of Operational Research*, 20(1):92–101, 1985.

[40] R. N. Gasimov. Augmented Lagrangian duality and nondifferentiable optimization methods in nonconvex programming. *Journal of Global Optimization*, 24:187–203, 2002.

[41] A. A. Goldstein. Convex programming in Hilbert space. *Bulletin of The American Mathematical Society*, 70:709–711, 1964.

[42] H. J. Greenberg and W. P. Pierskalla. A review of quasi-convex functions. *Operations Research*, 19(7):1553–1570, 1971.

[43] H. J. Greenberg and W. P. Pierskalla. Quasiconjugate functions and surrogate duality. *Cahiers Centre Études Recherche Opertionnelle*, 15:437–448, 1973.

[44] N. Hadjisavvas, S. Komlósi, and S. Schaible. *Handbook of Generalized Convexity and Generalized Monotonicity*. Springer-Verlag, New York, 2005.

[45] J. Haslinger and P. Neittaanmaki. *Finite Element Approximation for Optimal Shape, Material and Topology Design*. Wiley, Chichester, 1996.

[46] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer-Verlag, Berlin, 1996.

[47] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer-Verlag, Berlin, 2001.

[48] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, 1985.

[49] X. X. Huang and X. Q. Yang. Efficiency and approachability of nonconvex bicriteria programs. *Journal of Mathematical Analysis and Applications*, 261(1):134–150, 2001.

[50] V. Jeyakumar and G. Li. Robust duality for fractional programming problems with constraint-wise data uncertainty. *Journal of Optimization Theory and Applications*, 151:292–303, 2011.

[51] V. Jeyakumar, A. Rubinov, and Z. Wu. Non-convex quadratic minimization problems with quadratic constraints: Global optimality conditions. *Mathematical Programming*, 110:521–541, 2007.

[52] B. Johansson, M. Rabi, and M. Johansson. A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization*, 20(3):1157–1170, 2009.

[53] T. Kärkkäinen and E. Heikkola. Robust formulations for training multilayer perceptrons. *Neural Computation*, 16:837–862, 2004.

[54] A. Kashyap, T. Basar, and R. Srikant. Quantized consensus. *Automatica*, 43:1192–1203, 2007.

[55] K. C. Kiwiel. *Methods of Descent for Nondifferentiable Optimization*. Springer-Verlag, Berlin, 1985.

[56] K. C. Kiwiel. Proximity control in bundle methods for convex nondifferentiable minimization. *Mathematical Programming*, 46:105–122, 1990.

[57] K. C. Kiwiel. The efficiency of subgradient projection methods for convex optimization, part I: General level methods. *SIAM Journal on Control and Optimization*, 34(2):660–676, 1996.

[58] K. C. Kiwiel. The efficiency of subgradient projection methods for convex optimization, part II: Implementations and extensions. *SIAM Journal on Control and Optimization*, 34(2):677–697, 1996.

[59] K. C. Kiwiel. Efficiency of proximal bundle methods. *Journal of Optimization Theory and Applications*, 104:589–603, 2000.

[60] K. C. Kiwiel. Convergence and efficiency of subgradient methods for quasiconvex minimization. *Mathematical Programming*, 90:1–25, 2001.

[61] K. C. Kiwiel. Convergence of approximate and incremental subgradient methods for convex optimization. *SIAM Journal on Optimization*, 14(3):807–840, 2004.

[62] K. C. Kiwiel. Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 18(2):379–388, 2007.

[63] K. C. Kiwiel. A nonderivative version of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 20(4):1983–1994, 2010.

[64] I. V. Konnov. On properties of supporting and quasi-supporting vectors. *Journal of Mathematical Sciences*, 71:2760–2763, 1994.

[65] N. Langenberg and R. Tichatschke. Interior proximal methods for quasiconvex optimization. *Journal of Global Optimization*, 52(3):641–661, 2012.

[66] T. Larsson, Z. Liu, and M. Patriksson. A dual scheme for traffic assignment problems. *Optimization*, 42(4):323–358, 1997.

[67] T. Larsson, M. Patriksson, and A.-B. Ströberg. Ergodic, primal convergence in dual subgradient schemes for convex programming. *Mathematical Programming*, 86:283–312, 1999.

[68] T. Larsson, M. Patriksson, and A.-B. Strömberg. On the convergence of conditional epsilon-subgradient methods for convex programs and convex-concave saddle-point problems. *European Journal of Operational Research*, 151(3):461–473, 2003.

[69] C. Lemaréchal, A. Nemirovski, and Y. Nesterov. New variants of bundle methods. *Mathematical Programming*, 69:111–147, 1995.

[70] E. S. Levitin and B. T. Polyak. Constrained minimization problems. *USSR Computational Mathematics and Mathematical Physics*, 6:1–50, 1966.

[71] A. S. Lewis. The mathematics of eigenvalue optimization. *Mathematical Programming*, 97(1-2):155–176, 2003.

[72] S. J. Li, X. Q. Yang, and G. Y. Chen. Nonconvex vector optimization of set-valued mappings. *Journal of Mathematical Analysis and Applications*, 283(2):337–350, 2003.

[73] D. G. Luenberger. Quasi-convex programming. *SIAM Journal on Applied Mathematics*, 16(5):1090–1095, 1968.

[74] Z.-Q. Luo and P. Tseng. Analysis of an approximate gradient projection method with applications to the backpropagation algorithm. *Optimization Methods and Software*, 4(2):85–101, 1994.

[75] M. Makela. Survey of bundle methods for nonsmooth optimization. *Optimization Methods and Software*, 17(1):1–29, 2002.

[76] O. L. Mangasarian and M. V. Solodov. Serial and parallel backpropagation convergence via nonmonotone perturbed minimization. *Optimization Methods and Software*, 4:103–116, 1994.

[77] J. E. Martínez-Legaz and P. H. Sach. A new subdifferential in quasiconvex analysis. *Journal of Convex Analysis*, 6(1):1–12, 1999.

[78] E. Mijangos. Approximate subgradient methods for nonlinearly constrained network flow problems. *Journal of Optimization Theory and Applications*, 128:167–190, 2006.

[79] J. Moreau, P. Panagiotopoulos, and G. Strang. *Topics in Nonsmooth Mechanics*. Birkhauser, Basel, 1988.

[80] A. Nedić and D. P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001.

[81] A. Nedić and D. P. Bertsekas. The effect of deterministic noise in subgradient methods. *Mathematical Programming*, 125:75–99, 2010.

[82] A. Nedić and A. Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization*, 19(4):1757–1780, 2009.

[83] A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142:205–228, 2009.

[84] A. S. Nemirovsky and D. B. Yudin. Cesare convergence of gradient method approximation of saddle points for convex-concave functions. *Doklady Akademii Nauk SSSR*, 239:1056–1059, 1978.

[85] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, 1983.

[86] Y. Nesterov. *Effective Methods in Nonlinear Programming*. Radio i Svyaz, Moscow, 1989.

[87] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2009.

[88] E. S. H. Neto and A. R. D. Pierro. Incremental subgradients for constrained convex optimization: A unified framework and new methods. *SIAM Journal on Optimization*, 20(3):1547–1572, 2009.

[89] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Science+Business Media, New York, 2006.

[90] J. Outrata, M. Kocvara, and J. Zowe. *Nonsmooth Approach to Optimization Problems With Equilibrium Constraints*. Kluwert Academic Publisher, Dordrecht, 1998.

[91] J.-S. Pang. Error bounds in mathematical programming. *Mathematical Programming*, 79(1-3):299–332, 1997.

[92] B. T. Polyak. A general method for solving extremum problems. *Soviet Mathematics Doklady*, 8:593–597, 1967.

[93] B. T. Polyak. Nonlinear programming methods in the presence of noise. *Mathematical Programming*, 14:87–97, 1978.

[94] B. T. Polyak. *Introduction to Optimization*. Optimization Software, New York, 1987.

[95] M. G. Rabbat and R. D. Nowak. Quantized incremental algorithms for distributed optimization. *IEEE Journal on Selected Areas in Communications*, 23(4):798–808, 2005.

[96] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

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

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

[99] A. M. Rubinov and X. Q. Yang. *Lagrange-type Functions in Constrained Non-Convex Optimization*. Kluwer Academic Publishers, 2003.

[100] A. Ruszczyński. A merit function approach to the subgradient method with averaging. *Optimization Methods and Software*, 23:161–172, 2008.

[101] C. Sagastizábal and M. V. Solodov. An infeasible bundle method for nonsmooth convex constrained optimization without a penalty function or a filter. *SIAM Journal on Optimization*, 16(1):146–169, 2005.

[102] H. Schramm and J. Zowe. A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results. *SIAM Journal on Optimization*, 2(1):121–152, 1992.

[103] S. Sen and H. D. Sherali. A class of convergent primal-dual subgradient algorithms for decomposable convex programs. *Mathematical Programming*, 35:279–297, 1986.

[104] H. D. Sherali and G. Choi. Recovery of primal solutions when using subgradient optimization methods to solve Lagrangian duals of linear programs. *Operations Research Letters*, 19(3):105–113, 1996.

[105] Q. Shi, C. He, and L. Jiang. Normalized incremental subgradient algorithm and its application. *IEEE Transactions on Signal Processing*, 57(10):3759–3774, 2009.

[106] N. Z. Shor, K. C. Kiwiel, and A. Ruszczyński. *Minimization Methods for Non-differentiable Functions*. Springer-Verlag, New York, 1985.

[107] M. Studniarski and D. E. Ward. Weak sharp minima: Characterizations and sufficient conditions. *SIAM Journal on Control and Optimization*, 38(1):219–236, 1999.

[108] P. Tseng. An incremental gradient(-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, 1998.

[109] H. Xu. Level function method for quasiconvex programming. *Journal of Optimization Theory and Applications*, 108:407–437, 2001.

[110] X. Y. Zheng and K. F. Ng. Metric regularity and constraint qualifications for convex inequalities on banach spaces. *SIAM Journal on Optimization*, 14(3):757–772, 2003.

[111] X. Y. Zheng and K. F. Ng. Error bound moduli for conic convex systems on banach spaces. *Mathematics of Operations Research*, 29(2):213–228, 2004.