UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCES
DEPARTMENT OF MATHEMATICS
AND INFORMATICS

Modifications of Newton-type methods for solving semi-smooth stochastic optimization problems

- PhD thesis -

Modifikacije metoda Njutnovog tipa za rešavanje semi-glatkih problema stohastičke optimizacije

- Doktorska disertacija -

Supervisor:
Prof. Dr. Nataša Krejić

Candidate:
Tijana Ostojić

Novi Sad, 2023

УНИВЕРЗИТЕТ У НОВОМ САДУ **ОБРАЗАЦ – 5а**
НАВЕСТИ НАЗИВ ФАКУЛТЕТА ИЛИ ЦЕНТРА

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА[1]

Врста рада:	Докторска дисертација
Име и презиме аутора:	Тијана Остојић
Ментор (титула, име, презиме, звање, институција)	Др Наташа Крејић, редовни професор, Универзитет у Новом Саду Природно-математички факултет
Наслов рада:	Модификације метода Њутновог типа за решавање семи-глатких проблема стохастичке оптимизације
Језик публикације (писмо):	Енглески језик
Физички опис рада:	Унети број: Страница: 188 Поглавља: 7 Референци: 107 Табела: 3 Слика: 10 Графикона: 0 Прилога: 0
Научна област:	Математика
Ужа научна област (научна дисциплина):	Нумеричка математика
Кључне речи / предметна одредница:	Нумеричка оптимизација, семи-глатка оптимизација, стохастичка оптимизација, узорачко очекивање, методе линијског претраживања, променљива величина узорка
Резиме на српском језику:	У бројним проблемима оптимизације који потичу из стварних и научних примена, често се суочавамо са недиференцијабилношћу. У ову класу спада велики број проблема, од модела природних феномена који показују нагле промене, оптимизације облика, до функције циља у машинском учењу и дубоким неуронским мрежама. У пракси, решавање семи-глатких конвексних проблема обично је изазовније и захтева веће рачунске трошкове у односу на глатке проблеме. Циљ ове тезе је формулација и теоријска анализа метода Њутновог типа за решавање семи-глатких конвексних стохастичких проблема оптимизације. Разматрани су проблеми оптимизације са функцијом циља датом у облику математичког очекивања без претпоставке о диференцијабилности функције.

[1] Аутор докторске дисертације потписао је и приложио следеће Обрасце:
5б – Изјава о ауторству;
5в – Изјава о истоветности штампане и електронске верзије и о личним подацима;
5г – Изјава о коришћењу.
Ове Изјаве се чувају на факултету у штампаном и електронском облику и не кориче се са тезом.

	Како је врло тешко, па некад чак и немогуће одредити аналитички облик математичког очекивања, функција циља се апроксимира узорачким очекивањем. Имајући у виду да је тачност апроксимације функције циља и њених извода пропорционална рачунским трошковима – већа прецизност подразумева веће трошкове у општем случају, важно је дизајнирати ефикасан баланс између тачности и трошкова. Стога, главни фокус ове тезе је развој алгоритама базираних на одређивању оптималне динамике увећања узорка у семи-глатком окружењу, са посебном пажњом на контроли тачности и одабиру праваца претраге. По питању одабира правца, размотрено је неколико опција, док контрола тачности укључује јефтиније апроксимације функције циља (са мањом прецизношћу) током почетних фаза процеса да би се уштедели рачунски напори. Овај приступ има за циљ очување рачунских ресурса, резервишући примену апроксимација функције циља високе тачности за завршне фазе процеса оптимизације. Детаљан опис предложених метода представљен је у поглављима 5 и 6, где су анализиране и теоријске особине нумеричких поступака, тј. доказана је њихова конвергенција и приказана сложеност развијених метода. Поред теоријског оквира, потврђена је успешна практична имплементација датих алгоритама. Показано је да су предложене методе ефикасније у практичној примени у односу на постојеће методе из литературе. Поглавље 1 ове тезе служи као основа за праћење наредних поглавља пружајући преглед основних појмова. Поглавље 2 се односи на нелинеарну оптимизацију, при чему је посебан акценат стављен на технике линијског претраживања. У поглављу 3 фокус се помера на семи-глатке проблеме оптимизације и методе за њихово решавање и служи као преглед постојећих резултата из ове области. Преостали делови тезе, почевши од поглавља 4, где се уводи проблем изучавања ове тезе (минимизација функције дате у облику очекиване вредности), па надаље, представљају оригинални допринос аутора.
Датум прихватања теме од стране надлежног већа:	11.02.2021.
Датум одбране: (Попуњава одговарајућа служба)	
Чланови комисије: (титула, име, презиме, звање, институција)	Председник: др Наташа Крклец Јеринкић, ванредни професор, Универзитет у Новом Саду, Природно-математички факултет Члан: др Наташа Крејић, редовни професор, Универзитет у Новом Саду, Природно-математички факултет Члан: др Сања Рапајић, редовни професор, Универзитет у Новом Саду, Природно-математички факултет Члан: др Марко Виола, доцент, Универзитетски колеџ Даблин, Ирска
Напомена:	

UNIVERSITY OF NOVI SAD
FACULTY OR CENTER

KEY WORD DOCUMENTATION[2]

Document type:	Doctoral dissertation
Author:	Tijana Ostojić
Supervisor (title, first name, last name, position, institution)	Dr. Nataša Krejić, professor, University of Novi Sad Faculty of Sciences
Thesis title:	Modifications of Newton-type methods for solving semi-smooth stochastic optimization problems
Language of text (script):	English
Physical description:	Number of: Pages: 188 Chapters: 7 References: 107 Tables: 3 Illustrations: 10 Graphs: 0 Appendices: 0
Scientific field:	Mathematics
Scientific subfield (scientific discipline):	Numerical mathematics
Subject, Key words:	Numerical optimization, semi-smooth optimization, stochastic optimization, sample average approximation, line search methods, variable sample size
Abstract in English language:	In numerous optimization problems originating from real-world and scientific applications, we often face nonsmoothness. A large number of problems belong to this class, from models of natural phenomena that exhibit sudden changes, shape optimization, to hinge loss functions in machine learning and deep neural networks. In practice, solving a nonsmooth convex problem tends to be more challenging, usually more difficult and costly than a smooth one. The aim of this thesis is the formulation and theoretical analysis of Newton-type algorithms for solving nonsmooth convex stochastic optimization problems. The optimization problems with the objective function given in the form of a mathematical expectation without differentiability assumption of the function are considered.

[2] The author of doctoral dissertation has signed the following Statements:
5б – Statement on the authority,
5в – Statement that the printed and e-version of doctoral dissertation are identical and about personal data,
5г – Statement on copyright licenses.
The paper and e-versions of Statements are held at he faculty and are not included into the printed thesis.

	The Sample Average Approximation (SAA) is used to estimate the objective function. As the accuracy of the SAA objective functions and its derivatives is naturally proportional to the computational costs – higher precision implies larger costs in general, it is important to design an efficient balance between accuracy and costs. Therefore, the main focus of this thesis is the development of adaptive sample size control algorithms in a nonsmooth environment, with particular attention given to the control of the accuracy and selection of search directions. Several options are investigated for the search direction, while the accuracy control involves cheaper objective function approximations (with looser accuracy) during the initial stages of the process to save computational effort. This approach aims to conserve computational resources, reserving the deployment of high-accuracy objective function approximations for the final stages of the optimization process. A detailed description of the proposed methods is presented in Chapter 5 and 6. Also, the theoretical properties of the numerical procedures are analyzed, i.e., their convergence is proved, and the complexity of the developed methods is studied. In addition to the theoretical framework, the successful practical implementation of the given algorithms is presented. It is shown that the proposed methods are more efficient in practical application compared to the existing methods from the literature. Chapter 1 of this thesis serves as a foundation for the subsequent chapters by providing the necessary background information. Chapter 2 covers the fundamentals of nonlinear optimization, with a particular emphasis on line search techniques. In Chapter 3, the focus shifts to the nonsmooth framework. This chapter serves the purpose of reviewing the existing knowledge and established results in the field. The remaining sections of the thesis, starting from Chapter 4, where the framework for the subject of this thesis (the minimization of the expected value function) is introduced, onwards, represent the original contribution made by the author.
Accepted on Scientific Board on:	11.02.2021.
Defended: (Filled by the faculty service)	
Thesis Defend Board: (title, first name, last name, position, institution)	President: Dr. Nataša Krklec Jerinkić, associate professor, University of Novi Sad, Faculty of Sciences Member: Dr. Nataša Krejić, full professor, University of Novi Sad, Faculty of Sciences Member: Dr. Sanja Rapajić, full professor, University of Novi Sad, Faculty of Sciences Member: Dr. Marco Viola, assistant professor, University College Dublin, Ireland
Note:	

Abstract

In numerous optimization problems originating from real-world and scientific applications, we often face nonsmoothness. A large number of problems belong to this class, from models of natural phenomena that exhibit sudden changes, shape optimization, to hinge loss functions in machine learning and deep neural networks. In practice, solving a nonsmooth convex problem tends to be more challenging, usually more difficult and costly than a smooth one.

The aim of this thesis is the formulation and theoretical analysis of Newton-type algorithms for solving nonsmooth convex stochastic optimization problems. The optimization problems with the objective function given in the form of a mathematical expectation without differentiability assumption of the function are considered.

The Sample Average Approximation (SAA) is used to estimate the objective function. As the accuracy of the SAA objective functions and its derivatives is naturally proportional to the computational costs – higher precision implies larger costs in general, it is important to design an efficient balance between accuracy and costs. Therefore, the main focus of this thesis is the development of adaptive sample size control algorithms in a nonsmooth environment, with particular attention given to the control of the accuracy and selection of search directions. Several options are investigated for the search direction, while the accuracy control involves cheaper objective function approximations (with looser accuracy) during the initial stages of the process to save computational effort. This approach aims to conserve computational resources, reserving the deployment of high-accuracy objective function approximations for the final stages of the optimization process. A detailed description of the proposed methods is presented in Chapter 5 and 6. Also, the theoretical properties of the numeri-

cal procedures are analyzed, i.e., their convergence is proved, and the complexity of the developed methods is studied. In addition to the theoretical framework, the successful practical implementation of the given algorithms is presented. It is shown that the proposed methods are more efficient in practical application compared to the existing methods from the literature.

Chapter 1 of this thesis serves as a foundation for the subsequent chapters by providing the necessary background information. Chapter 2 covers the fundamentals of nonlinear optimization, with a particular emphasis on line search techniques. In Chapter 3, the focus shifts to the nonsmooth framework. This chapter serves the purpose of reviewing the existing knowledge and established results in the field. The remaining sections of the thesis, starting from Chapter 4, where the framework for the subject of this thesis (the minimization of the expected value function) is introduced, onwards, represent the original contribution made by the author.

Apstrakt

U brojnim problemima optimizacije koji potiču iz stvarnih i naučnih primena, često se suočavamo sa nediferencijabilnošću. U ovu klasu spada veliki broj problema, od modela prirodnih fenomena koji pokazuju nagle promene, optimizacije oblika, do funkcije cilja u mašinskom učenju i dubokim neuronskim mrežama. U praksi, rešavanje semi-glatkih konveksnih problema obično je izazovnije i zahteva veće računske troškove u odnosu na glatke probleme.

Cilj ove teze je formulacija i teorijska analiza metoda Njutnovog tipa za rešavanje semi-glatkih konveksnih stohastičkih problema optimizacije. Razmatrani su problemi optimizacije sa funkcijom cilja datom u obliku matematičkog očekivanja bez pretpostavke o diferencijabilnosti funkcije.

Kako je vrlo teško, pa nekad čak i nemoguće odrediti analitički oblik matematičkog očekivanja, funkcija cilja se aproksimira uzoračkim očekivanjem. Imajući u vidu da je tačnost aproksimacije funkcije cilja i njenih izvoda proporcionalna računskim troškovima – veća preciznost podrazumeva veće troškove u opštem slučaju, važno je dizajnirati efikasan balans između tačnosti i troškova. Stoga, glavni fokus ove teze je razvoj algoritama baziranih na određivanju optimalne dinamike uvećanja uzorka u semi-glatkom okruženju, sa posebnom pažnjom na kontroli tačnosti i odabiru pravaca pretrage. Po pitanju odabira pravca, razmotreno je nekoliko opcija, dok kontrola tačnosti uključuje jeftinije aproksimacije funkcije cilja (sa manjom preciznošću) tokom početnih faza procesa da bi se uštedeli računski napori. Ovaj pristup ima za cilj očuvanje računskih resursa, rezervišući primenu aproksimacija funkcije cilja visoke tačnosti za završne faze procesa optimizacije. Detaljan opis predloženih metoda predstavljen je u poglavljima 5 i 6, gde su analizirane i teorijske osobine numeričkih pos-

tupaka, tj. dokazana je njihova konvergencija i prikazana složenost razvijenih metoda. Pored teorijskog okvira, potvrđena je uspešna praktična implementacija datih algoritama. Pokazano je da su predložene metode efikasnije u praktičnoj primeni u odnosu na postojeće metode iz literature.

Poglavlje 1 ove teze služi kao osnova za praćenje narednih poglavlja pružajući pregled osnovnih pojmova. Poglavlje 2 se odnosi na nelinearnu optimizaciju, pri čemu je poseban akcenat stavljen na tehnike linijskog pretraživanja. U poglavlju 3 fokus se pomera na semi-glatke probleme optimizacije i metode za njihovo rešavanje i služi kao pregled postojećih rezultata iz ove oblasti. Preostali delovi teze, počevši od poglavlja 4, gde se uvodi problem izučavanja ove teze (minimizacija funkcije date u obliku očekivane vrednosti), pa nadalje, predstavljaju originalni doprinos autora.

Acknowledgments

I would like to take this opportunity to express my deepest gratitude and appreciation to all those who have supported me throughout my PhD studies and the completion of this thesis.

First and foremost, I would express my gratitude to my supervisor, Prof. Dr. Nataša Krejić, for her invaluable help, constant support, and guidance throughout my academic journey. I am immensely thankful for the opportunity to study and conduct research in nonsmooth optimization under her mentorship. Her guidance, insightful feedback, and encouragement have significantly shaped the trajectory of my research work, fostering both intellectual and personal growth.

Then, but no less importantly, I would like to express my sincere gratitude to Prof. Dr. Nataša Krklec Jerinkić for all the patience, energy, support, and generosity in sharing her knowledge and experiences with me. Her expertise, dedication, and belief in my abilities have continually inspired and motivated me to strive for excellence. I am also grateful to the members of my thesis committee, Prof. Dr. Sanja Rapajić, and Dr. Marco Viola for their time, effort, and valuable feedback which enhanced the quality of this thesis.

Last but certainly not least, I am grateful to my family and friends for the love, countless sacrifices, and support I have received on this remarkable journey. Your encouragement and belief in my dreams have been invaluable. Tara and Stefan, your love, hugs, and smiles provided the motivation I needed to keep going. I hope this journey has instilled in you the importance of pursuing your passions and never giving up on your dreams. This achievement is as much yours as it is mine.

With all my love and gratitude,
Tijana Ostojić
Novi Sad, September, 2023

Introduction

The optimization of complex systems under uncertainty is a challenging problem that arises in various fields, ranging from engineering and finance to machine learning and operations research. In many real-world scenarios, the presence of stochastic elements and the semi-smoothness of objective functions introduce complications in the optimization process. Therefore, specialized techniques are required to handle those complications that arisen due to the uncertainty and semi-smoothness.

This thesis aims to investigate and propose modifications of Newton-type methods specifically tailored for solving semi-smooth stochastic optimization problems given in the form of mathematical expectation. More precisely, for solving an SAA (Sample Average Approximation) reformulation of the original stochastic problems. Newton-type methods are widely recognized for their efficiency and rapid convergence properties in smooth optimization settings. However, their direct application to semi-smooth and stochastic problems is not possible, so Newton-type methods need to be modified significantly. The primary motivation behind this research is to enhance the existing optimization algorithms and develop novel techniques that can effectively handle the characteristics of semi-smoothness and stochasticity. By incorporating suitable modifications and adaptations, we aim to improve Newton-type methods' convergence behavior, robustness, and computational efficiency in the context of semi-smooth stochastic optimization. To achieve this, the thesis will delve into a comprehensive review of the existing literature on semi-smooth optimization, stochastic optimization, and Newton-type methods. The fundamental concepts, mathematical foundations, and theoretical fundamentals of these areas are explored establishing a solid framework for the subsequent research. Further-

more, we will analyze the challenges posed by semi-smooth stochastic optimization problems and identify the limitations of existing algorithms when applied to such scenarios. This analysis will serve as the basis for proposing innovative modifications to Newton-type methods to effectively address these challenges and improve their performance.

As the accuracy of the SAA approximate objective functions and its derivatives is naturally proportional to the computational costs – higher precision implies larger costs in general, it is important to design an efficient balance between accuracy and costs. Therefore, the main focus will be on the development of adaptive sample size control algorithms for solving semi-smooth stochastic optimization problems, with particular attention given to the control of the accuracy and selection of search directions. Several options will be investigated for the search direction, while the accuracy control will involve cheaper objective function approximations (with looser accuracy) during the initial stages of the process to save computational effort. This approach aims to conserve computational resources, reserving the deployment of high-accuracy objective function approximations for the final stages of the optimization process.

Outline of thesis

The remaining part of the thesis is structured as follows:

Chapter 1 contains notation and a brief overview of definitions and theorems essential for better comprehension and follow-up of the subsequent analysis of original results.

In **Chapter 2** the relevant concepts of smooth nonlinear optimization are summarized. The emphasis is on line search methods, and various strategies for choosing the search direction and step size are discussed.

Chapter 3 comprises an overview of the basic aspects of semi-smooth optimization methods.

In **Chapter 4** the framework for the main subject of this thesis, the minimization of the expected value function, is introduced. The

chapter begins by introducing the semi-smooth optimization problem and subsequently presents the essential assumptions required for the algorithms employed in solving the problem at hand.

The last two chapters are the centerpieces and constitute the original contribution of this thesis. They present novel algorithms including their convergence analysis and practical implementation.

In **Chapter 5** the stochastic spectral projected gradient method is adapted to the nonsmooth framework. The spectral step is employed in order to find a suitable direction that improves the performance of the first-order method. This coefficient approximates the average eigenvalue of the Hessian matrix providing at least some rough second-order information which is crucial for fast convergence. Moreover, an adaptive strategy that dynamically determines when to switch to the next level of accuracy is presented.

- N. Krejić, N. Krklec Jerinkić, T. Ostojić, Spectral projected subgradient method for nonsmooth convex optimization problems, *Numerical Algorithms (2022), pp. 1-19.* [56]

- N. Krklec Jerinkić, T. Ostojić, AN-SPS: Adaptive Sample Size Nonmonotone Line Search Spectral Projected Subgradient Method for Convex Constrained Optimization Problems, *arXiv preprint arXiv:2208.10616, (2022).* [62]

Chapter 6 is based on the Inexact Restoration framework, which was originally developed for constrained optimization problems and has already proven to be a powerful tool for problems with inexact objective functions. The main idea of this method is to consider feasibility and optimality separately and to balance them by a merit function such that eventually one gets a feasible optimal point. The problems with inexact objective functions can be easily transformed into constrained problems with approximate objective functions of different

levels of accuracy and a simple constraint that measures the level of accuracy. On the other hand, the famous BFGS methods from smooth optimization appear very efficient in semi-smooth settings, and recent theoretical developments have shown that they can also be an effective general-purpose tool for semi-smooth optimization. Therefore, the possibilities of combining an approximate objective function with a quasi-Newton method are investigated.

- N. Krejić, N. Krklec Jerinkić, T. Ostojić, An inexact restoration-nonsmooth algorithm with variable accuracy for stochastic nonsmooth convex optimization problems in machine learning and stochastic linear complementarity problems, *Journal of Computational and Applied Mathematics (2023), 423, 114943.* [55]

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Overview of the Background Material

This chapter is devoted to a brief overview of the relevant definitions, basic notation, and theoretical results to facilitate further reading of the thesis. First, we set the notation used in the thesis.

Notation

$\mathbb{N}$ – the set of positive integers;

$\mathbb{R}$ – the set of real numbers;

$\mathbb{R}_+$ – the set of nonnegative real numbers;

$\mathbb{R}^n$ – the space of n-dimensional vectors with real components;

$\mathbb{R}^{n\times m}$ – the space of real-valued matrices with n rows and m columns;

$x \in \mathbb{R}^n$ – column vector, $x = (x_1, x_2, \ldots, x_n)^T$, where $x_1, x_2, \ldots, x_n$ represent its components;

$x \geq 0$ – a vector whose components are nonnegative, i.e., $x_i \geq 0$ for every component x_i; The space of such vectors is denoted with $\mathbb{R}^n_+$;

$x = 0$ – a vector whose components are equal to zero, i.e., $x_i = 0$, $i = 1, \ldots, n$;

$A \in \mathbb{R}^{n \times m}$ – a matrix with n rows, m columns and entries $A_{i,j} \in \mathbb{R}$ ($A_{i,j}$ - the element in the ith row and jth column of the matrix A);

$A = 0$ – every component of the matrix A is zero, i.e., $A_{i,j} = 0$, $i, j = 1, 2, \ldots, n$;

$A^T \in \mathbb{R}^{m \times n}$ – the transpose of matrix $A \in \mathbb{R}^{n \times m}$;

$|A|$ – the determinant of the matrix A;

A^{-1} – the inverse matrix of the matrix A;

ι_{min}, ι_{max} – the smallest and the largest eigenvalue in the absolute value of matrix A, respectively;

$\mathrm{eig}(A) = \{\iota_i\}_{i=1,\ldots,n}$ – the set of eigenvalues of matrix A, where $\iota_1 \geq \ldots \geq \iota_n$;

I – the identity matrix;

$\|x\|$ – the Euclidean norm $\|x\|_2$, i.e., $\|x\|^2 = \sum_{i=1}^n x_i^2$;

$x^T y$ – the scalar product of vectors x and y, i.e., $x^T y = \sum_{i=1}^n x_i y_i$;

$\{x_k\} := \{x_k\}_{k\in\mathbb{N}}$ – the sequence $x_1, x_2, \ldots$;

$int(X)$ – a interior of a set X;

$cl(X)$ – a closure of an open set X;

$conv(X)$ – a convex hull of a set X;

$\mathcal{O}(x)$ – a neighborhood of a point x, i.e., any open subset of $\mathbb{R}^n$ that contains $x \in \mathbb{R}^n$;

$\mathcal{B}(x, r)$ – an open ball with center $x \in \mathbb{R}^n$ and radius $r > 0$, $\mathcal{B}(x, r) := \{y \in \mathbb{R}^n : \|x - y\| < r\}$;

$C(D)$ – the set of functions which are continuous on $D \subseteq \mathbb{R}^n$;

$C^1(D)$ – the set of functions that have continuous first derivatives on D (continuously-differentiable or smooth functions);

$C^k(D)$ – the set of functions that have k continuous derivatives on D, $k \geq 1$;

$\nabla f(x) \in \mathbb{R}^n$ – the gradient of a function $f : \mathbb{R}^n \to \mathbb{R}$ at point $x \in \mathbb{R}^n$;

$\nabla^2 f(x) \in \mathbb{R}^{n\times n}$ – the Hessian of a function $f : \mathbb{R}^n \to \mathbb{R}$ at point $x \in \mathbb{R}^n$;

$\partial f(x) \subseteq \mathbb{R}^n$ – a subdifferential of a function $f : \mathbb{R}^n \to \mathbb{R}$ at point $x \in \mathbb{R}^n$;

$\mathcal{A}$ – the set of all possible outcomes, $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$;

$\mathcal{P}(\mathcal{A})$ – the partitive set of $\mathcal{A}$;

$\bar{A}$ – the complementary set of the set A, $\bar{A} = \mathcal{A} \setminus A$;

$\varnothing$ – the empty set.

1.1 Linear Algebra and Functional Analysis

Definition 1.1.1 *The set X is open if $X = int\, X$. The set X is closed if $X = cl\, X$.*

Definition 1.1.2 *The set X is bounded if there exists a positive constant M such that for every $x \in X$ there holds $\|x\| \leq M$.*

Definition 1.1.3 *The set $X \subseteq \mathbb{R}^n$ is a compact set if it is closed and bounded.*

Definition 1.1.4 *The matrix $A \in \mathbb{R}^{n\times n}$ is symmetric if $A = A^T$.*

Definition 1.1.5 *The matrix $A \in \mathbb{R}^{n\times n}$ is positive semidefinite if for every $x \in \mathbb{R}^n$ we have that $x^T Ax \geq 0$. The matrix $A \in \mathbb{R}^{n\times n}$ is positive definite if for every $x \in \mathbb{R}^n$, $x \neq 0$ the inequality is strict, that is $x^T Ax > 0$.*

If the matrix A is symmetric positive definite, then the inverse matrix of A, A^{-1}, is also positive definite.
For $m, M \in \mathbb{R}$ the notation $mI \preceq A \preceq MI$ is used to indicate the fact that for every $\iota \in \mathrm{eig}(A)$ it holds $m \leq \iota \leq M$.

Inequality

$$||x + y|| \leq ||x|| + ||y|| \text{ for every } x, y \in \mathbb{R}^n$$

is called triangular inequality, while the following represented reversed triangular inequality, holds for every norm and every x, y :

$$||x - y|| \geq \Big|\,||x|| - ||y||\,\Big|.$$

The Cauchy–Schwarz inequality states that for all vectors $x, y \in \mathbb{R}^n$ the following inequality holds

$$|x^T y| \leq ||x||||y||.$$

Definition 1.1.6 *The sequence* $\{x_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ *is*

i) bounded if there exists $M \in \mathbb{R}$ *such that* $||x_k|| \leq M$ *for every* $k \in \mathbb{N}$*;*

ii) Cauchy if for every $\varepsilon > 0$ *there exists* $\bar{k} \in \mathbb{N}$ *such that* $||x_s - x_l|| \leq \varepsilon$ *for every* $s, l \geq \bar{k}$*;*

iii) convergent if there exists x^* *such that* $\lim_{k\to\infty} x_k = x^*$*. That is, if for every* $\varepsilon > 0$ *there exists* $\bar{k} \in \mathbb{N}$ *such that* $||x_k - x^*|| \leq \varepsilon$ *for every* $k \geq \bar{k}$*.*

Lemma 1.1.1 *For a sequence* $\{x_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ *we have that*

i) $\{x_k\}$ *is convergent if and only if it is a Cauchy sequence;*

ii) if $\{x_k\}$ *converges to* $x^* \in \mathbb{R}$ *then all its subsequences also converge to* x^**.*

Definition 1.1.7 *For a sequence* $\{x_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ *and a point* $\tilde{x} \in \mathbb{R}$*, we say that* $\tilde{x}$ *is an accumulation point of* $\{x_k\}$ *if for every open subset* $X \subseteq \mathbb{R}^n$ *such that* $\tilde{x} \in X$*, we have that* $x_k \in X$ *for infinitely many values of* $k \in \mathbb{N}$*.*

Definition 1.1.8 *For a sequence* $\{x_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ *and a point* $\tilde{x} \in \mathbb{R}$*, we say that* $\tilde{x}$ *is a strictly strong accumulation point if there exists a subsequence* $K \subseteq \mathbb{N}$ *and a constant* $b \in \mathbb{N}$ *such that* $\lim_{k_i\in K} x_{k_i} = \tilde{x}$ *and* $k_{i+1} - k_i \leq b$ *for any two consecutive elements* $k_i, k_{i+1} \in K$*.*

Lemma 1.1.2 *For a sequence* $\{x_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ *we have that if* $\{x_k\}$ *is bounded then it has at least one accumulation point. Moreover, if* $\tilde{x}$ *is an accumulation point of* $\{x_k\}$*, then there exists a subsequence of* $\{x_k\}$ *that converges to* $\tilde{x}$*.*

Definition 1.1.9 *Suppose that the sequence* $\{x_k\}_{k\in\mathbb{N}}$ *converges to* x^*. *The convergence is Q-linear if there is a constant* $\rho \in (0,1)$ *such that for all* k *sufficiently large*

$$\|x_{k+1} - x^*\| \leq \rho \|x_k - x^*\|.$$

The convergence is Q-superlinear if

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0.$$

The convergence is Q-quadratic if there exists a positive constant M *such that for all* k *sufficiently large*

$$\|x_{k+1} - x^*\| \leq M \|x_k - x^*\|^2.$$

The convergence is R-linear if for all k *sufficiently large*

$$\|x_k - x^*\| \leq a_k,$$

where $\{a_k\}_{k\in\mathbb{N}}$ *is a sequence which converges to zero Q-linearly.*

In this thesis, we will generally consider nonsmooth real-valued functions, more precisely the functions that are not necessarily differentiable. But first, let us give a basic definition in the case of smooth function.

Definition 1.1.10 *The function* $f : \mathbb{R}^n \to \mathbb{R}$ *is continuously differentiable if for every* $i = 1, \ldots, n$ *the partial derivative* $\partial_{x_i} f$ *exists and it is continuous everywhere in* $\mathbb{R}^n$. *Analogously, the function* f *is twice continuously differentiable, if* $\partial_{x_i}\partial_{x_j} f$ *exists and it is continuous everywhere in* $\mathbb{R}^n$ *for* $i, j = 1, \ldots, n$.

For a twice continuously-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we can define gradient $\nabla f(x) \in \mathbb{R}^n$ and Hessian $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$ of function f in the following way

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

and

$$\left(\nabla^2 f(x) \right)_{i,j} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i, j = 1, ..., n.$$

1.2 Convex Analysis

As many optimization problems arising in various applications require the minimization of an objective cost function that is convex but not differentiable, the main concern of this thesis is nonsmooth convex optimization methods. The development of nonsmooth optimization methods has begun with the analysis of convex functions. It has been shown that convex optimization algorithms are efficient for computing reliable solutions in a broad range of applications, as one of the most important features of convex functions is that every local minimizer is also a global minimizer of that function. In the following, we provide the basic concepts of convex analysis, where one can distinguish two types of convexity. The first one is the convexity of a function and the other one convexity of a set. This section mostly relies on [22].

Definition 1.2.1 *The set $C \subseteq \mathbb{R}^n$ is convex if for all real numbers $\lambda \in [0,1]$ it holds*

$$x, y \in C \Rightarrow \lambda x + (1-\lambda)y \in C.$$

Definition 1.2.2 *The convex combination of vectors $x_1, x_2, \ldots, x_n$ is given by*

$$\sum_{i=1}^{k} \lambda_i x_i,$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are nonnegative real numbers such that $\sum_{i=1}^{k} \lambda_i = 1$.

Definition 1.2.3 *The convex hull of a set $C \subseteq \mathbb{R}^n$ is*

$$conv(C) = \{x \in \mathbb{R}^n | x = \sum_{i=1}^{k} \lambda_i x_i, \sum_{i=1}^{k} \lambda_i = 1, x_i \in C,$$
$$\lambda_i \geq 0, i = 1, \ldots, k, k > 0\}.$$

Definition 1.2.4 *Function $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$ is convex on a convex set D if for every $x, y \in D$ and every $\lambda \in [0, 1]$*

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

Definition 1.2.5 *Function $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$ is strictly convex on a convex set D if for every $x, y \in D$ and every $\lambda \in (0, 1)$*

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y).$$

Definition 1.2.6 *Function $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$ is strongly convex with parameter $\mu > 0$ on a convex set D if for any $x, y \in D$ and any $\lambda \in [0, 1]$ there holds*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \mu\frac{1}{2}\lambda(1 - \lambda)||x - y||^2.$$

In addition, if the function f is differentiable, then we can state another characterization of convex and strongly convex functions.

Theorem 1.2.1 *Assume that $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$ and $f \in C^1(D)$. Then the function f is convex if and only if for every $x, y \in D$*

$$f(x) \geq f(y) + \nabla f(y)^T(x - y).$$

Furthermore, the function is $\mu-$strongly convex if and only if there exists a positive constant μ such that for every $x, y \in D$

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}||y - x||^2.$$

Theorem 1.2.2 *Assume that f is twice continuously differentiable on a convex and open set D. Then it holds that f is convex on D if and only if the Hessian of the function f is positive semidefinite on D, i.e., $\nabla^2 f(x) \succeq 0$ for all $x \in D$.*

Definition 1.2.7 *Let $C \subset \mathbb{R}^n$ be a closed convex set. The orthogonal projection map, denoted by $P_C : \mathbb{R}^n \to C$, is defined as follows*

$$P_C(y) := \arg\min\{||y - z|| : z \in C\}.$$

The next proposition presents important non-expansive properties of the projection.

Proposition 1.2.1 *Let $C \subset \mathbb{R}^n$ be a closed convex set, $x, y \in \mathbb{R}^n$ and $z \in C$. Then, we have*

i) $||P_C(y) - z||^2 \leq ||y - z||^2$;

ii) $||P_C(x) - P_C(y)||^2 \leq ||x - y||^2$.

Lipschitz Continuous Functions

Lipschitz continuous functions play a significant role in convex and nonsmooth analysis and therefore we give the following definition.

Definition 1.2.8 *A function $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$ is called locally Lipschitz continuous over D if for all $x_0 \in D$ there exists $\varepsilon > 0$ and a Lipschitz constant $L(x_0) \in \mathbb{R}_+$ such that*

$$||f(x_0) - f(x)|| \leq L(x_0)||x_0 - x||,$$

for all $x \in D \cap \mathcal{B}(x_0, \varepsilon)$.

If there exists $L \in \mathbb{R}_+$ such that for every $x, y \in D$

$$||f(x) - f(y)|| \leq L||x - y|| \tag{1.1}$$

holds, than we call f globally Lipschitz continuous on D.

In other words, we say that a function $f : D \to \mathbb{R}$ is locally Lipschitz if every point in D has a neighborhood on which f is Lipschitz continuous. It can be shown that every convex function is locally Lipschitz continuous. Furthermore, Rademacher's theorem states that Lipschitz continuous function f is differentiable almost everywhere, i.e., the set of points where f is not differentiable is of measure zero. In particular, every neighborhood of x contains a point y for which $\nabla f(y)$ exists. If the function is continuously differentiable then it is locally Lipschitz continuous. We state theorems for the sake of completeness.

Proposition 1.2.2 *Any convex function $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$ is locally Lipschitz continuous.*

Theorem 1.2.3 *(Rademacher) Let D be an open subset of $\mathbb{R}^n$, $f : D \to \mathbb{R}^m$ a L-Lipschitz function, i.e., (1.1) holds, and $\tilde{\mu}$ the Lebesgue measure. Then f is differentiable almost everywhere. That is, there is a set $E \subset D$ with $\tilde{\mu}(D \setminus E) = 0$ such that for every $x \in E$ there is a linear function $L_x : \mathbb{R}^n \to \mathbb{R}^m$ with*

$$\lim_{y \to x} \frac{f(x) - f(y) - L_x(y - x)}{\|y - x\|} = 0.$$

Lemma 1.2.1 *Assume that the function f is given by $f(x) = \sum_{i=1}^{N} f_i(x)$, with $f_i : \mathbb{R}^n \to \mathbb{R}$. If for every $i = 1, \ldots, N$ the function f_i is L_i-Lipschitz continuous with $L_i \geq 0$, then f is Lipschitz continuous with constant $L = \sum_{i=1}^{N} L_i$.*

Lemma 1.2.2 *If $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable function, the following properties hold*

i) the Hessian matrix $\nabla^2 f(x)$ is symmetric;

ii) if ∇f is L-Lipschitz continuous, then for every $x \in \mathbb{R}^n$ we have that $\iota_{max}\left(\nabla^2 f(x)\right) \leq L$;

iii) f is μ-strongly convex if and only if $\iota_{min}\left(\nabla^2 f(x)\right) \geq \mu$.

1.3 Probability Theory

In this section, we will give an overview of some definitions and results from probability theory [32, 86]. Recall that we deal only with real-valued random variables.

A probability space is a triple $(\mathcal{A}, \mathcal{F}, P)$, where $\mathcal{A}$ is a set of outcomes, $\mathcal{F}$ is a set of events, and P is a function that assigns probabilities to events. We say that the event happens almost surely (a.s.) if it happens with probability 1. Next, the definition of independent events is stated.

Definition 1.3.1 *The sequence of events $A_1, A_2, \ldots$ from $\mathcal{F}$ is independent if for every finite sequence of indices $k_1 < \ldots < k_s$ the following equality holds*

$$P(A_{k_1} \cap \ldots \cap A_{k_s}) = P(A_{k_1}) \cdots P(A_{k_s}).$$

In order to define random variables, first we need to define Borel's σ-field.

Definition 1.3.2 *Borel's σ-field $\mathcal{B}$ in topological space $(\mathbb{R}, \tau)$ is the smallest σ-field that contains τ.*

Next, we provide the definition of the random variable.

Definition 1.3.3 *Mapping $X : \mathcal{A} \to \mathbb{R}$ is a random variable on the space $(\mathcal{A}, \mathcal{F}, P)$ if*

$$X^{-1}(S) \in \mathcal{F} \text{ for every } S \in \mathcal{B}.$$

Definition 1.3.4 *The cumulative distribution function for the random variable X, $F_X : \mathbb{R} \to [0, 1]$, is defined by*

$$F_X(x) = P\{X \leq x\}.$$

One can distinguish two types of random variables - discrete and continuous.

Definition 1.3.5 *The random variable X is discrete if there exists a countable set S such that $P(X \in S) = 1$.*

Definition 1.3.6 *The random variable X is continuous if there exists a nonnegative function φ_X such that for every $S \in \mathcal{B}$*

$$P(X \in S) = \int_S \varphi_X(x)dx.$$

The function $\varphi_X(\cdot)$ is called the probability density function.

Definition 1.3.7 *Random variables $X_1, X_2, \ldots$ are independent if the events $X_1^{-1}(S_1), X_2^{-1}(S_2), \ldots$ are independent for all $S_i \in \mathcal{B}, i = 1, 2 \ldots$.*

Now, the numerical characteristics of random variables, more precisely the definition of mathematical expectation and variance, will be introduced.

Definition 1.3.8 *If X is a discrete random variable, then the mathematical expectation $E[X]$ exists if and only if*

$$\sum_{k=1}^{\infty} |x_k| P(X = x_k) < \infty,$$

where $x_1, x_2, \ldots$ are the values that X may take and it is given by

$$E[X] = \sum_{k=1}^{\infty} x_k P(X = x_k).$$

If X is absolutely continuous, the mathematical expectation exists if

$$\int_{-\infty}^{\infty} |x|\varphi_X(x)dx < \infty$$

and it is defined by

$$E[X] = \int_{-\infty}^{\infty} x\varphi_X(x)dx.$$

In the following, some characteristics of the mathematical expectation are stated and their significance is pointed out.

Theorem 1.3.1 *Let $X_1, X_2, \ldots, X_n$ be random variables that poses the mathematical expectations and $c \in \mathbb{R}$. Then the following holds*

i) $|E[X_k]| \leq E[|X_k|]$;

ii) $E[c] = c$;

iii) $E[cX_k] = cE[X_k]$;

iv) $E[X_1 + \ldots + X_n] = E[X_1] + \ldots + E[X_n]$;

v) If $X_k \geq 0$ almost surely, then $E[X_k] \geq 0$;

vi) If $X_1, X_2, \ldots, X_n$ are independent, then

$$E\left[\prod_{k=1}^{n} X_k\right] = \prod_{k=1}^{n} E[X_k];$$

vii) If $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is a random vector, then

$$E[\mathbf{X}] = (E[X_1], \ldots, E[X_n]).$$

Definition 1.3.9 *The function $f : \mathbb{R}^n \to \mathbb{R}^m$ is a Borel's function if for every S from Borel's σ-field $\mathcal{B}(\mathbb{R}^m)$ the inverse $f^{-1}(S)$ belongs to Borel's σ-field $\mathcal{B}(\mathbb{R}^n)$.*

Theorem 1.3.2 *Let $f : \mathbb{R} \to \mathbb{R}$ be a Borel's function. Then, if X is discrete the mathematical expectation of $f(X)$ is*

$$E\left[f(X)\right] = \sum_{k=1}^{\infty} f(x_k) P(X = x_k)$$

and if X is continuous

$$E\left[f(X)\right] = \int_{-\infty}^{\infty} f(x)\varphi_X(x)dx.$$

Definition 1.3.10 *The random variable X is P-integrable if $E\left[X\right]$ is well defined and finite.*

Now, we will define the variance, which represents a measure of spread for the distribution of a random variable that determines the degree to which the values of a random variable differ from the expected value. It is denoted with $D\left[X\right]$ or $Var^2\left[X\right]$. But, first, we need to define the moments and the central moments.

Definition 1.3.11 *Let X be a random variable and $k \in \mathbb{N}$. Then the moment of order k of X is given by $E\left[X^k\right]$, while the central moment of order k is*

$$E\left[\left(X - E\left[X\right]\right)^k\right].$$

Definition 1.3.12 *The variance of a random variable X is the second-order central moment of that random variable, i.e.,*

$$D\left[X\right] = E\left[\left(X - E\left[X\right]\right)^2\right].$$

The following formula, which can easily be obtained from the previous definition, is often used for the calculation of variance

$$D[X] = E\left[X^2\right] - E^2[X].$$

Theorem 1.3.3 *Let $X_1, X_2, \ldots, X_n$ be random variables with the variances $D[X_1], D[X_2], \ldots, D[X_n]$ and $c \in \mathbb{R}$. Then the following holds*

i) $D[X_k] \geq 0$;

ii) $D[X_k] = 0$ *if and only if* X_k *is a constant almost surely;*

iii) $D[cX_k] = c^2 D[X_k]$;

iv) $D[X_k + c] = D[X_k]$;

v) If $X_1, X_2, \ldots, X_n$ *are independent, then*

$$D\left[\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n} D[X_k];$$

vi) If $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ *is a random vector, then*

$$D[\mathbf{X}] = (D[X_1], \ldots, D[X_n]).$$

As the behavior of sequences of random variables is one of the most important parts of probability theory, we define four basic types of convergence concerning random variables and discuss how they are related.

Definition 1.3.13 *A sequence of random variables $X_1, X_2, \ldots$ converges in probability towards random variable X if for every $\varepsilon > 0$*

$$\lim_{k\to\infty} P(|X_k - X| \geq \varepsilon) = 0.$$

The basic idea behind this type of convergence is that the probability of an "unusual" outcome becomes smaller and smaller as the sequence progresses.

Definition 1.3.14 *A sequence of random variables $X_1, X_2, \ldots$ converges almost surely (a.s.) towards random variable X if*

$$P(\lim_{k\to\infty} X_k = X) = 1.$$

This is the type of stochastic convergence that is most similar to pointwise convergence.

Definition 1.3.15 *A sequence of random variables $X_1, X_2, \ldots$ converges in mean square towards random variable X if the following conditions hold*

i) $E[X_k^2] < \infty$ *for every* $k \in \mathbb{N}$;

ii) $\lim_{k\to\infty} E[(X_k - X)^2] = 0$.

Definition 1.3.16 *A sequence of random variables $X_1, X_2, \ldots$ converges in distribution towards random variable X if, for every $x \in \mathbb{R} \cup \{-\infty, \infty\}$ such that $F_X(x)$ is continuous, the following holds*

$$\lim_{k\to\infty} F_{X_k}(x) = F_X(x).$$

The relationships between the mentioned convergences are formulated in the following theorems.

Theorem 1.3.4 *If a sequence of random variables $X_1, X_2, \ldots$ converges in mean square towards random variable X, then it converges in probability.*

Theorem 1.3.5 *If a sequence of random variables $X_1, X_2, \ldots$ converges almost surely towards random variable X, then it converges in probability.*

Theorem 1.3.6 *If a sequence of random variables $X_1, X_2, \ldots$ converges in probability towards random variable X, then it converges in distribution. Moreover, if a sequence of random variables converges to a constant, then convergence in distribution implies convergence in probability.*

Convergence in distribution is the weakest form of mentioned convergence since it is implied by all other types of convergence.

Finally, at the end of this section, we state two fundamental laws that deal with limiting behavior of the sequence of independent random variables throughout the Weak Law of Large Numbers and Strong Law of Large Numbers. Convergence in probability is stated in the first law, while the other one considers almost sure convergence.

Let S_n be the sum of n random variables, i.e.,

$$S_n = X_1 + X_2 + \ldots + X_n.$$

Theorem 1.3.7 *(Weak Law of Large Numbers) Let $X_1, X_2, \ldots$ be independent random variables. If there exists a constant C such that $D[X_i] \leq C$ for every $i \in \mathbb{N}$, then*

$$\lim_{n\to\infty} \frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} E[X_i] \quad \textit{in probability.}$$

Theorem 1.3.8 *(Strong Law of Large Numbers) Let $X_1, X_2, \ldots$ be independent random variables. If the random variables $X_1, X_2, \ldots$ have the same distribution and the finite mathematical expectation $E[X_k] = a$, then*

$$\lim_{n\to\infty} \frac{S_n}{n} = a \ \ a.s.$$

Chapter 2

Nonlinear Optimization

In this chapter, a short summary of smooth nonlinear optimization is given. The fundamentals of deterministic optimization are presented - optimization problem and optimality conditions are introduced, and a general framework of line search method as core globalization strategy is presented. Due to the nature of considered nonsmooth problems in this thesis, where the search direction is not necessarily descent, the special class within the line search framework, the nonmonotone line search method, is presented at the end of this chapter. This chapter mostly relies on [61, 82].

2.1 Problem Statement and Optimality Conditions

The problem that we consider is given by

$$\min_{x \in \Omega} f(x), \tag{2.1}$$

where function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable function and set $\Omega \subseteq \mathbb{R}^n$ is called the feasible set and it can be represented in the

form

$$\Omega = \{x \in \mathbb{R}^n | \quad g_i(x) \leq 0, i = 1, \ldots, s, \quad h_i(x) = 0, \quad i = 1, \ldots, m\},$$

where $g_1, \ldots, g_s, h_1, \ldots, h_m$ are real-valued functions that represent inequality and equality constraints.

It is assumed that $f(x)$ is nonlinear and bounded from below on Ω. In that case, the optimal value $f^* := \inf_{x \in \Omega} f(x)$ is finite and our goal is to find $x^* \in X^*$ with X^* defined as

$$X^* := \{x \in \Omega | f(x) = f^*\}.$$

Now, we state the important result that gives the conditions for the existence of x^*.

Theorem 2.1.1 *(Weierstrass) If $\Omega \subseteq \mathbb{R}^n$ is non-empty and compact and $f : \Omega \to \mathbb{R}$ is continuous, then there exists a global minimizer of the considered optimization problem.*

In the rest of this chapter, we will focus on the special case of problem (2.1) - the unconstrained optimization problem. That is, we assume that $\Omega = \mathbb{R}^n$ and therefore we want to find an optimal solution x^* such that

$$f(x^*) = \min_{x \in \mathbb{R}^n} f(x). \tag{2.2}$$

Let us formally give the definition of the optimal point.

Definition 2.1.1 *(Global minimizer) The point $x^* \in \mathbb{R}^n$ is a global minimizer of $f : \mathbb{R}^n \to \mathbb{R}$ if $f(x^*) \leq f(x)$ for all $x \in \mathbb{R}^n$. If this inequality is strict, then x^* is a strict global minimizer.*

In general, it is very hard to recognize a global minimizer. Thus, local minimizers are considered as an alternative.

Definition 2.1.2 *(Local minimizer) The point $x^* \in \mathbb{R}^n$ is a local minimizer of $f : \mathbb{R}^n \to \mathbb{R}$ if there is an open neighborhood $\mathcal{O}$ of x^* such that $f(x^*) \leq f(x)$ for all $x \in \mathcal{O}$. If this inequality is strict, then x^* is a strict local minimizer.*

Now, we are stating necessary and sufficient conditions for a solution of problem (2.2).

Theorem 2.1.1 *(First-order necessary conditions) If x^* is a local minimizer and $f(x)$ is continuously differentiable in an open neighbourhood $\mathcal{O}$ of x^*, then $\nabla f(x^*) = 0$.*

The points which satisfy the first-order necessary conditions are called stationary points. It is important to note that $\nabla f(x^*) = 0$ does not necessarily mean that x^* is a local minimizer. From the previous theorem, we can only claim that any local minimizer must be a stationary point. As the first derivatives do not provide enough information to detect whether the point is a minimizer, the second derivatives are needed.

Theorem 2.1.2 *(Second-order necessary conditions) If x^* is a local minimizer of $f(x)$ and $\nabla^2 f(x)$ exists and is continuous in an open neighborhood $\mathcal{O}$ of x^*, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite matrix.*

Finally, let us state the second-order sufficient conditions, which make guarantees that x^* is a local minimizer.

Theorem 2.1.3 *(Second-order sufficient conditions) Suppose that $\nabla^2 f(x)$ is continuous in an open neighbourhood $\mathcal{O}$ of x^*, $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite. Then, x^* is a strict local minimizer of f.*

We say that the unconstrained optimization problem is convex if the objective function is convex. In the following, we state the important fact regarding convex problems.

Theorem 2.1.2 *Suppose that function f is convex on a convex set S. Then, every local minimizer of the function f is also a global minimizer.*

The vast majority of optimization problems are difficult to solve directly. Thus, numerical algorithms for optimization are used to find an approximation of the optimal solution (2.2). These algorithms are called iterative methods, because they start from some initial point $x_0 \in \mathbb{R}^n$ and according to a certain iterative rule form a sequence $\{x_k\}_{k\in\mathbb{N}}$ recursively, where the elements of this sequence represent estimates of the optimal solution and are called iterations. The goal of numerical algorithms is to construct a sequence of iterates $\{x_k\}_{k\in\mathbb{N}}$ that converges to a solution x^* of the considered problem.

Let s_k be the step from the current iteration x_k to a new iteration x_{k+1}, i.e.,

$$x_{k+1} = x_k + s_k.$$

Regarding the choice of s_k, there are two fundamental strategies that guarantee convergence towards a local minimum: the line search and the trust region, where these two approaches differ in the way they calculate s_k, more precisely in the order in which they choose the direction and magnitude of that direction.

In the line search method, we first choose a search direction p_k from the current point x_k and then compute the step size α_k such that $s_k = \alpha_k p_k$. Opposite to line search methods, at an arbitrary iteration, trust region methods determine the step size bound first (i.e., the trust region radius), and then solve a sub-problem to find a suitable direction.

In the thesis, we consider only line search methods, so in the following section basic notations and review of significant results are presented. More about trust region methods can be found in [23, 82].

2.2 Line Search Methods

Let us formally state the rule which gives us the following iteration in the line search method

$$x_{k+1} = x_k + \alpha_k p_k, \tag{2.3}$$

where $\alpha_k > 0$ is called the step size and p_k is the search direction. The main idea of this method is to obtain a new iteration x_{k+1} with a lower function value. More precisely, after choosing a direction p_k we should determine the optimal scaling of p_k such that

$$f(x_k + \alpha_k p_k) < f(x_k).$$

Therefore, we want to solve the following problem exactly or approximately

$$\alpha_k = \arg\min_{\alpha>0} f(x_k + \alpha p_k). \tag{2.4}$$

So, the question is how to choose the search direction p_k and how to solve the problem (2.4) in order to get a suitable step size α_k. The model algorithm of the line search method, Algorithm 1, is stated in the following.

As the different choices of p_k and α_k seriously affect the efficiency of the iterative method, the most frequently used search directions, as well as a few strategies for the choice of the step size α_k will be outlined in further sections. Notice that Step S1 of Algorithm 1 is written in the most general form without imposing any condition on the decrease of the objective function, so that monotone and nonmonotone

Algorithm 1: LS (**L**ine **S**earch)

S0 *Initialization.* Specify an initial point $x_0 \in \mathbb{R}^n$.
Set $k = 0$.

S1 *Search direction.* Choose the search direction p_k.

S2 *Step size.* Find $\alpha_k > 0$ as a (approximate) solution of (2.4).

S3 *Update.* Set $x_{k+1} = x_k + \alpha_k p_k$.

S4 **If** some termination criterion is satisfied, then stop.
Else $k = k + 1$ and go to **S1**.

line search strategies can fit into that framework. More precisely, in Step S1, the necessary condition which search direction should satisfy is not specified. In Section 2.2.1 we will present what kind of a search direction p_k is desirable in the case of the monotone line search method, while in Section 2.3 the nonmonotone line search method will be analyzed. Also, in Section 2.2.2 we will present some practical strategies for choosing a suitable step size α_k for a given direction p_k such that adequate reduction in function value is achieved.

2.2.1 Search Directions

Most line search methods require p_k to be a descent direction in order to get improvement. Thus, we give a formal definition.

Definition 2.2.1 *For a given point $x_k \in \mathbb{R}^n$ a direction $p_k \in \mathbb{R}^n$ is called a descent direction, if there exists $\overline{\alpha}$ such that*

$$f(x + \alpha p) < f(x), \ \forall \alpha \in (0, \overline{\alpha}).$$

If the function f is continuously-differentiable it can be shown that p_k is a descent direction from point x_k if the following condition is satisfied

$$\nabla f(x_k)^T p_k < 0. \tag{2.5}$$

This property guarantees that the function f can be reduced along this direction p_k, is a direct consequence of the first-order Taylor expansion of continuously differentiable function f around x_k

$$f(x_k + \alpha p_k) = f(x_k) + \alpha p_k^T \nabla f(x_k) + O(\alpha^2).$$

It is obvious that as long as the step size α is chosen to be small enough it is ensured that

$$f(x_k + \alpha p_k) < f(x_k),$$

which is the primary goal in optimization methods - to obtain a point that is better than the current one at every iteration.

Now, for continuously differentiable function f we can state a characterization of the descent search directions which are frequently used.

Newton-type Methods

The method that uses the search direction of the form

$$p_k = -B_k^{-1} \nabla f(x_k),$$

with a symmetric and nonsingular matrix B_k is called a *Newton-type method* for optimization. The following result holds.

Lemma 2.2.1 *(Descent direction): If matrix B_k is positive definite, i.e., $B_k \succ 0$, then $p_k = -B_k^{-1} \nabla f(x_k)$ is a descent direction.*

Notice that previous lemma holds because for the direction $p_k = -B_k^{-1} \nabla f(x_k)$ we have

$$\nabla f(x_k)^T p_k = -\nabla f(x_k)^T B_k^{-1} \nabla f(x_k) < 0,$$

i.e., the condition (2.5) is satisfied.

Gradient Descent direction is the simplest and it is obtained by choosing $B_k = I$, i.e., the descent direction is defined as $p_k = -\nabla f(x_k)$. This method is also called the Steepest Descent method because along this direction the objective function decreases most rapidly. If we assume that step size is fixed at each iteration, the following result holds.

Theorem 2.2.1 *Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable convex function and that ∇f is $L-$Lipschitz continuous. Let $\{x_k\}$ be the sequence generated by* (2.3) *with search direction $p_k = -\nabla f(x_k)$ and step size $\alpha_k = \alpha$ for every k. If $\alpha \leq 1/L$, then x_k converges linearly to a solution of the problem* (2.2)*.*

The Gradient Descent algorithm with fixed step size is very applicable and widely used due to its simplicity and low cost since it only requires first-order derivatives and no additional computation for the choice of the step size. However, the main drawback of this method is the convergence rate which is at most linear. Accordingly, it might be very slow and it may require many iterations to find a solution with good accuracy.

Let us assume that the function is twice continuously differentiable. Then, we can obtain more sophisticated choices of the descent direction p_k. The objective function $f \in C^2(\mathbb{R}^n)$ can be approximated around the current iteration x_k using the second-order Taylor expansion such that

$$f(x_k + p) \approx f(x_k) + p^T \nabla f(x_k) + \frac{1}{2} p^T \nabla^2 f(x_k) p. \tag{2.6}$$

If we denote the right-hand side of (2.6) with $m_k(p)$, then the goal is to compute the direction p at iteration k by minimizing the quadratic function m_k. If it is assumed that $\nabla^2 f(x_k) \succ 0$, then the function m_k

has the unique minimizer

$$p_k = -\left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k).$$

In other words, for $B_k = \nabla^2 f(x_k)$ we get another important method - Newton's method. This method can be seen as the opposite of the Gradient Descent method, as it is expensive and fast. It achieves local quadratic convergence under suitable regularity assumptions on the function f (see Theorem 2.2.2). However, it can be too computationally expensive since it requires the computation of the second derivatives at every iteration. Now, we state the main convergence result.

Theorem 2.2.2 *Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously-differentiable and that the Hessian $\nabla^2 f(x)$ is Lipschitz continuous in a neighbourhood of a solution x^* at which the sufficient optimality conditions are satisfied. Consider the sequence of iterates generated by the pure Newton's algorithm*[1]*, $\{x_k\}$. Then*

i) if the starting point x_0 is sufficiently close to x^*, the sequence of iterates $\{x_k\}$ converges to x^*;

ii) if the method converges, the rate of convergence of $\{x_k\}$ is quadratic;

iii) if the method converges, the sequence of gradient norms $||\nabla f(x_k)||$ converges to zero quadratically.

As mentioned before, in order to be sure that p_k is a descent direction, the condition of Lemma 2.2.1 should be satisfied. More precisely,

[1]The term "pure" refers that no notion of a step size α is involved, i.e., in (2.3) we have $\alpha = 1$.

the matrix $\nabla^2 f(x_k)$ has to be positive definite at each iteration. It is important to point out that every iteration of Newton's method requires the computation of the Hessian in order to solve the linear system of equations, which might be expensive itself. The method could also become unstable even if $\nabla^2 f(x_k) \succ 0$ in the case of ill-conditioned $\nabla^2 f(x_k)$ at some iteration.

To avoid the shortcomings of Gradient Descent and Newton's method, algorithms that imitate Newton's idea have been proposed. In these methods, referred to as Quasi-Newton (QN) methods, the Hessian matrix is replaced by a matrix $B_k \in \mathbb{R}^{n \times n}$, such that

$$B_k \approx \nabla^2 f(x_k)$$

is a good approximation of the true Hessian and has lower evaluation and linear algebra costs since the approximation is based only on the first-order information. The rate of convergence in QN methods is no more than superlinear, thus the convergence is slower with respect to the Newton method, but on the other hand, the cost is significantly smaller. The idea behind the QN methods is that two successive iterations x_k and x_{k+1} together with the gradients $\nabla f_k := \nabla f(x_k)$ and $\nabla f_{k+1} := \nabla f(x_{k+1})$ contain curvature (i.e., Hessian) information.

Several strategies for computation of the matrix B_k have been proposed in the literature based on the conditions that B_{k+1} should satisfy. The main condition is known as the secant equation

$$B_{k+1} s_k = y_k, \tag{2.7}$$

where the difference between two iterations and the discrepancy between the gradients in two neighboring iterations is given by

$$s_k = x_{k+1} - x_k \text{ and } y_k = \nabla f_{k+1} - \nabla f_k.$$

However, a unique solution for B_{k+1} is not provided from the condition (2.7). Thus, the additional requirements on B_{k+1} are imposed,

such as symmetry and a restriction that the difference between successive approximation B_k to B_{k+1} has a low rank. That is, B_{k+1} is a solution of the following problem

$$\min ||B - B_k||_* \tag{2.8}$$
$$\text{s.t. } B^T = B, Bs_k = y_k.$$

Different updating formulas for B_{k+1} are obtained by solving problem (2.8) depending on the matrix norm $||\cdot||_*$.

One of mostly used updating formulas of this type is proposed by Davidon, Fletcher, and Powell. It is obtained by using the weighted Frobenius norm and it is defined by

$$B_{k+1} = (I - \frac{y_k s_k^T}{y_k^T s_k}) B_k (I - \frac{y_k s_k^T}{y_k^T s_k}) + \frac{y_k y_k^T}{y_k^T s_k}.$$

Another one is the BFGS formula, proposed by Broyden, Fletcher, Goldfarb, and Shanno

$$H_{k+1} = (I - \frac{s_k y_k^T}{y_k^T s_k}) H_k (I - \frac{y_k s_k^T}{y_k^T s_k}) + \frac{s_k s_k^T}{y_k^T s_k},$$

where H_{k+1} represents the inverse Hessian approximation, i.e., $H_k = B_k^{-1}$ ($H_{k+1} y_k = s_k$). The initial approximation H_0 is chosen by the user and it is often defined as $H_0 = \gamma I, \gamma > 0$. We note that the BFGS formula preserves positive definiteness, more precisely, if H_k is positive definite and $y_k^T s_k > 0$ then H_{k+1} is positive definite as well.

Spectral gradient methods

At the end of this section, we introduce a slightly modified version of the QN search direction used later in Chapter 5 - Spectral Gradient (SG) method. This method is well-known for its efficiency and simplicity and has been widely used and developed as a solver of optimization

problems [9, 37, 54, 95]. It was originally proposed by Barzilai and Borwein [5], so it is often referred to as the BB method. The step length selection strategy in the SG method is crucial for faster convergence compared to classical gradient methods, as it incorporates second-order information related to the spectrum of the Hessian matrix. More precisely, it relies on a simple approximation of the second-order derivative (Hessian), which takes the form of an identity matrix multiplied by the so-called spectral coefficient. Roughly speaking, this coefficient approximates the average eigenvalue of the Hessian matrix and provides at least some kind of second-order information, which is crucial for fast convergence. Although the theoretical results are not as strong as for Newton-like methods that rely on true second-order derivatives or better approximations of the Hessian matrix, spectral gradient methods are cheap, easy to implement, and they provide very good numerical results, making them popular in practice.

The spectral coefficient is constructed to best fit the secant equation (2.7), where one of the key ingredients is the difference between the two consecutive gradient values y_k. Thus, we want to find a diagonal matrix of the special form

$$D_k = \lambda_k I, \quad \lambda_k \in \mathbb{R}$$

that best fits the secant equation

$$H_{k+1} y_k = s_k,$$

where matrix D_k represents an approximation of the inverse Hessian $(\nabla^2 f(x_k))^{-1}$. Therefore, the search direction is parallel to the direction of the negative gradient, i.e., we have

$$p_k = -\lambda_k I \nabla f(x_k) = -\lambda_k \nabla f(x_k).$$

The spectral coefficient λ_k is defined by a secant condition, imposing either $\lambda_k = \tilde{\lambda}_k^{-1}$ with

$$\tilde{\lambda}_k = \arg\min_{\lambda \in \mathbb{R}} ||y_{k-1} - \lambda s_{k-1}||^2 \tag{2.9}$$

or

$$\lambda_k = \arg\min_{\lambda \in \mathbb{R}} ||\lambda y_{k-1} - s_{k-1}||^2. \tag{2.10}$$

The following coefficients are obtained from (2.9) and (2.10), respectively:

$$\lambda_k^{BB1} = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}} \tag{2.11}$$

and

$$\lambda_k^{BB2} = \frac{y_{k-1}^T s_{k-1}}{y_{k-1}^T y_{k-1}}. \tag{2.12}$$

In addition to these two rules for calculating λ_k, in the literature, several spectral gradient methods have been proposed which generalize the BB methods. In Chapter 5 we will consider Adaptive Barzilai–Borwein (ABB) [104] and its modification ABBmin [34], which are based on adaptive criteria used to switch between λ_k^{BB1} and λ_k^{BB2}. The step lengths are defined by the following rules:

$$\lambda_k^{ABB} := \begin{cases} \lambda_k^{BB2}, & \frac{\lambda_k^{BB2}}{\lambda_k^{BB1}} < \tau, \\ \lambda_k^{BB1}, & \text{otherwise}, \end{cases} \tag{2.13}$$

and

$$\lambda_k^{ABBmin} := \begin{cases} \min\{\lambda_j^{BB2} : j = \max\{1, k - m_a\}, ..., k\}, & \frac{\lambda_k^{BB2}}{\lambda_k^{BB1}} < \tau, \\ \lambda_k^{BB1}, & \text{otherwise}, \end{cases} \tag{2.14}$$

where m_a is a nonnegative integer and $\tau \in (0, 1)$.

The case when the curvature condition $s_{k-1}^T y_{k-1} > 0$ does not hold leads to the fact that λ_k can be negative, so the search direction is not the descent one. This drawback can be overcome by using the safeguard [96]

$$\bar{\lambda}_k = \min\{\lambda_{max}, \max\{\lambda_k, \lambda_{min}\}\},$$

where $0 < \lambda_{min} << 1 << \lambda_{max} < \infty$. Thus, setting $p_k = -\bar{\lambda}_k \nabla f(x_k)$ it is ensured that the direction is descent and numerical stability can be controlled.

2.2.2 Step Size

In line search methods, after the search direction p_k is chosen at the k-th iteration, the next task is to find a step size α_k along the search direction.

The exact step size is the one that represents the exact solution of the problem (2.4). However, except in certain very special cases, the exact step size is difficult or even impossible to find in practical computation. Therefore, the inexact line search rules considering the value of the function and its derivatives along the direction p_k are constructed such that the global convergence of such methods is ensured. The idea behind these methods is to find an approximate solution of (2.4) which decreases the value of the objective function. More precisely, the goal is to find α_k such that

$$f(x_k + \alpha_k p_k) < f(x_k).$$

The methods where the above condition is satisfied are called monotone line search methods, while in nonmonotone line search methods, this cannot be guaranteed.

Since we want to ensure the decrease of the function, the step sizes should be small enough to get sufficient decrease and on the other hand, long enough to make progress. The sufficient decrease condition - often called the Armijo condition - is imposed in order to ensure the convergence

$$f(x_k + \alpha_k p_k) \leq f(x_k) + \eta \alpha_k (\nabla f(x_k))^T p_k, \tag{2.15}$$

where $\eta \in (0, 1)$ and it is usually set to $\eta = 10^{-4}$. Furthermore, the second imposed condition - curvature condition - prevents the step size

from becoming too small. It is stated as follows

$$(\nabla f(x_k + \alpha_k p_k))^T p_k \geq c\,(\nabla f(x_k))^T p_k, \tag{2.16}$$

where c is some constant that satisfies $0 < \eta < c < 1$. The conditions (2.15) and (2.16) together are called the Wolfe conditions.

The condition (2.16) can be written as $\Phi'(\alpha_k) \geq \Phi'(0)$, where $\Phi(\alpha) = f(x_k + \alpha p_k)$. As there is no guarantee that the step length α_k which satisfies the Wolfe conditions is the local minimum of function Φ, the curvature condition (2.16) could be modified in such a way to force α_k to lie in at least a broad neighborhood of a local minimizer or stationary point of Φ. This can be done by imposing the strong Wolfe conditions, which consist of the Armijo condition and

$$|(\nabla f(x_k + \alpha_k p_k))^T p_k| \leq |c\,(\nabla f(x_k))^T p_k|,$$

instead of (2.16).

Now, with suitable assumptions over the function f it can be shown there exist step sizes that satisfy the (strong) Wolfe conditions.

Lemma 2.2.2 *Suppose that the function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and let p_k be a descent direction for the function f at point x_k. Also, suppose that f is bounded from below on $\{x_k + \alpha p_k | \alpha > 0\}$. Then if $0 < \eta < c < 1$, there exist intervals of step lengths satisfying the (strong) Wolfe conditions.*

Notice that both versions of the second Wolfe condition are far too costly to compute since multiple evaluations of the derivative of f at each iteration are required. Because of that, this condition is usually not checked directly and instead, a backtracking strategy is usually employed. Algorithm 2 describes the backtracking strategy in more detail. The main idea is to consider a decreasing sequence of possible values of the step sizes $\{\alpha_j\}$, where $\alpha_{j+1} := \beta\alpha_j$ and $0 < \beta < 1$, until the suitable step size is found such that condition (2.15) is satisfied.

Algorithm 2: Backtracking.

S0 *Initialize* $0 < \alpha_0$, $\beta \in (0, 1)$, $\eta > 0$, x_k, p_k.
Set $j = 0$.

S1 **While** $\alpha_j > 0$ and $f(x + \alpha_j p_k) > f(x_k) + \eta \alpha_j \nabla f(x_k)^T p_k$ **do**

set $\alpha_{j+1} = \beta \alpha_j$.

End

S2 Return α_j.

2.3 Nonmonotone Strategy

At the end of this chapter, we will consider various nonmonotone techniques. In monotone line search methods, as we have already mentioned, α_k is chosen so that $f(x_{k+1}) < f(x_k)$, while in nonmonotone line search methods, some growth in the function value is allowed, i.e., a strict decrease of function value is not required in each iteration. In that manner, the set of admissible search directions is significantly enlarged and unnecessarily small steps at the beginning of the iterative procedure are prevented. It has been shown that nonmonotone line search methods especially outperform the monotone ones in the case where the iterative sequence $\{x_k\}$ is trapped near a narrow curved valley so that very short steps or zigzags occur [39, 40]. The nonmonotone methods could also be a better option for the stochastic optimization problems allowing in general larger step sizes, which are highly desirable when for example the Quasi-Newton or Newton methods are employed. Moreover, it is pointed out that the nonmonotone line search could improve the likelihood of finding a global optimal

solution and the rate of convergence [29].

Nonmonotone line search strategies are a well-developed class of methods for classical optimization, where the dominant three nonmonotone rules are originally presented by Grippo et al. [39], Zhang and Hager [103] and Li and Fukushima [69].

The first one was proposed for unconstrained optimization problems, where Newton's method was considered [39]. Unlike (2.15), this line search rule is given as follows

$$f(x_k + \alpha_k p_k) \leq \max_{i \in [\max\{1, k-c\}, k]} f(x_i) + \eta \alpha_k (\nabla f(x_k))^T p_k, \tag{2.17}$$

where $c \geq 1$, and $\eta \in (0, 1)$, while p_k has to be the descent direction.

Then, instead of using the maximum of previous function values, in the second approach it is suggested that a convex combination of previously computed function values should be used [103] such that for a descent direction p_k a step size α_k satisfies the condition

$$f(x_k + \alpha_k p_k) \leq D_k + \eta \alpha_k (\nabla f(x_k))^T p_k, \tag{2.18}$$

where D_k is defined with $D_0 = f(x_0)$ and

$$D_{k+1} = \frac{\eta_k q_k}{q_{k+1}} D_k + \frac{1}{q_{k+1}} f(x_{k+1}),$$

where $q_0 = 1$ and

$$q_{k+1} = \eta_k q_k + 1$$

with $\eta_k \in [\eta_{min}, \eta_{max}]$ and $0 \leq \eta_{min} \leq \eta_{max} \leq 1$.

Notice that the level of monotonicity is determined with parameter η_k in the following way. If $\eta_k = 1$ for every k, then the algorithm treats all previous function values equally, i.e.,

$$D_k = \frac{1}{k+1} \sum_{i=0}^{k} f(x_i),$$

while for $\eta_k = 0$ we have the standard Armijo rule. In [103] it is emphasized that the best numerical results are obtained for the value of η_k close to 1 when the iteration x_k is far from the solution and closer to 0 when we achieve neighborhood of the minimizer. The numerical results for $\eta_k = 0.85$ are reported and it is shown that one can provide satisfactory performance. Moreover, it has been proved that $D_k \geq f(x_k)$ so the line search rule is well defined.

Finally, the requirement that search direction p_k has to be descent, as in the above-stated nonmonotone line search rules (2.17) and (2.18), can be relaxed. In [69] a new line search rule for arbitrary search direction p_k is proposed for solving the system of nonlinear equations $F(x) = 0$, $F : \mathbb{R}^n \to \mathbb{R}^n$. [2]The rule can be expressed as

$$||F(x_k + \alpha_k p_k)|| \leq ||F(x_k)|| - \sigma_1 ||\alpha_k p_k||^2 + \varepsilon_k ||F(x_k)||,$$

where $\sigma_1 > 0$ and the following property of the parameter sequence $\{\varepsilon_k\}_{n \in \mathbb{N}}$ is assumed

$$\varepsilon_k > 0, \ \sum_k \varepsilon_k = \varepsilon < \infty.$$

This nonmonotone rule is successfully applied in many papers for deterministic and stochastic problems.

[2]Notice that for $f(x) = ||F(x)||$ or $f(x) = ||F(x)||^2$, the problems min $f(x)$ and $F(x) = 0$ are equivalent.

Chapter 3

Nonsmooth Optimization Methods

Nonsmooth optimization problems appear as important mathematical models today, starting with models of natural phenomena that exhibit sudden changes, shape optimization, to hinge loss functions in machine learning and deep neural networks. In practice, solving a nonsmooth convex problem tends to be more challenging usually more difficult and costly than a smooth one. Several approaches for nonsmooth problems involving high-dimensional data are available in the literature, starting with subgradient methods, cutting-plane, and bundle methods, gradient sampling, etc. Due to the general nature of the nonsmooth property, which requires employing subgradient directions and various convergence analysis concepts, the number of open problems is still very large. The purpose of this chapter is to give a brief overview of the optimization methods that can be used to solve nonsmooth optimization problems.

3.1 Subgradient Methods

The subgradient method was first introduced in the mid-sixties by N. Z. Shor, [93], and since then it has been extensively studied, [13, 31, 50, 85, 88]. The main advantage of this method is its simplicity and ease of implementation for a wide range of problems where the subdifferential of the nondifferentiable convex objective function can be easily computed. The basic idea is very similar to one that is used in the gradient descent algorithm for differentiable functions, but with some notable exceptions. First, the subgradient method can be applied directly to nondifferentiable functions. The next difference is in the step size strategy. While in the gradient method, the step sizes are usually chosen via an exact or approximate line search, in the subgradient method they are in most cases fixed in advance. And finally, unlike the gradient method, the subgradient method is not a descent method, i.e., the function value can increase at the new iteration.

3.1.1 Subgradient and Optimality Condition

Let us start with recalling some definitions that will be used throughout this thesis. The required concepts, such as subgradient and optimality conditions, were introduced by T.R. Rockafellar [88].

Definition 3.1.1 *A vector $g \in \mathbb{R}^n$ is a subgradient of $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$, at $x \in D$ if for all $z \in D$,*

$$f(z) \geq f(x) + g^T(z - x).$$

If f is convex and differentiable, then its gradient at x is a subgradient.

Definition 3.1.2 *The set of subgradients of f at the point x is called the subdifferential of f at x and it is denoted by*

$$\partial f(x) = \left\{ g \in \mathbb{R}^n | (\forall z \in \mathbb{R}^n) f(z) \geq f(x) + g^T(z - x) \right\}.$$

A function f is called subdifferentiable at x if there exists at least one subgradient at x. A function f is called subdifferentiable on D if it is subdifferentiable at all $x \in D$.

Definition 3.1.3 *The ε-subdifferential of function f at point x is defined as*

$$\partial_\varepsilon f(x) := conv\ (\partial f(\mathcal{B}(x, \varepsilon))).$$

Notice that the subdifferential $\partial f(x)$ is always closed and convex set, even if f is not convex, due to the fact that it is the intersection of an infinite set of halfspaces

$$\partial f(x) = \cap_{z \in D}\{g | f(z) \geq f(x) + g^T(z - x)\}.$$

In addition, if f is continuous at x, then the subdifferential $\partial f(x)$ is bounded.

Proposition 3.1.1 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then, for all $x \in \mathbb{R}^n$ the set $\partial f(x)$ is a nonempty, convex, compact subset of $\mathbb{R}^n$. In addition, f is L-Lipschitz function on $S \subset \mathbb{R}^n$ if and only if $||g|| \leq L$ for all $g \in \partial f(x)$ and $x \in S$.*

Lemma 3.1.1 *Suppose $f(x) = \sum_{i=1}^N f_i(x)$, where $f_1, \ldots, f_N$ are convex functions. Then we have*

$$\partial f(x) = \sum_{i=1}^{N} \partial f_i(x).$$

This property can be extended to infinite-sums, integrals, and expectations (provided they exist).

A point x^* is a minimizer of a function $f : D \to \mathbb{R}, D \subseteq \mathbb{R}^n$, if and only if f is subdifferentiable at x^* and

$$0 \in \partial f(x^*),$$

i.e., $g = 0$ is a subgradient of f at x^*. This follows directly from the fact that $f(x) \geq f(x^*)$ for all $x \in D$. And clearly if f is subdifferentiable at x^* with $0 \in \partial f(x^*)$, then

$$f(x) \geq f(x^*) + 0^T(x - x^*) = f(x^*) \text{ for all } x.$$

Notice that $0 \in \partial f(x^*)$ reduces to $\nabla f(x^*) = 0$ if convex function f is differentiable at x^*.

3.1.2 The Method

The idea behind the subgradient method is to generalize the classical gradient method to nonsmooth functions $f : \mathbb{R}^n \to \mathbb{R}$, such that starting from an initial point x_0, at each step k we choose $g_k \in \partial f(x_k)$ and set

$$x_{k+1} = x_k - \alpha_k g_k.$$

However, generalizing the gradient method is not straightforward, as the subgradient $g_k \in \partial f(x_k)$ does not have to be uniquely defined, even for convex f, and the choice of the vector $-g_k$ as search direction does not guarantee descent in f. Thus, this method may take steps that increase the value of function f.

Subgradient Descent

Let us start by showing how we can compute the direction of the steepest descent of a convex nonsmooth function. It is known that in the case when x is not a minimizer of f, the subdifferential $\partial f(x)$ always contains a vector g such that $-g$ is a descent direction for f [100]. Let us denote by g_{min} the vector that has the smallest norm of all the vectors in $\partial f(x)$. Then the vector $-g_{min}$ represents the direction of steepest descent for f at x. The vector g_{min} exists because $\partial f(x)$ is

nonempty and compact, and it can be expressed in the following way

$$g_{min} := \arg\min_{g \in \partial f(x)} ||g||. \tag{3.1}$$

Proposition 3.1.2 *[100] For a convex function $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$, and $x \in D$ that is not a minimizer of f, the vector $-g_{min}$ defined in* (3.1) *is a descent direction at x.*

Thus, we get a natural algorithm for minimizing convex, nonsmooth functions computing the minimum norm element of the subdifferential, and searching along the negative of this direction. However, computing the minimum norm element might be prohibitively expensive.

In the following, we will show that an algorithm that simply follows arbitrary subgradients can converge, under appropriate selections of steplengths. However, the convergence of these methods is quite slow, both in theory and practice.

Step Sizes

The classical subgradient method employs a predefined sequence of step sizes, which is a very different selection strategy with respect to the standard gradient method. Many different types of step size rules are used, where standard choices include a constant step size and also sequences that converge to zero sublinearly. Let us list only the basic possibilities for $\{\alpha_k\}$.

1) **Constant step size**

   a) $\alpha_k = \alpha$ is a positive constant, independent of k.

   b) $\alpha_k = \dfrac{\gamma}{||g_k||}$, where $\gamma > 0$. This means that the length of each step is constant, i.e., $||x_{k+1} - x_k|| = \gamma$.

2) **Square summable but not summable**
The step sizes satisfy

$$\alpha_k \geq 0, \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty, \quad \sum_{k=1}^{\infty} \alpha_k = \infty.$$

3) **Nonsummable diminishing**

a) The step sizes satisfy

$$\alpha_k \geq 0, \quad \lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty.$$

Step sizes that satisfy this condition are called diminishing step size rules.

b) The step sizes are chosen as

$$\alpha_k = \frac{\gamma_k}{||g_k||},$$

where

$$\gamma_k \geq 0, \quad \lim_{k\to\infty} \gamma_k = 0, \quad \sum_{k=1}^{\infty} \gamma_k = \infty.$$

Diminishing step size rules guarantee convergence to the optimal value of f as the number of iterations k goes to infinity, while for constant step sizes method yields a suboptimal function value with the approximation error. Now, we give convergence results and convergence rate estimates for the method using the above-defined stepsize rules.

Convergence Analysis

For the convergence analysis of the subgradient method, it is assumed that $||g|| \leq L$, $L > 0$, for all $g \in \partial f(x)$ and all x. Notice that this assumption implies that f must be Lipschitz with constant L. Also, denote by x^* a minimizer of f and define the distance between an initial point and the minimizer by R, i.e., $R := ||x_1 - x^*||$. In addition, as we have already pointed out that this method may take steps that increase the value of f, we need to keep track of the best point so far, i.e., the one with the smallest function value, so we define

$$f_k^{best} = \min\{f(x_1), \ldots, f(x_k)\}. \tag{3.2}$$

While in the standard gradient descent method the convergence proof is based on the function value decreasing at each step, in the subgradient method the key quantity is represented by the Euclidean distance to the optimal set. So, we have

$$\begin{aligned} ||x_{k+1} - x^*||_2^2 &= ||x_k - \alpha_k g_k - x^*||^2 \\ &= ||x_k - x^*||^2 - 2\alpha_k g_k^T (x_k - x^*) + \alpha_k^2 ||g_k||^2 \\ &\leq ||x_k - x^*||^2 - 2\alpha_k (f(x_k) - f^*) + \alpha_k^2 ||g_k||^2, \end{aligned}$$

where $f^* = f(x^*)$. Moreover, applying the inequality above recursively, as well as definition (3.2) and the stated assumptions, one can show that

$$f_k^{best} - f^* \leq \frac{R^2 + \sum_{i=1}^k \alpha_i^2 ||g_i||^2}{2 \sum_{i=1}^k \alpha_i},$$

i.e.,

$$f_k^{best} - f^* \leq \frac{R^2 + L^2 \sum_{i=1}^k \alpha_i^2}{2 \sum_{i=1}^k \alpha_i}. \tag{3.3}$$

Finally, from inequality (3.3) the convergence results for different step size rules can be obtained.

1) **Constant step size**

   a) For $\alpha_k = \alpha$ we have

$$f_{best}^k - f^* \leq \frac{R^2 + L^2\alpha^2 k}{2\alpha k},$$

   so

$$\lim_{k\to\infty} f_k^{best} \leq f^* + \frac{L^2\alpha}{2},$$

   i.e., f_k^{best} converges to $\frac{L^2\alpha}{2}$-vicinity of the optimal value.

   b) For $\alpha_k = \frac{\gamma}{||g_k||}$, $\gamma > 0$, we have

$$\lim_{k\to\infty} f_k^{best} \leq f^* + \frac{L\gamma}{2},$$

   i.e., f_k^{best} converges to $\frac{L\gamma}{2}$-vicinity of the optimal value.

2) **Square summable but not summable**
   For the step sizes which satisfy

$$\alpha_k \geq 0, \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty, \quad \sum_{k=1}^{\infty} \alpha_k = \infty, \tag{3.4}$$

   we have

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + L^2 \sum_{i=1}^k \alpha_i^2}{2\sum_{i=1}^k \alpha_i},$$

   so it holds

$$\lim_{k\to\infty} f_k^{best} = f^*.$$

3) **Nonsummable diminishing**

a) If the step sizes satisfy

$$\alpha_k \geq 0, \quad \lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty,$$

then

$$\lim_{k\to\infty} f_k^{best} = f^*.$$

b) If the step sizes are chosen as $\alpha_k = \frac{\gamma_k}{||g_k||}$, where

$$\gamma_k \geq 0, \quad \lim_{k\to\infty} \gamma_k = 0, \quad \sum_{k=1}^{\infty} \gamma_k = \infty,$$

then

$$\lim_{k\to\infty} f_k^{best} = f^*.$$

3.2 Cutting-Plane Methods

Another nonsmooth optimization method is the so-called cutting-plane method. This method was introduced by J.E. Kelley, [48]. While the simplicity of the subgradient algorithm comes at the price of ignoring past information, the idea of this method is to use this information obtained by an oracle in order to build a model of the function f itself.

The following optimization problem is considered

$$\min_{x\in\Omega} f(x),$$

where $\Omega \subseteq \mathbb{R}^n$ is a nonempty, closed and convex set, and $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function. The cutting-plane method relays on the approximation of the objective function from below. More precisely, it is

based on the following observation: from convexity, we know that

$$f(x) \geq f(y) + g^T(x - y),\ \forall x, y,$$

so we can approximate the objective function from below by a piecewise affine function

$$\bar{f}_k(x) = \max_{j=0,\ldots,k} (f(x_j) + g_j^T(x - x_j)), \tag{3.5}$$

where x_k represents the current iteration, $x_j, j = 0, \ldots, k-1$, are the auxiliary points, it is assumed that for each of them, a subgradient $g_j \in \partial f(x_j)$ is available, and $f(x_j) + g_j^T(x - x_j)$ corresponds to a cutting-plane. Then, for all $x \in \Omega$ it holds

$$f(x) \geq \bar{f}_k(x) = \max_{j=0,\ldots,k} (f(x_j) + g_j^T(x - x_j)).$$

In order to get a new iteration, we can replace f by its approximation $\bar{f}_k$ and solve the following problem

$$\min_{x \in \Omega} \max_{j=0,\ldots,k} (f(x_j) + g_j^T(x - x_j)) - f(x_k). \tag{3.6}$$

Furthermore, due to the subtraction of the function value $f(x_k)$ from the objective function, one is allowed to reformulate this nondifferentiable optimization problem (3.6) into a linear problem with constraints

$$\begin{aligned} &\min_{v \in \mathbb{R},\ x_k + p \in \Omega} v \\ &\text{s.t. } v \geq -a_j + g_j^T p, \forall j = 0, \ldots, k, \end{aligned}$$

where $p = x - x_k$ and the linearization error is given by $a_j := f(x_k) - f(x_j) - g_j^T(x_k - x_j)$. Starting with $x_0 \in \Omega$, the update is the following

$$x_{k+1} = x_k + p.$$

Notice that, obtaining a more precise model such that holds $\bar{f}_k(x) \leq \bar{f}_{k+1}(x)$, can be achieved by adding the corresponding cutting plane to the approximation $\bar{f}_k(x)$ given by (3.5). Furthermore, it is worth noting that in addition to obtaining an estimate x_{k+1} of the optimal point x^*, we get a lower bound $\overline{f}_{k+1}(x_{k+1})$ on the optimal value $f(x^*)$ as well. It is reasonable to terminate the process when $a_{k+1} \leq \varepsilon$, i.e., $f(x_{k+1}) - \overline{f}_{k+1}(x_{k+1}) \leq \varepsilon$, which guarantees that $f(x_{k+1}) \leq f(x^*) + \varepsilon$. So, the advantage of the cutting-plane method compared to the subgradient method is that the model (3.5) provides a stopping criterion (based on the linearization error), which did not exist for the subgradient method.

However, the cutting-plane method may converge slowly in practice as subsequent solutions can be very distant, exhibiting a zig-zag behavior. Therefore, many cutting planes do not actually contribute to the approximation of f around the optimum x^*. To overcome this shortage, bundle methods were developed in order to reduce this behavior by adding a stabilization term to (3.5) [50].

3.3 Bundle Methods

Another important class of nonsmooth optimization methods, which can be seen as a stabilization of the cutting-plane method, are bundle methods. They attempt to combine the practical advantages of the cutting-plane method with the theoretical strengths of a proximal point method. The main difference refers to adding an extra point called the center, $\bar{x}_k$, to the bundle of information. The same piecewise-linear model for the function (3.5) is used, but without solving a linear problem at each iteration. Instead, the next iteration is computed by the Moreau-Yosida regularization for $\bar{f}_k$ at $\bar{x}_k$ in the

following way

$$x_{k+1} = \arg\min_{x \in \Omega} \bar{f}_k(x) + \frac{\mu_k}{2} ||x - \bar{x}_k||^2. \tag{3.7}$$

The quadratic term in the relation (3.7) serves the purpose of stabilizing the cutting-plane method. More precisely, it makes the next iteration closer to the current center $\overline{x}_k$ by avoiding drastic movements as in the case of cutting planes. The parameter μ_k controls the trade-off between minimizing $\bar{f}$ and staying close to a point $\bar{x}_k$ which is known to be good. These methods offer an advantage over classical subgradient methods as they use more information about the local behavior of the function. This is achieved by approximating the subdifferential of the objective function using a bundle of subgradients from previous iterations. Thus, instead of using only one arbitrary subgradient at each point, the idea is to make an approximation of the whole subdifferential of the objective function. Moreover, it is possible to define a stopping criterion, which is an additional advantage compared to subgradient methods. However, the drawback of these methods is the requirement of solving at least one quadratic programming subproblem in each iteration, which can be time-consuming, especially for large-scale problems.

A comprehensive review of the history and development of bundle methods can be found in [97]. A great number of bundle methods in combination with Newton-type methods [73] and Trust Region methods [1] have been developed. The modifications of bundle methods have been used for nonconvex problems [80], constrained problems [89], and multi-criteria problems [79]. Proximal bundle methods (see for example [42, 64, 75]) are based on the bundle methodology and have the ability to provide exact solutions even if most of the time the available information is inaccurate, unlike their forerunner variants. The proximal bundle method takes ideas from the subgradient method and the proximal method [47, 83]. The first one can be seen

as an extension of gradient methods in smooth optimization and the second one is a variant of the proximal point method, which minimizes the original function plus a quadratic part.

3.4 Gradient Sampling Methods

One of the key characteristics of gradient sampling (GS) methods, which represents some of the latest approaches in nonsmooth optimization, is that no subgradient information is required. Thus, this is simple descent method for solving nonsmooth, nonconvex optimization problems with a solid theoretical foundation and it has been employed in a wide variety of applications. The underlying motivation of this method is that some nonsmooth objective functions $f : \mathbb{R}^n \to \mathbb{R}$ (for example locally Lipschitz continuous functions) are differentiable almost everywhere. So, f is differentiable at a randomly generated point $x \in \mathbb{R}^n$ with probability one. In other words, an algorithm can obtain, as in the case when f is a smooth function, the objective function value $f(x)$ and the gradient $\nabla f(x)$, instead of requiring an oracle to compute a subgradient.

The original method was first introduced by J.V. Burke, A.S. Lewis, and M.L. Overton in [18]. They considered a locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$, which is continuously differentiable on an open dense subset $D \subset \mathbb{R}^n$, and has bounded level sets. At each iteration, the gradients of f at the current iteration x_k and at $m \geq n+1$ points from the $\mathcal{B}(x_k, \varepsilon)$ are computed in order to find an optimal descent direction - p_k. The next step is the computation of a step size - α_k. Thus, employing the Armijo line search one can calculate the step size and obtain the next iteration candidate, $x_{k+1} = x_k + \alpha_k p_k$, such that the condition of the sufficient descent holds. In the case $x_{k+1} \notin D$, a new, slightly perturbed point from D is chosen such that the Armijo condition is preserved.

One can distinguish two convergence results regarding the ε radius. In the first one, for the constant radius ε, it can be shown with probability 1 that the GS algorithm generates a sequence with an accumulation point, which is ε-stationary (i.e., $0 \in \partial_\varepsilon f(x)$). In another case, where ε is dynamically reducing, if the algorithm converges to a point, the limit of the sequence is a stationary point for f with probability 1. Another important reference on this topic is [51], where the stronger convergence results of the gradient sampling method can be found. It has been shown that every accumulation point generated by the GS algorithm is ε-stationary, almost always. While in the case where ε is dynamically reduced, every accumulation point of an arbitrary subsequence is stationary, without the assumption that the whole sequence converges.

In conclusion, the GS method offers a valuable approach to optimization problems characterized by nonsmooth and nonconvex objective functions, where traditional gradient-based methods may face challenges. By incorporating gradient information and random sampling, the method enables efficient exploration of the search space and can yield an improved approximate solution. However, it is essential to consider some limitations of the method. These include potential slower convergence rates, sensitivity to sampling strategies, the need for careful selection of sample sizes, dependence on initialization, limited theoretical guarantees, and the requirement for tuning parameters. Despite these drawbacks, the GS method remains a promising tool for a wide range of optimization applications. Thus, there exists a considerable body of literature related to the GS method. For example, an approach for nonconvex, nonsmooth constrained problems is considered in [24], while in [25] an adaptive gradient sampling approach which reduces significantly the number of required gradient evaluations is proposed. Moreover, in [26] the adaptive gradient sampling idea is combined with quasi-Newton methods.

3.5 BFGS Method

The famous BFGS method for minimization of smooth convex functions has been popular for decades due to its superior practical performance [82]. Despite the fact that the BFGS direction generated at a nonsmooth point is not necessarily a descent direction, recent theoretical developments have revealed that it can also be an effective general-purpose tool for nonsmooth optimization [41, 65, 66, 67, 102].

Now, we give an overview of results from [102] that will be used in the algorithms introduced later in Chapters 5 and 6.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a nonsmooth convex function and recall that

$$\partial f(x) = \{g \in \mathbb{R}^n : f(y) \geq f(x) + g^T(y - x),\ y \in \mathbb{R}^n\}$$

is the subdifferential of f at a point $x \in \mathbb{R}^n$, where the vectors $g \in \partial f(x)$ represent subgradients. Clearly, if f is a differentiable convex function then $\partial f(x) = \nabla f(x)$. Moreover, recall that a function f attains its global minimum at $x^* \in \mathbb{R}^n$ if and only if $0 \in \partial f(x^*)$.

The BFGS algorithm defined in [102] relays on the descent direction property defined for nonsmooth functions as follows. The direction p is a descent direction for f at $x \in \mathbb{R}^n$ if

$$g^T p < 0 \text{ for all } g \in \partial f(x),$$

or equivalently, if

$$\sup_{g \in \partial f(x)} g^T p < 0.$$

However, generating such a direction is not an easy task in general.

Let us now briefly recall the algorithm for finding the descent direction presented in [102, Algorithm 2, p. 1155]. The pseudo-quadratic model of f at $x \in \mathbb{R}^n$ is given by

$$Q(p) = f(x) + Y(p),$$

where

$$Y(p) = \frac{1}{2}p^T B^{-1}p + \sup_{g\in\partial f(x)} g^T p,$$

and $B \in \mathbb{R}^{n\times n}$ is a nonsingular matrix. The linear part in the above model is $\sup_{g\in\partial f(x)} g^T p$ and it is assumed that an oracle for computing the supremum is available.

The iterative procedure, represented through Algorithm 3, guarantees to find a quasi-Newton descent direction p, assuming an oracle that supplies $\sup_{g\in\partial f(x)} g^T p$ for a given direction. The input parameters are a subgradient $\tilde{g}_0 \in \partial f(x)$, a direction-finding tolerance $\epsilon \geq 0$, iteration bound $i_{max} \in \mathbb{N}$, matrix B from the quadratic model and an oracle to calculate $\arg\sup_{g\in\partial f(x)} g^T p$ for any given x and p.

The initial subgradient $\tilde{g}_0 \in \partial f(x)$ in Algorithm 3 is chosen arbitrarily and it is assumed that $\tilde{g}_{i+1} = \arg\sup_{g\in\partial f(x)} g^T p_i$ is provided by an oracle. The following Lemma ensures that if the point x is not optimal, then there exists a direction-finding tolerance $\epsilon \geq 0$ for Algorithm 3 such that the returned search direction p is descent direction.

Lemma 3.5.1 *[102, Lemma 3, p. 1187] Let $B \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix with all eigenvalues larger than $m > 0$. If the point x is not stationary and if the number of iterations i_{max} in Algorithm 3 is unbounded, then there exists a tolerance $\epsilon \geq 0$ such that the descent direction*

$$p = -B\overline{g}, \overline{g} \in \partial f(x)$$

returned by Algorithm 3 satisfies

$$\sup_{g\in\partial f(x)} g^T p < 0.$$

The descent direction p generated through Algorithm 3 satisfies the inequality below under the conditions stated in Lemma 3.5.1 [102,

Algorithm 3: p =descentDirection($\tilde{g}_0 \in \partial f(x), \epsilon\ , i_{max}, B$)

S0 Initialize $i = 0$, $\overline{g}_0 = \tilde{g}_0$, $p_0 = -B\tilde{g}_0$.

S1 Calculate the next subgradient $\tilde{g}_1 = \underset{g \in \partial f(x)}{\arg\sup}\, g^T p_0$.

S2 Compute $\epsilon_0 := p_0^T \tilde{g}_1 - p_0^T \overline{g}_0$.

S3 **While** $\left(\tilde{g}_{i+1}^T p_i > 0 \quad \text{or} \quad \epsilon_0 > \epsilon\right) \quad \text{and} \quad \epsilon_i > 0 \quad \text{and} \quad i < i_{max}$ **do**

$$\mu^* := \min\left[1, \frac{(\overline{g}_i - \tilde{g}_{i+1})^T B \overline{g}_i}{(\overline{g}_i - \tilde{g}_{i+1})^T B (\overline{g}_i - \tilde{g}_{i+1})}\right]$$

$$\overline{g}_{i+1} = (1 - \mu^*)\overline{g}_i + \mu^* \tilde{g}_{i+1}$$

$$p_{i+1} = (1 - \mu^*) p_i - \mu^* B \tilde{g}_{i+1}$$

$$\tilde{g}_{i+2} = \underset{g \in \partial f(x)}{\arg\sup}\, g^T p_{i+1}$$

$$\epsilon_{i+1} = \min_{j \le i+1}\left[p_j^T \tilde{g}_{j+1} - \frac{1}{2}\left(p_j^T \overline{g}_j + p_{i+1}^T \overline{g}_{i+1}\right)\right]$$

$$i := i + 1$$

End

S4 Compute $p = \arg\min_{j \le i} Y(p_j)$.

S5 **If** $\sup_{g \in \partial f(x)} g^T p < 0$ then return p,
else return failure.

Corollary 4, p. 1188],

$$\sup_{g \in f(x)} g^T p \leq -\frac{1}{2}\overline{g}^T B \overline{g} \leq -\frac{m}{2}||\overline{g}||^2 < 0. \tag{3.8}$$

Furthermore, the following also holds.

Theorem 3.5.1 *[102] Let p be a descent direction at an iteration x. If $\varphi(\alpha) := f(x + \alpha p)$ is bounded from below, then there exists $\alpha' > 0$ such that the subgradient Armijo condition*

$$f(x + \alpha p) \leq f(x) + c_1 \alpha \sup_{g \in \partial f(x)} g^T p,$$

holds for all $\alpha \in [0, \alpha']$, where $0 < c_1 < 1$.

Chapter 4

Minimization of Expected Value Function

In many optimization problems, the objective function cannot be computed exactly due to some kind of random noise. A typical example would be the minimization of a function stated in the form of mathematical expectation given that the exact analytical expression for the expectation is rarely available and, even if it is available, it is not computable exactly. Thus one has to approximate the objective function with some sample-based approximation.

The importance of the stochastic optimization problem arising from various scientific fields generated a large amount of literature in recent years. A number of important problems can be stated in this form - starting from data analytics with huge data sets which require working with subsamples or machine learning problems with online data sets that continually increase and change [20], to simulations of natural and industrial processes with a number of random parameters [19, 76, 78, 98].

Within this chapter, we are going to set the framework for the main subject of this thesis - the minimization of the expected value function,

where we will actually consider the stochastic optimization problems with underlying randomness rather than the stochastic problems with intentionally imposed noise. In order to do that, first in Section 4.1 we will present in detail the optimization problem and then provide some necessary assumptions for algorithms used to solve the problem under consideration, which will be described later in Chapters 5 and 6. In Section 4.2 we then focus on the approximation of the original problem using Sample Average Approximation (SAA) and the quality of the approximate solution thus obtained.

4.1 Problem Description

Let us begin by introducing the following constrained optimization problem

$$\min_{x\in\Omega} f(x) = E\left[F(x,\xi)\right], \tag{4.1}$$

where $\Omega \subset \mathbb{R}^n$ is a convex, closed set, $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is continuous and convex function with respect to x, bounded from below, $\xi : \mathcal{A} \to \mathbb{R}^m$ is random vector and $(\mathcal{A}, \mathcal{F}, P)$ is a probability space. Notice that F is locally Lipschitz as a consequence of convexity [4] but possibly nonsmooth. Thus, the main issues that arise in iterative methods for solving (4.1) are the approximation of the objective function and the choice of search directions. The step size is a challenging issue in stochastic analysis as well and it was a subject of research in many papers, [31, 36, 43, 49, 84, 90]. Line search methods, which are an important tool in deterministic optimization, are not easily extended to the stochastic case due to the mutual dependence of step size and search direction, which are both random variables in the stochastic framework. An important study on this topic is given in [84] where the approximations of the objective function and its gradient are assumed to be good enough with a fixed high probability. Under these settings,

the complexity analysis in terms of the expected number of iterations to reach a near-optimal solution is provided. In [27] a second-order direction is considered but an additional sampling is used in Armijo-like condition to overcome the bias issue.

Assumptions

Now, we will introduce the standard assumptions on the objective function, which summarize the problem properties.

Assumption A 1 *Assume that functions* $f_i(x) = F(x, \xi_i)$, $i = 1, 2, \ldots,$ *are continuous, convex, and bounded from below with a constant* C.

Assumption A 2 *Assume that the function* F *is dominated by a* P*-integrable function on any compact subset of* $\mathbb{R}^n$.

The previous Assumption A2 is satisfied if there exists a nonnegative function $\bar{F}(\xi)$ such that $E\left[\bar{F}(\xi)\right] < \infty$ and $\mathrm{P}\big(|F(x,\xi)| \leq \bar{F}(\xi)\big) = 1$ for every $x \in \Omega$. Notice that this condition holds if the function F is bounded with some finite constant $\bar{F}$, i.e., if $|F(x,\xi)| \leq \bar{F}$ for every $x \in \Omega$ and almost every ξ.

4.2 Sample Average Approximation

Due to the difficulty in computing the mathematical expectation in general, the most common approach is to approximate the original objective function $f(x)$ by applying the SAA function. SAA method is an approach for solving stochastic optimization problems, where the expected objective function of the stochastic problem is approximated by a sample average estimate derived from a random sample. This random sample can be viewed as historical data of N observations of

ξ, or it can be generated by Monte Carlo sampling techniques. More precisely, for a given independent and identically distributed (i.i.d.) sample vectors $\xi_i, i \in \mathcal{N}$, the SAA approximate objective function is defined as

$$f_{\mathcal{N}}(x) = \frac{1}{N} \sum_{i \in \mathcal{N}} f_i(x), \tag{4.2}$$

where $f_i(x) = F(x, \xi_i)$ and $N = |\mathcal{N}|$ determines the size of a sample used for approximation. The sample vectors $\xi_i, i \in \mathcal{N}$ are assumed to be i.i.d. and the sample size N determines the accuracy of the approximation (4.2), [92]. Naturally, larger N implies higher accuracy of the approximate function $f_{\mathcal{N}}$, but makes any optimization algorithm more costly as the cost of computing $f_{\mathcal{N}}$, as well as search directions, increases with N. Even if the original problem is already in the SAA form, i.e., if we are dealing with finite-sum problems, the costs of employing the full sample at each iteration can be large, and thus variable sample size (VSS) strategies are often applied. There is a vast literature dealing with VSS methods for SAA approximations, [6, 7, 44, 54, 52, 63], which range from simple heuristics to complex schemes, all of them with the idea of using cheaper, lower accuracy approximations of the objective function whenever possible, in order to save the computational effort. In the following two chapters, we will be concerned with the directions for choosing sample size N such that the solution of the SAA problem provides a good approximation of the original problem solution.

The second issue one needs to address is the choice of search directions. In the case of smooth problems we can choose between relatively slow but cheap first-order methods or more elaborate and more costly second-order methods, depending on a particular problem structure, needed accuracy, etc. In the case of nonsmooth problems, as we have already mentioned in the previous chapter, the gradient is generally replaced by a subgradient or more elaborate schemes like gradient

sampling, [25, 51], bundle methods, [73], proximal methods, [64], and so on. A number of recent papers deal with second-order search directions [2, 3, 46].

An important special case of a constrained optimization problem (4.1) that we will consider is when $\Omega = \mathbb{R}^n$. Then, this is actually an unconstrained optimization problem with the objective function in the form of mathematical expectation. Let us formally state this problem that we will consider in Chapter 6

$$\min_x f(x) = E\left[F(x, \xi)\right]. \tag{4.3}$$

where $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is continuous and convex function with respect to x, bounded from below, $\xi : \Omega \to \mathbb{R}^m$ is random vector and $(\mathcal{A}, \mathcal{F}, P)$ is probability space, as we have stated before. Convexity implies that F is locally Lipschitz, [4] and no additional smoothness assumption is imposed.

Also, we will be interested in the problem of finite-sum, as a special case of problem (4.2), which can be expressed in the following form

$$\min_{x \in \Omega} f_{\mathrm{N}}(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x), \tag{4.4}$$

where the functions $f_i(x) = F(x, \xi_i)$ are again continuous, convex and possibly nonsmooth.

4.2.1 SAA error

SAA is a powerful tool for solving optimization problems given in the form (4.1) and (4.3), as it can provide a good approximation of the expected value of the function without calculating it exactly. On the other hand, a solution of the SAA problem (4.2) serves only as an approximate solution of the original problem, thus the main concern is the quality measure of SAA solutions. More precisely, we are

interested in the quality of the solution of the SAA problem (4.2) and the needed conditions for achieving the convergence to the optimal solution of the original problem. It has been shown that under mild conditions an optimal solution and the optimal value of the SAA problem converge exponentially fast to their true counterparts as the sample size N increases. Comprehensive material about the statistical properties of the SAA estimators can be found in [92].

For well-defined and finite function $f(x) = E[F(x,\xi)]$ and random variables ξ_i, $i = 1, 2, \ldots$ with the same distribution as ξ the function $f_{\mathcal{N}}(x) = \frac{1}{N}\sum_{i\in\mathcal{N}} F(x,\xi_i)$ is also a random variable since it depends on a random sample, where the sample used to approximate the objective function is denoted by $\mathcal{N}$, while N denotes its cardinality. Moreover, if the sample is i.i.d., then by the (strong) Law of Large Numbers the almost sure convergence of $f_{\mathcal{N}}(x)$ is obtained. More precisely, for every x it holds

$$\lim_{N\to\infty} f_{\mathcal{N}}(x) = f(x) \quad \text{a.s.} \tag{4.5}$$

This result represents the consistency of the SAA estimator, which is important because it gives a certain assurance that the error of the estimation approaches zero in the limit a.s. as the sample size grows to infinity. On the other hand, $f_{\mathcal{N}}$ is unbiased estimator of $f(x)$ due to

$$E[f_{\mathcal{N}}(x)] = \frac{1}{N}\sum_{i\in\mathcal{N}} E[F(x,\xi_i)] = \frac{1}{N}\sum_{i\in\mathcal{N}} E[F(x,\xi)] = f(x).$$

Also, the variance can be expressed in the following way

$$D[f_{\mathcal{N}}(x)] = \frac{1}{N^2}\sum_{i\in\mathcal{N}} D[F(x,\xi_i)] = \frac{1}{N^2}\sum_{i\in\mathcal{N}} D[F(x,\xi)] = \frac{1}{N} D[F(x,\xi)].$$

The next theorem provides a stronger result than (4.5) - the uniform convergence (Uniform Law of Large Numbers - ULLN).

Theorem 4.2.1 *[92] Suppose that S is nonempty, compact subset of $\mathbb{R}^n$ and that for any $x \in S$ the function $F(\cdot, \xi)$ is continuous at x for almost every ξ. Furthermore, suppose that the sample $\xi_1, \ldots, \xi_N$ is i.i.d. and that the function $F(x, \xi)$, $x \in S$, is dominated by an integrable function. Then $f(x) = E[F(x, \xi)]$ is finite valued and continuous on S and*

$$\lim_{N\to\infty} f_{\mathcal{N}}(x) = f(x) \quad \textit{a.s. uniformly on } S,$$

i.e.,

$$\lim_{N\to\infty} \sup_{x\in S} |f_{\mathcal{N}}(x) - f(x)| = 0 \quad \textit{a.s.} \tag{4.6}$$

For the problems (4.1) and (4.3) notice that Assumption A1 implies that f is a convex and continuous function as well as $f_{\mathcal{N}}$ for any given $\mathcal{N}$. Moreover, if the conditions of the previous theorem are satisfied, we have that $f_{\mathcal{N}}$ a.s. converges uniformly to f on any compact subset $S \subseteq \mathbb{R}^n$. Also, notice that $\sup_{x\in S} |f_{\mathcal{N}}(x) - f(x)| = 0$ holds trivially if the sample is finite and the full sample is eventually achieved and retained.

Two approaches are distinguished depending on the sample size. In the first one, we have a finite sample size and it is assumed to be determined before the process of optimization starts. The second one deals with unbounded sample size. In both cases, the main issue is how to change the sample size N_k during the optimization process, i.e., across the iterations. In order to define the rule for updating the sample size, we introduce the SAA error measure $h(N_k)$, i.e., a proxy for $|f(x_k) - f_{\mathcal{N}_k}(x_k)|$, as follows.

In the finite-sum case with the full sample size $N_{max} < \infty$ we define

$$h(N_k) = \frac{N_{max} - N_k}{N_{max}}, \tag{4.7}$$

while in general (unbounded sample size) case we define

$$h(N_k) = \frac{1}{N_k}. \tag{4.8}$$

Notice that in both cases we have that function $h : \mathbb{N} \to [0, 1]$ is monotonically decreasing and strictly positive if the full sample is not attained. Moreover, in the finite-sum case we have $h(N_k) = 0$ if and only if $N_k = N_{max}$, while in unbounded sample case we have $\lim_{N_k \to \infty} h(N_k) = 0$. Other choices are eligible as well, but we will keep these for simplicity.

In addition to the sample size, it is important to point out that there is also a difference in how the new sample is chosen, in terms of whether it contains the previous one or represents a different realization of the sample. In the case where we are dealing with a priory realized sample, the cumulative sample means that at each iteration a new sample is appended to the previous one. On the other hand, in the VSS methods one can use different sample realizations in each iteration, i.e., the samples in two consecutive iterations are independent-noncumulative.

Recall that consistency of f_N, i.e., almost sure convergence of $f_N(x)$ towards $f(x)$, is achieved if the sample is i.i.d. and the considered functions are well defined and finite. In [44] it is shown that the condition on the identically distributed sample can be relaxed if the sample size increases at a certain rate. More precisely, the sequence of sample sizes $\{N_k\}$ should increase such that one of the following properties is satisfied:

$$\sum_{k=1}^{\infty} \alpha^{N_k} < \infty \text{ for all } \alpha \in (0, 1) \tag{4.9}$$

or

$$\sum_{k=1}^{\infty} \frac{1}{N_k} < \infty.$$

The considered condition (4.9) holds for $N_k \geq k$ for example. However, too fast an increase in the sample size can result in an inefficient algorithm. Therefore, the consistency estimator, more precisely, the

upper bound of the error $|f_{\mathcal{N}_k}(x) - f(x)|$, plays an important role in VSS methods.

For the case of cumulative i.i.d samples, one possible bound derived in [44] is

$$|f_{\mathcal{N}_k}(x) - f(x)| \leq C\sqrt{\frac{\ln(\ln N_k)}{N_k}}, \tag{4.10}$$

where $C = C(x)$ is positive parameter related to the variance of function $F(x, \xi)$. On the other hand, in the case of a noncumulative sample, then if the sample size increases fast enough, i.e.,

$$N_k \geq ak^b \text{ for } a > 0, b > 2,$$

it follows

$$|f_{\mathcal{N}_k}(x) - f(x)| \leq C\frac{\ln N_k}{N_k}$$

for k large enough, where $C = C(x)$ is again a positive parameter related to the variance of function $F(x, \xi)$.

Chapter 5

Nonsmooth Methods with Variable Sample Size for Constrained Optimization Problems

Within this chapter, we consider constrained optimization problems with a nonsmooth objective function in the form of a mathematical expectation, previously defined as (4.1). The framework that uses SAA to approximate the objective function, which is either unavailable or too costly to compute, is proposed and this part of the thesis represents the original contribution [56, 62]. The proposed algorithms combine a SAA subgradient with the spectral coefficient in order to find a suitable direction that improves the performance of the first-order method, as shown by numerical results in Sections 5.1.3 and 5.2.3. The step sizes are chosen from the predefined interval and the almost sure convergence of the methods is proved under the standard assumptions in a stochastic environment.

Two approaches are distinguished within this chapter depending

on the variable sample size strategy. Although finding a good way of varying the sample size during the optimization process may affect the algorithm a lot, a suitable sample size strategy will not be the main concern in Section 5.1, where the first Algorithm SPS (Spectral Projected Subgradient) is proposed [56]. In order to prove a.s. convergence it will be assumed only that the sample size tends to infinity. The problem of determining the optimal VSS strategy in this framework is considered in Algorithm AN-SPS (Adaptive Sample Size Nonmonotone Line Search Spectral Projected Subgradient) proposed in Section 5.2 [62].

We propose a framework for solving nonsmooth constrained optimization problem (4.1) assuming that the feasible set Ω is easy to project on (for example a box or a ball in $\mathbb{R}^n$). This allows us to apply a method of the Spectral Projected Gradient (SPG) type.

5.1 Spectral Projected Subgradient Method

Spectral Projected Gradient (SPG) is a well-developed method with an abundance of literature covering theory and applications. For example, SPG method for finite-sum problems has been studied in [9, 95]. In [95] SPG is used in combination with the stochastic gradient method and convergence is proved under the assumption that the full gradient is computed in every m iterations. In [9], the subsampled spectral gradient methods are analyzed and the effects of the choice of spectral coefficient are investigated. In [37] the SPG direction is employed within Inexact Restoration framework to address nonlinear optimization problems with nonconvex constraints. The SPG methods for problems with continuously differentiable objective function given in the form of mathematical expectation have been analyzed in [54].

The method proposed in this part of the thesis is a subgradient method, but it differs in several ways from the methods available in the literature. We propose a way to plug VSS-SPG ideas into the nonsmooth framework. Since the objective function may be nonsmooth, we have to use subgradients instead of gradients. Thus, in Section 5.1.1 we refer to the core algorithm as SPS - Spectral Projected Subgradient method. The spectral coefficient is calculated by employing consecutive subgradients of possibly different SAA functions and the safeguard which provides positive, bounded spectral coefficients is used. We prove a.s. convergence under the standard assumptions for the stochastic environment in Section 5.1.2. Moreover, in order to improve the performance of the algorithm, we also propose a line search variant of SPS named LS-SPS. The line search is defined in such a way that LS-SPS falls into the SPS framework and thus the same convergence results hold. In spite of the fact that the descent property of the search direction is desirable, it is not necessary in each iteration to ensure the convergence result. The proposed line search is well-defined and the a.s. convergence is achieved even if the search direction is not a descent one for the SAA function.

Although the proposed algorithms are constructed to cope with unbounded sample sizes, they can also be applied to finite-sum problems and we devote part of the consideration to this important special class as well.

In Section 5.1.1, the stochastic SPG method is adapted to the nonsmooth framework. The a.s. convergence of the proposed SPS method is proved under the standard assumptions in Section 5.1.2 and the SPS is further upgraded by introducing a specific line search technique resulting in LS-SPS in Section 5.1.2. In Section 5.1.3, numerical results on machine learning problems show the efficiency of the proposed method, especially LS-SPS.

5.1.1 The Algorithm

Now we describe the subsampled spectral projected subgradient framework algorithm for nonsmooth problems - SPS. For any given $z \in \mathbb{R}^n$, we denote by $P_\Omega(z)$ the orthogonal projection of z onto Ω. Recall that $\mathcal{N}_k$ denotes the sample used to approximate the objective function and N_k denotes its cardinality.

Algorithm 4: SPS (**S**pectral **P**rojected **S**ubgradient Method for Nonsmooth Optimization)

S0 *Initialization.* Given $N_0 \in \mathbb{N}$, $x_0 \in \Omega$, $0 < C_1 < 1 < C_2 < \infty$, $0 < \underline{\zeta} \leq \overline{\zeta} < \infty$, $\zeta_0 \in \left[\underline{\zeta}, \overline{\zeta}\right]$.
Set $k = 0$.

S1 *Direction.* Choose $\bar{g}_k \in \partial f_{\mathcal{N}_k}(x_k)$ and set $p_k = -\zeta_k \bar{g}_k$.

S2 *Step size.*
If $k = 0$, set $\alpha_0 = 1$.
Else, choose $\alpha_k \in [C_1/k, C_2/k]$.

S3 *Main update.* Set $x_{k+1} = P_\Omega(x_k + \alpha_k p_k)$ and $s_k = x_{k+1} - x_k$.

S4 *Sample size update.* Chose $N_{k+1} \in \mathbb{N}$.

S5 *Spectral coefficient update.*
Calculate $y_k = g_{\mathcal{N}_k}(x_{k+1}) - \bar{g}_k$, where $g_{\mathcal{N}_k}(x_{k+1}) \in \partial f_{\mathcal{N}_k}(x_{k+1})$.
Set $\zeta_{k+1} = \min\{\overline{\zeta}, \max\{\underline{\zeta}, \frac{s_k^T s_k}{s_k^T y_k}\}\}$.

S6 Set $k := k + 1$ and go to **S1**.

Let us now comment on the Algorithm 4. In Step S1 we calculate the direction by choosing a subgradient of the SAA function $f_{\mathcal{N}_k}$ and

taking the opposite direction multiplied by the spectral coefficient. Notice that the safeguard in Step S5 ensures that the negative subgradient direction is retained. However, this direction does not have to be descent for the function $f_{\mathcal{N}_k}$ since we take an arbitrary subgradient.

Within Step S2 we choose the step size α_k from the given interval. The constants C_1 and C_2 can be arbitrary small and large, respectively, allowing a wide range of feasible step sizes. This choice was motivated by the common assumption on the step size sequence (see (3.4)):

$$\alpha_k \geq 0, \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty, \quad \sum_{k=1}^{\infty} \alpha_k = \infty. \tag{5.1}$$

Notice that the choice in Step S2 ensures that the sequence of step sizes of SPS algorithm satisfies (5.1). After finding the direction and the step size, we project the point $x_k + \alpha_k p_k$ onto the feasible set Ω and thus we retain feasibility in all the iterations of the algorithm.

In Step S4, we choose the sample size to be used in the subsequent iteration. To prove the convergence result, we assume that N_k tends to infinity or achieves and retains the maximal sample size in the case of finite-sums. Thus, the simplest way to ensure this is to increase the sample size at each iteration. However, we formulate this step as generally as possible to emphasize that other choices are feasible as well, including some adaptive strategies.

In Step S5 y_k is calculated as a difference of two subgradients of the same approximate function $f_{\mathcal{N}_k}$, but different approaches are feasible as well. For instance, one can use subgradients of different functions $y_k = g_{\mathcal{N}_{k+1}}(x_{k+1}) - g_{\mathcal{N}_k}(x_k)$. This can reduce the costs, especially if the sample is not cumulative, but also brings additional noise into the spectral coefficient since the subgradients are calculated for two different functions in general. However, if we have a finite-sum problem and the full sample is reached, the cost of calculating the subgradient may be reduced to one subgradient per iteration since one can

obviously take $\bar{g}_{k+1} = g_{\mathcal{N}_k}(x_{k+1})$. In general, another choice could be $y_k = g_{\mathcal{N}_{k+1}}(x_{k+1}) - g_{\mathcal{N}_{k+1}}(x_k)$. This reduces the influence of noise and usually provides a better approximation of the spectral coefficient of the true objective function, but it requires additional evaluations. Although the choice of y_k was addressed in the literature (see [9] for example), in general, it remains an open question that requires thorough analysis before drawing the final conclusions. It is important to point out that the choice of y_k does not affect the convergence analysis and the theoretical results obtained in the next section, but it may affect the algorithm's performance significantly.

5.1.2 Convergence Analysis

In this subsection, we analyze conditions needed for a.s. convergence of the SPS algorithm. Recall that samples are assumed to be i.i.d. In the convergence theorems, it will be assumed that the standard assumptions for the stochastic environment, Assumptions A1 and A2 stated in the previous chapter within Section 4.1, are satisfied. More precisely, it will be assumed that $f_i(x) = F(x, \xi_i)$, $i = 1, 2, \ldots$, are continuous, convex, and bounded from below with a constant C, and the function F is dominated by a P-integrable function on any compact subset of $\mathbb{R}^n$.

The main result, a.s. convergence of Algorithm SPS, is stated in the following theorem. We assume also that the feasible set is compact, although this assumption may be relaxed as we will show in the sequel. Moreover, recall that the convexity implies that the functions f_i are locally Lipschitz continuous, and thus for every x and i there exists $L_i(x)$ such that for all $g \in \partial f_i(x)$ there holds $\|g\| \leq L_i(x)$. However, since there can be infinitely many functions f_i in general we assume that the chosen subgradients are uniformly bounded.

Let X^* and f^* be the set of solutions and the optimal value of

problem (4.1), respectively. Define the SAA errors sequence as

$$e_k = \max_{x \in \Omega} |f_{\mathcal{N}_k}(x) - f(x)|.$$

The convergence result is as follows.

Theorem 5.1.1 *Suppose that Assumptions A1 and A2 hold and* $\{x_k\}$ *is a sequence generated by Algorithm SPS where* $N_k \to \infty$. *Assume also that* Ω *is compact and convex and there exists* L *such that* $||\bar{g}_k|| \leq L$ *for all* k. *Then*

$$\liminf_{k \to \infty} f(x_k) = f^* \quad a.s. \tag{5.2}$$

Moreover, if $\sum_{k=0}^{\infty} e_k/k < \infty$, *then*

$$\lim_{k \to \infty} x_k = x^* \quad a.s. \tag{5.3}$$

for some $x^* \in X^*$.

Proof.

Denote by $\mathcal{W}$ the set of all possible sample paths of SPS algorithm. Suppose that (5.2) does not hold, i.e., $\liminf_{k\to\infty} f(x_k) = f^*$ does not happen with probability 1. In that case there exists a subset of sample paths $\tilde{\mathcal{W}} \subseteq \mathcal{W}$ such that $P(\tilde{\mathcal{W}}) > 0$ and for every $w \in \tilde{\mathcal{W}}$ there holds

$$\liminf_{k \to \infty} f(x_k(w)) > f^*,$$

i.e., there exists $\varepsilon(w) > 0$ small enough such that

$$f(x_k(w)) - f^* \geq 2\varepsilon(w)$$

for all k. Since f is continuous on the feasible set Ω, there exists $\tilde{y}(w) \in \Omega$ such that $f(\tilde{y}(w)) = f^* + \varepsilon(w)$. This further implies

$$f(x_k(w)) - f(\tilde{y}(w)) = f(x_k(w)) - f^* - \varepsilon(w) \geq 2\varepsilon(w) - \varepsilon(w) = \varepsilon(w).$$

Let us take an arbitrary $w \in \tilde{\mathcal{W}}$. Denote $z_{k+1}(w) := x_k(w) + \alpha_k(w)p_k(w)$. Notice that non-expansivity of orthogonal projection and the fact that $\tilde{y} \in \Omega$ together imply

$$||x_{k+1}(w) - \tilde{y}(w)|| = ||P_\Omega(z_{k+1}(w)) - P_\Omega(\tilde{y}(w))|| \leq ||z_{k+1}(w) - \tilde{y}(w)||. \tag{5.4}$$

Furthermore, using the fact that $\bar{g}_k$ is subgradient of the convex function $f_{\mathcal{N}_k}$, $\bar{g}_k \in \partial f_{\mathcal{N}_k}(x_k)$, we have

$$f_{\mathcal{N}_k}(x_k) - f_{\mathcal{N}_k}(\tilde{y}) \leq \overline{g}_k^T(x_k - \tilde{y})$$

and dropping w in order to facilitate the reading we obtain

$$\begin{aligned}
||z_{k+1} - \tilde{y}||^2 &= ||x_k + \alpha_k p_k - \tilde{y}||^2 = ||x_k - \alpha_k\zeta_k\overline{g}_k - \tilde{y}||^2 \\
&= ||x_k - \tilde{y}||^2 - 2\alpha_k\zeta_k\overline{g}_k^T(x_k - \tilde{y}) + \alpha_k^2\zeta_k^2||\overline{g}_k||^2 \\
&\leq ||x_k - \tilde{y}||^2 + 2\alpha_k\zeta_k(f_{\mathcal{N}_k}(\tilde{y}) - f_{\mathcal{N}_k}(x_k)) + \alpha_k^2\zeta_k^2||\overline{g}_k||^2 \\
&\leq ||x_k - \tilde{y}||^2 + 2\alpha_k\zeta_k(f(\tilde{y}) - f(x_k) + 2e_k) + \alpha_k^2\zeta_k^2||\overline{g}_k||^2 \\
&\leq ||x_k - \tilde{y}||^2 - 2\alpha_k\zeta_k(f(x_k) - f(\tilde{y})) + 4e_k\alpha_k\overline{\zeta} + \alpha_k^2\overline{\zeta}^2L^2 \\
&\leq ||x_k - \tilde{y}||^2 - 2\alpha_k\underline{\zeta}\varepsilon + 4e_k\alpha_k\overline{\zeta} + \alpha_k^2\overline{\zeta}^2L^2 \\
&= ||x_k - \tilde{y}||^2 - \alpha_k\left(2\underline{\zeta}\varepsilon - 4e_k\overline{\zeta} - \alpha_k\overline{\zeta}^2L^2\right).
\end{aligned} \tag{5.5}$$

By ULLN we have $\lim_{k\to\infty} e_k = 0$ a.s., or more precisely, $\lim_{k\to\infty} e_k(w) = 0$ for almost every $w \in \mathcal{W}$. Since $P(\tilde{\mathcal{W}}) > 0$, there must exist a sample path $\tilde{w} \in \tilde{\mathcal{W}}$ such that

$$\lim_{k\to\infty} e_k(\tilde{w}) = 0.$$

This further implies the existence of $\tilde{k}(\tilde{w}) \in \mathbb{N}$ such that for all $k \geq \tilde{k}(\tilde{w})$ we have

$$\alpha_k(\tilde{w})\overline{\zeta}^2L^2 + 4e_k(\tilde{w})\overline{\zeta} \leq \varepsilon(\tilde{w})\underline{\zeta} \tag{5.6}$$

because Step S2 of SPS algorithm implies that $\lim_{k\to\infty} \alpha_k = 0$ for any sample path. Furthermore, since (5.5) holds for all $w \in \tilde{\mathcal{W}}$ and thus for $\tilde{w}$ as well, from (5.4)-(5.6) we obtain

$$\begin{aligned}
||x_{k+1}(\tilde{w}) - \tilde{y}(\tilde{w})||^2 &\leq ||x_k(\tilde{w}) - \tilde{y}(\tilde{w})||^2 \\
&- \alpha_k(\tilde{w}) \left(2\underline{\zeta}\varepsilon(\tilde{w}) - 4e_k(\tilde{w})\overline{\zeta} - \alpha_k(\tilde{w})\overline{\zeta}^2 L^2\right) \\
&= ||x_k(\tilde{w}) - \tilde{y}(\tilde{w})||^2 \\
&+ \alpha_k(\tilde{w}) \left(4e_k(\tilde{w})\overline{\zeta} + \alpha_k(\tilde{w})\overline{\zeta}^2 L^2 - 2\underline{\zeta}\varepsilon(\tilde{w})\right) \\
&\leq ||x_k(\tilde{w}) - \tilde{y}(\tilde{w})||^2 + \alpha_k(\tilde{w}) \left(\underline{\zeta}\varepsilon(\tilde{w}) - 2\underline{\zeta}\varepsilon(\tilde{w})\right) \\
&= ||x_k(\tilde{w}) - \tilde{y}(\tilde{w})||^2 - \alpha_k(\tilde{w})\underline{\zeta}\varepsilon(\tilde{w})
\end{aligned}$$

and

$$||x_{k+s}(\tilde{w}) - \tilde{y}(\tilde{w})||^2 \leq ||x_k(\tilde{w}) - \tilde{y}(\tilde{w})||^2 - \varepsilon(\tilde{w})\underline{\zeta} \sum_{j=0}^{s-1} \alpha_j(\tilde{w}).$$

Letting $s \to \infty$ yields a contradiction since

$$\sum_{k=0}^{\infty} \alpha_k \geq \sum_{k=0}^{\infty} C_1/k = \infty$$

for any sample path and we conclude that (5.2) holds.

Now, let us prove (5.3), i.e., $\lim_{k\to\infty} x_k = x^*$ a.s., under the additional assumption $\sum_{k=0}^{\infty} e_k/k < \infty$. Notice that this assumption implies that $\sum_{k=0}^{\infty} \alpha_k e_k < \infty$ since $\alpha_k \leq C_2/k$. Since (5.2) holds, we know that

$$\liminf_{k\to\infty} f(x_k(w)) = f^*, \tag{5.7}$$

for almost every $w \in \mathcal{W}$. In other words, there exists $\overline{\mathcal{W}} \subseteq \mathcal{W}$ such that $P(\overline{\mathcal{W}}) = 1$ and (5.7) holds for all $w \in \overline{\mathcal{W}}$. Let us consider arbitrary $w \in \overline{\mathcal{W}}$. We will show that

$$\lim_{k\to\infty} x_k(w) = x^*(w) \in X^*$$

which will imply the result (5.3). Once again let us drop w to facilitate the notation.

Let $K_1 \subseteq \mathbb{N}$ be a subsequence of iterations such that

$$\lim_{k\in K_1} f(x_k) = f^*.$$

Since $\{x_k\}_{k\in K_1} \subseteq \{x_k\}_{k\in\mathbb{N}}$ and $\{x_k\}_{k\in\mathbb{N}}$ is bounded because of feasibility and the compactness of the feasible set Ω, there follows that $\{x_k\}_{k\in K_1}$ is also bounded and there exist $K_2 \subseteq K_1$ and $\tilde{x}$ such that

$$\lim_{k\in K_2} x_k = \tilde{x}. \tag{5.8}$$

Then, we have

$$f^* = \lim_{k\in K_1} f(x_k) = \lim_{k\in K_2} f(x_k) = f(\lim_{k\in K_2} x_k) = f(\tilde{x}).$$

Therefore, $f(\tilde{x}) = f^*$ and we have $\tilde{x} \in X^*$.

Now, we show that the whole sequence of iterates converges. Let

$$\{x_k\}_{k\in K_2} := \{x_{k_i}\}_{i\in\mathbb{N}}.$$

Following the steps of (5.5) and using the fact that $f(x_k) \geq f(\tilde{x})$ for all k, we obtain that the following holds for any $s \in \mathbb{N}$

$$\begin{aligned} ||x_{k_i+s} - \tilde{x}||^2 &\leq ||x_{k_i} - \tilde{x}||^2 + 4\overline{\zeta}\sum_{j=0}^{s-1} e_{k_i+j}\alpha_{k_i+j} + \overline{\zeta}^2 L^2 \sum_{j=0}^{s-1} \alpha_{k_i+j}^2 \\ &\leq ||x_{k_i} - \tilde{x}||^2 + 4\overline{\zeta}\sum_{j=0}^{\infty} e_{k_i+j}\alpha_{k_i+j} + \overline{\zeta}^2 L^2 \sum_{j=0}^{\infty} \alpha_{k_i+j}^2 \\ &= ||x_{k_i} - \tilde{x}||^2 + 4\overline{\zeta}\sum_{j=k_i}^{\infty} e_j\alpha_j + \overline{\zeta}^2 L^2 \sum_{j=k_i}^{\infty} \alpha_j^2 =: a_i. \end{aligned}$$

Thus, for any $s, m \in \mathbb{N}$ there holds

$$||x_{k_i+s} - x_{k_i+m}||^2 \leq 2||x_{k_i+s} - \tilde{x}||^2 + 2||x_{k_i+m} - \tilde{x}||^2 \leq 4a_i.$$

Since $\sum_{j=k_i}^{\infty} e_j\alpha_j$ and $\sum_{j=k_i}^{\infty} \alpha_j^2$ are residuals of the convergent sums and (5.8) holds, we have

$$\lim_{i\to\infty} a_i = 0.$$

Therefore, for every $\varepsilon > 0$ there exists $k_i \in \mathbb{N}$ such that for all $t, l \geq k_i$ there holds $||x_t - x_l|| \leq \varepsilon$, i.e., the sequence $\{x_k\}_{k\in\mathbb{N}}$ is a Cauchy sequence and thus convergent. This, together with (5.8), implies

$$\lim_{k\to\infty} x_k = \tilde{x}. \blacksquare$$

Let us comment on e_k first. We obtain (5.2) provided that the sample size tends to infinity in an arbitrary manner. The stronger result (5.3) is achieved under the assumption of fast enough increase of the sample size, i.e., if it holds $\sum_{k=0}^{\infty} e_k/k < \infty$. Having in mind the interval for step size α_k, we conclude that the assumption $\sum_{k=0}^{\infty} e_k/k < \infty$ is satisfied if $e_k \leq C_3 k^{-\nu}$ holds for all k large enough and arbitrary $C_3 > 0$, where $\nu > 0$ can be arbitrary small. While in general, this can be hard to guarantee, for some classes of functions F (e.g., function

plus noise with finite variance, $F = f + \xi$, $\mathrm{Var}(\xi) < \infty$), the error bound for cumulative samples (4.10) stated in Section 4.2.1 yields $e_k \leq C\sqrt{\frac{\ln \ln N_k}{N_k}}$, for all k large enough and some positive constant C directly dependent on the noise variance. In that case, it can be shown that the simple choice of $N_k = k$ provides the sufficient growth needed for (5.3).

In the case of finite-sum problem (4.4), we do not need Assumption A2 since (4.6) is trivially satisfied if the full sample size is eventually achieved and retained. Thus, $e_k = 0$ for all k large enough and $\sum_{k=0}^{\infty} e_k/k < \infty$ trivially holds. Moreover, the compactness of the feasible set and convexity of f imply the Lipschitz continuity of each f_i on Ω, and thus the subgradients $\bar{g}_k$ are uniformly bounded. Furthermore, we also know that the functions f_i are uniformly bounded from below on Ω. Although the sample paths may differ, the convergence result is deterministic since the original objective function $f_{\mathcal{N}}$ is eventually used. We summarize the result in the following corollary of the previous theorem.

Corollary 5.1.1 *Suppose that Assumption A1 holds and* $\{x_k\}$ *is a sequence generated by Algorithm SPS applied to* (4.4)*. Suppose that* $N_k = N$ *for all* k *large enough and* Ω *is compact and convex. Then*

$$\lim_{k \to \infty} x_k = x^* \in X^*.$$

Since the compactness of Ω excludes the important class of constrained problems, such as $x \geq 0$, which appear as subproblems in many cases, it is important to comment on the alternatives. The assumption of bounded Ω may be replaced with the assumption of bounded iterate sequence generated by Algorithm SPS. We state the result for completeness.

Theorem 5.1.2 *Suppose that Assumptions A1 and A2 hold and* $\{x_k\} \subseteq \bar{\Omega}$, *where* $\bar{\Omega} \subseteq \Omega$ *is bounded and* $N_k \to \infty$. *Assume that* Ω *is closed and convex and there exists* L *such that* $||\bar{g}_k|| \leq L$ *for all* k. *Then*

$$\liminf_{k\to\infty} f(x_k) = f^* \, a.s.$$

Moreover, if $\sum_{k=0}^{\infty} e_{k,\bar{\Omega}}/k < \infty$ *we have*

$$\lim_{k\to\infty} x_k = x^* \in X^* \ a.s.,$$

where $e_{k,\bar{\Omega}} := \max_{x\in\bar{\Omega}} |f_{\mathcal{N}_k}(x) - f(x)|$.

Let us now see under which conditions we obtain the boundedness of iterations. But first, define the SAA error sequence as

$$\bar{e}_k = |f_{\mathcal{N}_k}(x_k) - f(x_k)| + |f_{\mathcal{N}_k}(x^*) - f(x^*)|, \tag{5.9}$$

where $x^* \in X^*$ is an arbitrary solution point. We have the following result.

Proposition 5.1.1 *Suppose that Assumptions A1 and A2 hold and* $\{x_k\}$ *is a sequence generated by Algorithm SPS where* $N_k \to \infty$. *Assume that* Ω *is closed and convex and there exists* L *such that* $||\bar{g}_k|| \leq L$ *for all* k. *Then if* $\sum_{k=0}^{\infty} \bar{e}_k/k \leq C_4 < \infty$, *there exists a compact set* $\bar{\Omega} \subseteq \Omega$ *such that* $\{x_k\} \subseteq \bar{\Omega}$.

Proof. Let x^* be an arbitrary solution of the problem (4.1). Then, by following similar steps as in (5.5) and using the non-expansivity of

the projection (5.4), we obtain the following for an arbitrary k,

$$\begin{aligned}
||x_{k+1} - x^*||^2 &\leq ||z_{k+1} - x^*||^2 = ||x_k - \alpha_k \zeta_k \overline{g}_k - x^*||^2 \\
&= ||x_k - x^*||^2 - 2\alpha_k \zeta_k \overline{g}_k^T (x_k - x^*) + \alpha_k^2 \zeta_k^2 ||\overline{g}_k||^2 \\
&\leq ||x_k - x^*||^2 + 2\alpha_k \zeta_k (f_{\mathcal{N}_k}(x^*) - f_{\mathcal{N}_k}(x_k)) + \alpha_k^2 \zeta_k^2 ||\overline{g}_k||^2 \\
&\leq ||x_k - x^*||^2 + 2\alpha_k \zeta_k (f(x^*) - f(x_k) + \bar{e}_k) + \alpha_k^2 \zeta_k^2 ||\overline{g}_k||^2 \\
&\leq ||x_k - x^*||^2 + 2\bar{e}_k \overline{\zeta} C_2 / k + \alpha_k^2 \overline{\zeta}^2 L^2 \\
&\leq ||x_0 - x^*||^2 + 2C_2 \overline{\zeta} \sum_{k=0}^{\infty} \frac{\bar{e}_k}{k} + \overline{\zeta}^2 L^2 \sum_{k=0}^{\infty} \frac{C_2^2}{k^2}.
\end{aligned}$$

Thus, there exists a constant C_5 such that $||x_k - x^*|| \leq C_5$ for all k, which completes the proof. ■

We summarize the convergence result for the unbounded feasible set in the following theorem.

Theorem 5.1.3 *Suppose that Assumptions A1 and A2 hold, the feasible set Ω is convex and closed and $\{x_k\}$ is a sequence generated by Algorithm SPS where N_k tends to infinity fast enough to provide $\sum_{k=0}^{\infty} \bar{e}_k / k \leq C_4 < \infty$ with $\bar{e}_k$ given by (5.9). Then*

$$\liminf_{k \to \infty} f(x_k) = f^* \; a.s.$$

Moreover, if $\sum_{k=0}^{\infty} e_k / k < \infty$ then

$$\lim_{k \to \infty} x_k = x^* \; a.s.$$

for some $x^* \in X^*$, where $e_k = \max_{x \in \bar{\Omega}} |f_{\mathcal{N}_k}(x) - f(x)|$ and $\bar{\Omega}$ is a compact set containing $\{x_k\}$.

Improving the Efficiency - Line Search SPS

Notice that SPS algorithm works with an arbitrary subgradient direction related to the current SAA function. However, in some applications such as Hinge Loss binary classification, it is possible to provide

a descent direction with respect to the SAA function by applying the procedure proposed in [102] or gradient subsampling technique [17] for instance. On the other hand, it is well known that applying line search may improve the performance of the algorithm significantly, even in a stochastic environment. Thus, in order to make the SPS algorithm more efficient, we propose a line search technique adapted to the nonsmooth VSS framework such that the SPS convergence analysis still holds.

The proposed line search does not require a descent search direction in order to be well defined, nor the convergence analysis depends on the descent property. So, the following sufficient decrease property is desirable, but not necessary in order to prove the convergence of the Line Search SPS (LS-SPS) algorithm presented in the sequel,

$$\sup_{g\in\partial f_{\mathcal{N}_k}(x_k)} g^T p_k \leq -\frac{m}{2}\|p_k\|^2 \quad \text{for some } m>0. \tag{5.10}$$

In general case, it is not an easy task to find a direction p_k that satisfies the previous condition, but in some important cases, it can be done (see Algorithm 3).

The LS Procedure

Since we employ the spectral subgradient method, we use nonmonotone Armijo-type line search condition

$$f_{\mathcal{N}_k}(x_k+\alpha_k p_k) \leq F_k - \eta\alpha_k||p_k||^2, \tag{5.11}$$

where p_k is the search direction as in Step S1 of Algorithm 4 and

$$F_k = \max_{i\in[\max\{1,k-c\},k]} f_{\mathcal{N}_i}(x_i).$$

The candidates for α_k that we consider are:

$$\overline{\alpha}_k \text{ and } \frac{\overline{\alpha}_k + 1/k}{2},$$

where $\overline{\alpha}_k = \min\{1, C_2/k\}$, $1 < C_2 < \infty$. The reasoning behind this is the following. The choice of $\alpha_k = 1/k$ is a typical choice that is suitable for obtaining a.s. convergence, so we put $C_1 = 1$. The line search is employed to estimate if the larger value of α_k may be used. Since the backtracking techniques usually start with 1, we take the minimum of 1 and C_2/k as the initial choice. Although C_2/k must be included to ensure the theoretical requirements of Step S2, one can take C_2 arbitrary large such that $\overline{\alpha}_k = 1$ even for the large values of k. Thus, practically, 1 would be the initial choice in all practical applications. We set the middle of the interval $[1/k, \overline{\alpha}_k]$ as the second possible choice for step size in line search. Although other strategies are feasible as well, we reduce to these two guesses to avoid the computational costs of unsuccessful line search attempts. Thus, if $\alpha_k = \overline{\alpha}_k$ satisfies (5.11), we take this as a step size. If not, we check (5.11) with the medium value $\alpha_k = (\overline{\alpha}_k + 1/k)/2$. If the condition is satisfied, we retain this choice, otherwise, we set $\alpha_k = 1/k$.

Remark. LS-SPS algorithm falls into the framework of SPS algorithm as α_k satisfies the condition (5.11). Thus, the whole convergence analysis presented for the SPS algorithm also holds for LS-SPS.

5.1.3 Numerical Results

We performed numerical experiments on the set of binary classification problems listed in Table 5.1. The problems are modeled by the L_2-regularized Hinge Loss. More precisely, we consider the following optimization problem for learning with a Support Vector Machine introduced in [91]

$$\min_{x \in \Omega} f(x) := 10||x||^2 + \frac{1}{N}\sum_{i=1}^{N} \max\{0, 1 - z_i x^T w_i\}, \qquad (5.12)$$
$$\Omega := \{x \in \mathbb{R}^n \ : ||x||^2 \leq 0.1\},$$

Algorithm 5: LS-SPS (**L**ine **S**earch **S**pectral **P**rojected **S**ubgradient Method for Nonsmooth Optimization)

S0 *Initialization.* Given $N_0 \in \mathbb{N}$, $x_0 \in \Omega$, $1 < C_2 < \infty$, $0 < \underline{\zeta} \leq \overline{\zeta} < \infty$, $\zeta_0 \in \left[\underline{\zeta}, \overline{\zeta}\right]$, $\eta \in (0,1)$, $c \in \mathbb{N}$. Set $k = 0$.

S1 *Direction.* Choose $\bar{g}_k \in \partial f_{\mathcal{N}_k}(x_k)$ satisfying (5.10) if possible and set $p_k = -\zeta_k \bar{g}_k$.

S2 *Step size.*
If $k = 0$, set $\alpha_0 = 1$.
Else,

choose $\alpha_k \in \{\overline{\alpha}_k, (\overline{\alpha}_k + 1/k)/2\}$ such that

$$f_{\mathcal{N}_k}(x_k + \alpha_k p_k) \leq F_k - \eta \alpha_k ||p_k||^2,$$

holds if possible.

Otherwise set $\alpha_k = \frac{1}{k}$.

S3 *Main update.* Set $x_{k+1} = P_\Omega(x_k + \alpha_k p_k)$ and $s_k = x_{k+1} - x_k$.

S4 *Sample size update.* Chose $N_{k+1} \in \mathbb{N}$.

S5 *Spectral coefficient update.*
Calculate $y_k = g_{\mathcal{N}_k}(x_{k+1}) - \bar{g}_k$, where $g_{\mathcal{N}_k}(x_{k+1}) \in \partial f_{\mathcal{N}_k}(x_{k+1})$.
Set $\zeta_{k+1} = \min\{\overline{\zeta}, \max\{\underline{\zeta}, \frac{s_k^T s_k}{s_k^T y_k}\}\}$.

S6 Set $k := k + 1$ and go to **S1**.

where $w_i \in \mathbb{R}^n$ are the input features and $z_i \in \{1, -1\}$ are the corresponding labels. Thus, we have a convex problem with the compact feasible set easy to project on. Moreover, for this kind of problems, it is possible to calculate the descent direction and we use the procedure proposed in Section 3.5, Algorithm 3, as a subroutine that provides the descent property (5.10). We employ this subroutine in all the tested algorithms to ensure a fair comparison.

	Data set	N	n
1	SPLICE [105]	3175	60
2	MUSHROOMS [70]	8124	112
3	ADULT9 [105]	32561	123
4	MNIST(binary) [106]	70000	784

Table 5.1: Properties of the datasets used in the experiments.

Our numerical study has several goals. It is designed to investigate:

i) whether the variable sample size approach remains beneficial in the nonsmooth environment with a bounded full sample;

ii) whether introducing the line search pays off;

iii) whether the spectral coefficient improves the efficiency of the projected subgradient method.

We set the experiments as follows. The main criterion for comparison of the methods will be the computational cost modeled by FEV (number of function evaluations). More precisely, FEV_k^m represents the number of scalar products needed for method m to calculate x_k (starting from x_0). We also track the value of the true objective function across the iterations to observe the progress of the considered method.

To answer question i), we compare the VSS methods to their full sample counterparts. The extension $-F$ (e.g., SPS-F) will indicate that the full sample is used at every iteration, i.e., $N_k = N$ for all k. On the other hand, we assume the following sample size increase for the VSS methods

$$N_{k+1} = \lceil \min\{1.1N_k, N\} \rceil$$

with $N_0 = 0.1N$. Obviously, there are many other choices that can be more efficient, but we choose this simple increase to be tested in the initial phase of the method evaluation. To address question ii) we compare LS-SPS algorithm to SPS algorithm with the standard choice of the step size $\alpha_k = 1/k$. Finally, to address iii), we compare the proposed methods to the first-order subgradient method denoted by LS-PS (Line Search Projected Subgradient), which can be viewed as a special case of LS-SPS with $\underline{\zeta} = \overline{\zeta} = 1$. We also test the VSS and the full sample alternatives of the projected subgradient method: LS-PS and LS-PS-F, respectively. The results of the subgradient method with the choice of $\alpha_k = 1/k$ were poor and thus they are not reported here.

The relevant parameters are as follows. The initial points are chosen randomly $x_0 \in \Omega$ and we use the same initial point for all the tested methods within one run. The step size parameters are $C_1 = 10^{-2}, C_2 = 10^2$ and the line search is performed with $\eta = 10^{-4}$ and $c = 5$. For the proposed spectral methods, the safeguard parameters are $\underline{\zeta} = 10^{-4}$ and $\overline{\zeta} = 10^4$.

We perform 5 independent runs for each of the data sets which yields 20 runs of each method in total. A demonstrative run is presented in Figure 5.1 where the objective function $f(x_k)$ is plotted against the FEV_k. It reveals that LS-SPS methods outperform other methods reaching a tighter vicinity of the solution. Even if we use the spectral coefficient, the predetermined step size was not enough to bring the sequence to the same vicinity as obtained by the LS coun-

terparts. On the other hand, observing the LS-PS method which uses line search but without a spectral coefficient, we can see that the line search itself (without second-order information) was not enough to push the subgradient method towards the solution. Finally, notice that the computational cost is reduced significantly by employing the VSS scheme in LS-SPS method.

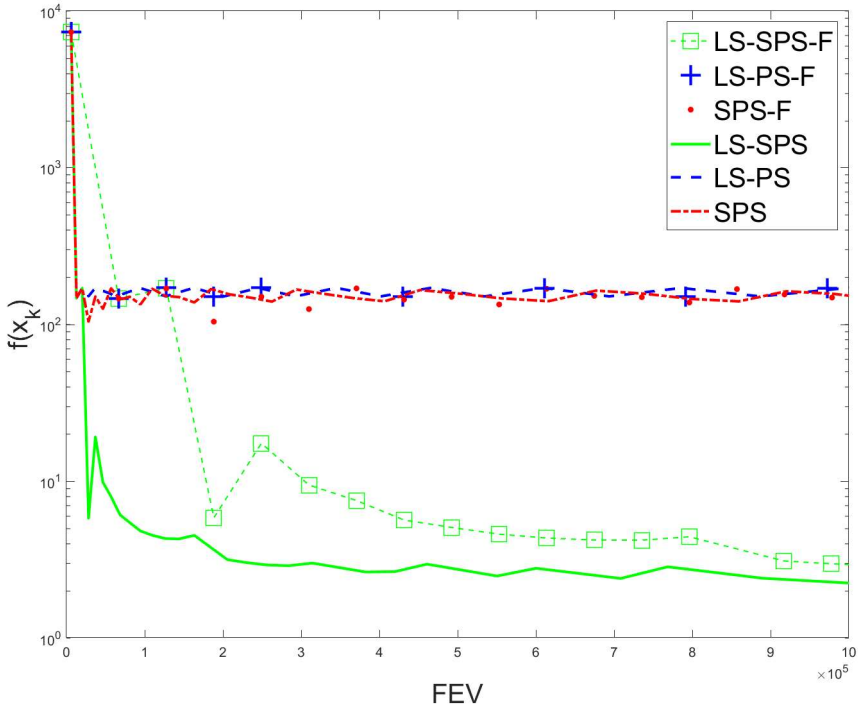

Figure 5.1: MNIST data set

In order to compare the tested methods taking into account all the

runs and data sets, we employ the metric based on the ideas of the performance profile [30] adapted to the stochastic environment in [63]. Roughly speaking, we estimate the probability of winning for each of the tested methods. For the considered run, the method m wins if it reaches the vicinity τ of the solution with the smallest costs. Since the theoretical stopping criterion is non-existing, we stop the methods if the maximal number of scalar products is achieved. At each iteration k we measure the distance from the solution by observing the relative error of method m with respect to the optimal value, i.e.,

$$r_k^m := \frac{f(x_k^m) - f^*}{f^*}.$$

The optimal value f^* on each data set is computed using SPS-F algorithm such that $f^* \approx f_{\mathrm{N}}(x_{1000})$. For each method m and each run l we register the first iteration $k(m,l)$ at which we have $r^m_{k(m,l)} \leq \tau$ and read the corresponding $FEV^m_{k(m,l)}$. Then, the method m earns a point in run l if

$$FEV^m_{k(m,l)} = \min_j FEV^j_{k(j,l)}.$$

Finally, we estimate the probability of winning, denoted by π by

$$\pi = \frac{t}{T},$$

where t is the number of earned points and T is the total number of runs. Notice that in the described situation we can have more winners, in other words, more methods can share first place if they reach the goal with the same costs.

The results are presented in Figure 5.2 for different relative errors $\tau \in [0.01, 3.5]$. They reveal that the VSS methods clearly outperform their full sample counterparts and that LS-SPS method turns out to be the best possible choice according to the conducted experiments. The algorithms LS-PS and SPS reach 1 for very large values of the

relative error τ which is not relevant, so we do not show this part of the graph.

Figure 5.3 represents a classical performance profile (PP) graph for fixed relative error $\tau = 1$. The FEV is kept as the criterion for the classical PP as well. On the y-axes we plot the probability that the method is close enough to the best one, where "close enough" is determined by the value on x-axes denoted by q. More precisely, retaining the same notation as above, the method m earns a PP(q) point in run l if $FEV^m_{k(m,l)} \leq q \min_j FEV^j_{k(j,l)}$ and the plotted values correspond to

$$\pi_{PP}(q) = \frac{t_{PP}(q)}{T},$$

where $t_{PP}(q)$ is the number of earned PP(q) points for the considered method. Again, from this figure, it is clear that LS-SPS is the most robust, i.e., it has the highest probability of being the optimal solver.

Additional Comparison

Now we show additional numerical results in order to compare the proposed algorithm with the proximal bundle method (PBM). It is known that PBM gives the best result under a fixed number of iterations. The reason behind this is the fact that the number of constraints in the quadratic program solved by PBM may grow linearly with the number of iterations [47]. Accordingly, PBM may become significantly slower when the number of iterations becomes larger. For that reason, we compare the fixed number of iterations of LS-SPS-F with PBM that use full sample size in all iterations, and after that LS-SPS with VSS PBM (where the sample is changed in the same way as in LS-SPS through iterations).

In order to ensure a fair comparison, we choose several different combinations of initial parameters for PBM. Table 5.2 summarizes the properties of the observed methods, where γ is the proximity control

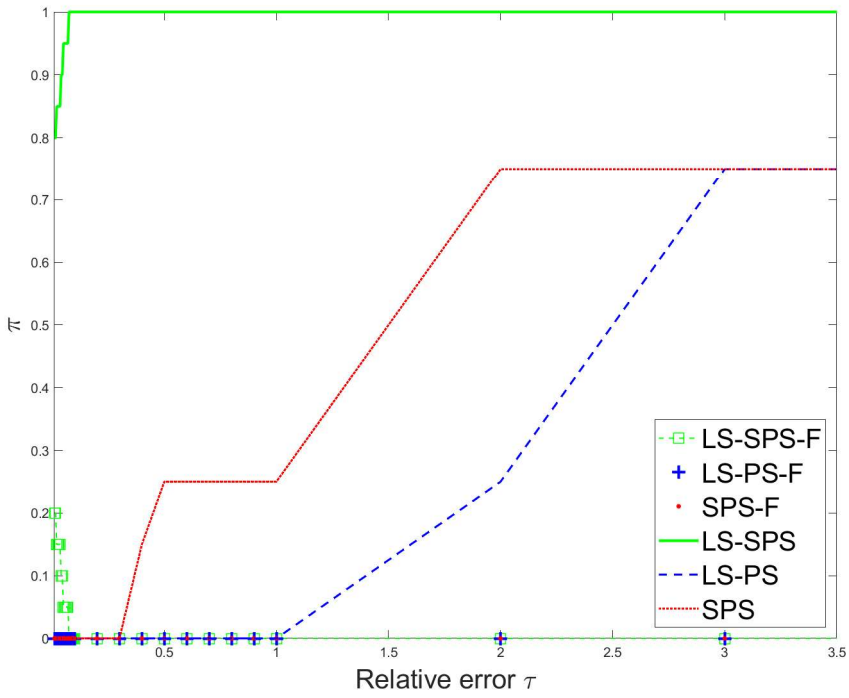

Figure 5.2: Empirical probabilities of winning (π) for different relative errors (τ).

parameter, m is the descent coefficient, ϵ is the tolerance parameter and ω is the decay coefficient. Detailed information about these parameters and implementations in Matlab of PBM is available at [107].

Figure 5.4 shows the results for Full and VSS versions of LS-SPS and PBM algorithms on MNIST data set, while the results on the other three data sets are similar. The objective function $f(x_k)$ is

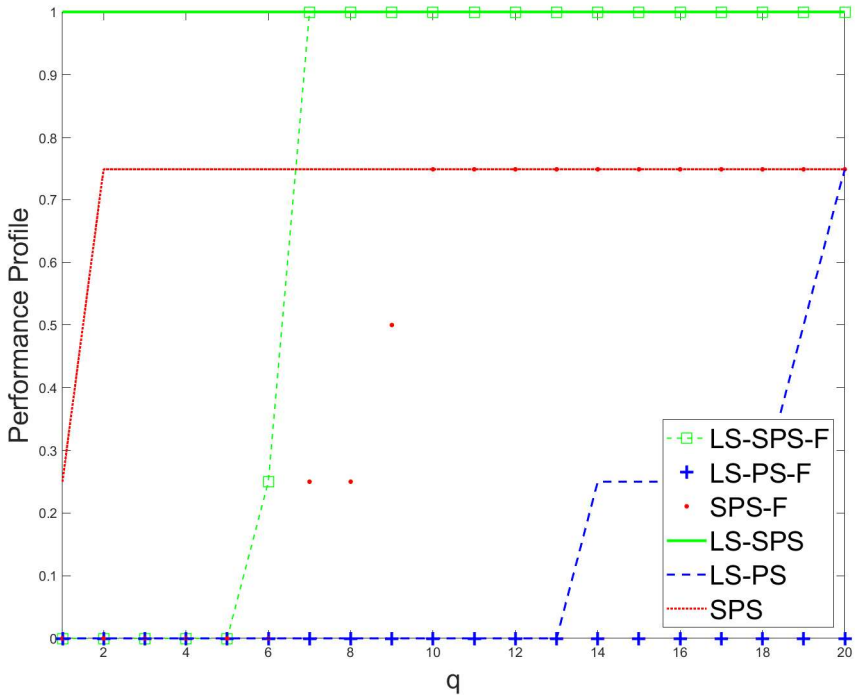

Figure 5.3: Performance profile for level of accuracy $\tau = 1$

plotted against the iteration k and the y-axes are in logarithmic scale. The results show that the proposed algorithms (LS-SPS-F and LS-SPS) outperform the observed PBM counterparts.

PBM	1	2	3	4	5
γ	1	1	1	1	1
m	0.01	0.1	0.01	0.01	0.01
ϵ	0.1	0.1	0.01	0.1	0.1
ω	0.5	0.5	0.5	0.9	0.1

Table 5.2: The initial parameters for proximal bundle method.

5.2 Adaptive Sample Size Nonmonotone Line Search Spectral Projected Subgradient Method

In order to solve problem (4.1), where the function f is continuous and bounded from below on Ω, we are going to present now AN-SPS - Adaptive sample size Nonmonotone line search Spectral Projected Subgradient method. It assumes subgradient directions, not necessarily descent, which may be combined with spectral coefficients. Both subgradients and spectral coefficients are calculated by employing SAA functions that vary across the iterations in general. In this setup, as it is mentioned in the previous section, it is needed to have a safeguard for the spectral coefficients to make sure that the resulting coefficients are positive and bounded. We also allow different nonmonotone line search rules described in Chapter 2, although the method is suitable for the monotone rule as well. The step size follows the idea of the SPS framework - line search over predefined intervals.

One of the key points lies in the adaptive sample size strategy. Roughly speaking, the main idea is to balance two types of errors - the one that measures how far is the iteration from the current SAA function's constrained optimum and the one that estimates the SAA

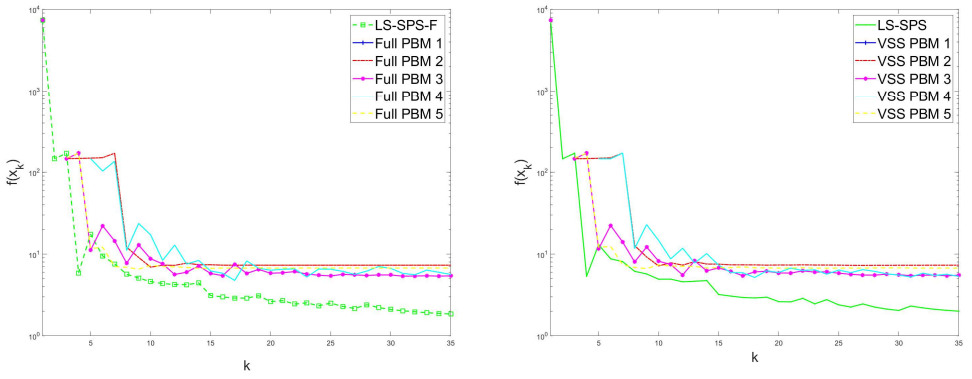

Figure 5.4: LS-SPS-F against Full PBM (left) and LS-SPS against VSS PBM (right). MNIST data set.

error. More precisely, we present an adaptive strategy that determines when to switch to the next level of accuracy and prove that this strategy pushes the sample size to infinity (or to the full sample size in a finite-sum case). In the SPS framework, the convergence result was proved under the assumption of the sample size increase at each iteration, while in AN-SPS this is a consequence of the algorithm's construction rather than the assumption.

We believe that one more important advantage with respect to SPS is the proposed scaling of the subgradient direction. The scaling strategy is not new in general [18], but it is a novelty within the SPS framework. One of the most important consequences of this modification is that the convergence result is proved without boundedness assumptions - we do not impose any assumption on uniformly bounded subgradients, feasible set, nor the iterations. Instead, we prove that AN-SPS generates the bounded sequence of iterates under a mild sample size growth condition. However, since feasible set Ω does not have

to be compact in general and thus the boundedness is not enough for claiming the solution exists, the Assumption 1 is modified and replaced with Assumption A3 in the sequel. More precisely, we assume that there exists a solution of $\min_{x\in\Omega} f_{\mathcal{N}}(x)$ for any given $\mathcal{N}$ instead of just assuming that the relevant functions are bounded from below.

The main result - almost sure convergence of the whole sequence of iterates - is proved under rather standard conditions for stochastic analysis. Moreover, in the finite-sum problem case, the convergence is deterministic, and it is proved under a significantly reduced set of assumptions with respect to the general case (4.1). Furthermore, we proved that the worst-case complexity can achieve the order of ε^{-1}. Although the worst-case complexity result stated in Theorem 5.2.4 is comparable to the complexity of standard subgradient methods with a predefined step size sequence and its stochastic variant (both of order ε^{-2}, see [14, 81] for instance), we believe that the advantage of the proposed method lies in its ability to accept larger steps and employ spectral coefficients combined with a nonmonotone line search, which can significantly speed up the method. Furthermore, the proposed method provides a wide framework for improving computational cost complexity since it allows different sampling strategies to be employed.

Numerical tests on Hinge loss problems and common data sets for machine learning show the advantages of the proposed adaptive VSS strategy. We also present the results of a study that investigates how different spectral coefficients combine with different nonmonotone rules.

Our contributions are the following. This part of the thesis may be seen as a continuation of the work presented in Section 5.1 and further development of algorithm LS-SPS. In this light, the main contributions are the following. In Section 5.2.1 an adaptive sample size strategy is proposed and in Section 5.2.2 we prove that this strategy pushes the sample size to infinity (or to the maximal sample size for finite-sum case). We show that the scaling can relax the boundedness

assumptions on subgradients, iterations, and feasible set. For finite-sum problems, we provide the worst-case complexity analysis of the proposed method. The LS-SPS is generalized in the sense that we allow different nonmonotone line search rules. Although important for the practical behavior of the algorithm, this change does not affect the convergence analysis and it is investigated mainly through numerical experiments presented in Section 5.2.3. Considering the spectral coefficient, we investigate different strategies for its formulation (previously defined rules within Section 2.2.1 - BB1, BB2, ABB, and ABBmin) in a stochastic framework. Moreover, different combinations of spectral coefficient and nonmonotone rules defined in Section 2.3 are evaluated within numerical experiments conducted on machine learning Hinge loss problems.

5.2.1 The Algorithm

In this section, we state the proposed AN-SPS framework algorithm. Recall that the sample used to approximate the objective function via (4.2) at iteration k is denoted by $\mathcal{N}_k$, while N_k denotes its cardinality. We use the SAA error measure $h(N_k)$ defined in Chapter 4 with (4.8), $h(N_k) = 1/(N_k)$, for unbounded sample size case and with (4.7), $h(N_k) = (N_{max} - N_k)/(N_{max})$, for the case where the full sample size is N_{max}. Other choices are eligible as well, but we keep these ones for simplicity. Furthermore, we define as in Algorithm 5 the upper bound of predefined interval for the line search by

$$\bar{\alpha}_k = \min\{1, C_2/k\},$$

where $C_2 > 0$ can be arbitrarily large.

First, notice that the initialization and Step S3 ensure the feasibility of the iterations. In Step S1, we choose an arbitrary subgradient of the current approximation function $f_{\mathcal{N}_k}$ at point x_k. Further, scaling

Algorithm 6: AN-SPS (**A**daptive Sample Size **N**onmonotone Line Search **S**pectral **P**rojected **S**ubgradient Method)

S0 *Initialization.* Given $N_0, m \in \mathbb{N}$, $x_0 \in \Omega$, $1 < C_2 < \infty$, $0 < \underline{\zeta} \leq \overline{\zeta} < \infty$, $\zeta_0 \in \left[\underline{\zeta}, \overline{\zeta}\right]$.
Set $k = 0$ and $F_0 = f_{\mathcal{N}_0}(x_0)$.

S1 *Search direction.* Choose $\bar{g}_k \in \partial f_{\mathcal{N}_k}(x_k)$.
Set $q_k = \max\{1, \|\bar{g}_k\|\}$, $v_k = \bar{g}_k / q_k$ and $p_k = -\zeta_k v_k$.

S2 *Step size.*

If $k = 0$, set $\alpha_0 = 1$.
Else, choose m points $\{\tilde{\alpha}_k^1, ..., \tilde{\alpha}_k^m\}$ such that

$$\frac{1}{k} < \tilde{\alpha}_k^1 < \tilde{\alpha}_k^2 < \ldots < \tilde{\alpha}_k^m = \bar{\alpha}_k.$$

If the condition

$$f_{\mathcal{N}_k}(x_k + \tilde{\alpha}_k^j p_k) \leq F_k - \eta \tilde{\alpha}_k^j ||p_k||^2 \tag{5.13}$$

is satisfied for some $j \in \{m, m-1, \ldots, 1\}$, set $\alpha_k = \tilde{\alpha}_k^j$ with the largest possible j.
Else, set $\alpha_k = \frac{1}{k}$.

S3 *Main update.* Set $z_{k+1} = x_k + \alpha_k p_k$, $x_{k+1} = P_\Omega(z_{k+1})$, $s_k = x_{k+1} - x_k$ and $\theta_k = \|s_k\|$.

S4 *Spectral coefficient update.* Choose $\zeta_{k+1} \in [\underline{\zeta}, \bar{\zeta}]$.

S5 *Sample size update.*
If $\theta_k < h(N_k)$, choose $N_{k+1} > N_k$ and a new sample $\mathcal{N}_{k+1}$.
Else, $\mathcal{N}_{k+1} = \mathcal{N}_k$.

S6 *Nonmonotone line search update.* Determine F_{k+1} such that

$$f_{\mathcal{N}_{k+1}}(x_{k+1}) \leq F_{k+1} < \infty.$$

S7 *Iteration update.* Set $k := k + 1$ and go to **S1**.

with q_k implies that $\|v_k\| \leq 1$. Moreover, the boundedness of the spectral coefficient ζ_k yields uniformly bounded search directions p_k. This is very important from the theoretical point of view since it helps us to overcome the boundedness assumptions mentioned at the beginning of this chapter.

For the step size selection, we practically use a backtracking-type procedure over the predefined interval $(\frac{1}{k}, \bar{\alpha}_k]$, which represents the generalization of Step S2 of Algorithm 5. Notice that C_2 can be arbitrarily large so that in practice $\bar{\alpha}_k$ is equal to 1 in most of the iterations. However, the upper bound C_2/k is needed to ensure theoretical convergence results. The lower bound, $1/k$, is known as a good choice from the theoretical point of view, and often a bad choice in practice. Thus, roughly speaking, the line search checks if larger, but still theoretically sound steps are eligible. Since the Armijo-like condition (5.13) is checked in at most m points, the procedure is well defined since if none of these candidate points satisfies condition (5.13), the step size is set to $1/k$. This allows us to use nondescent directions and practically arbitrary nonmonotone (or monotone) rule determined by the choice of F_k (see Section 2.3). For instance, F_k can be set to $f_{\mathcal{N}_k}(x_k) + 0.5^k$, but various other choices are possible as well. The choice of nonmonotone rule does not affect the theoretical convergence of the algorithm, but it can be very important in practice. Parameter m influences the per-iteration cost of the algorithm since it upper bounds the number of the function $f_{\mathcal{N}_k}$ evaluations within one line search procedure, i.e., within one iteration. Having in mind that the function $f_{\mathcal{N}_k}$ is just an estimate of the objective function in general, we suggest that m should be relatively small in order to avoid an unnecessarily precise line search and high computational costs. On the other hand, having m too small may yield smaller step sizes since $1/k$ is more likely to be accepted in general. Numerical results presented in Section 5.2.3 are obtained by taking $m = 2$ in all conducted experiments. However, tuning this parameter or even making it adaptive may be an interesting topic to

investigate.

We will test the performance of some choices for the spectral coefficients, where from a theoretical point of view the only requirement is the safeguard stated in the Step S4 of the algorithm - ζ_k must remain within positive, bounded interval $[\underline{\zeta}, \bar{\zeta}]$.

Finally, the adaptive sample size strategy is determined within Step S5. The overall step length θ_k may be considered as a measure of stationarity related to the current objective function approximation $f_{\mathcal{N}_k}$. In particular, we will show that, if the sample size is fixed, θ_k tends to zero and the sequence of iterates is approaching a minimizer of the current SAA function (see the proof of Theorem 5.2.1 in the sequel). When θ_k is relatively small (smaller than the measure of SAA error $h(N_k)$), we decide that the two errors are in balance and that we should improve the level of accuracy by enlarging the sample. Notice that Step S5 allows a completely different sample $\mathcal{N}_{k+1}$ in general with respect to $\mathcal{N}_k$ in the case where the sample size is increased. However, if the sample size is unchanged, the sample is unchanged, i.e., $\mathcal{N}_{k+1} = \mathcal{N}_k$, which allows non-cumulative samples to fit within the proposed framework as well.

AN-SPS algorithm detects the iteration within which the sample size needs to be increased, but it allows full freedom in the choice of the subsequent sample size as long as it is larger than the current one. After some preliminary tests, we end up with the following selection: when the sample size is increased, it is done as

$$N_{k+1} = \lceil \max\{(1 + \theta_k)N_k, rN_k\} \rceil, \tag{5.14}$$

with $r = 1.1$. Although some other choices such as direct balancing of θ_k and $h(N_{k+1})$ seemed more intuitive, they were all outperformed by the choice (5.14). Disregarding the safeguard part where, in case of $\theta_k = 0$, the sample size is increased by 10%, the relation becomes

$$1 + \frac{N_{k+1} - N_k}{N_k} \approx 1 + \theta_k.$$

Thus, the relative increase of the sample size is balanced with the stationarity measure. Furthermore, since we know that in these iterations $\theta_k < h(N_k)$, we obtain that the relative increase is bounded above by $h(N_k)$. Apparently, this helps the algorithm to overcome the problems caused by the non-beneficiary fast growth of the sample size.

5.2.2 Convergence Analysis

This subsection is devoted to the convergence analysis of the proposed AN-SPS method. One of the most important results lies in Theorem 5.2.1, where we prove that $h(N_k)$ tends to zero. More precisely, in the unbounded sample, this points to the sample size tending to infinity, while in the case of finite-sum, it means that the full sample is eventually reached. After that, it is shown that we can relax the common assumption of uniformly bounded subgradients stated in the convergence analysis in Section 5.1.2. Normalized subgradients have been used in the literature, but they represent a novelty with respect to the SPS framework. Therefore, we need to show that this type of scaling does not deteriorate the relevant convergence results.

We state the boundedness of iterations within Proposition 5.2.1. Although the convergence result stated in Theorem 5.2.2 mainly follows from the analysis of SPS (see Theorem 5.1.1 in Section 5.1.2), we provide the proof since it is based on different reasoning. Therefore, we show that AN-SPS retains almost sure convergence under relaxed assumptions with respect to LS-SPS, while, on the other hand, it brings more freedom to the choice of nonmonotone line search and the spectral coefficient. Finally, we formalize the conditions needed for the convergence in the finite-sum case within Theorem 5.2.3 and provide the worst-case complexity analysis.

We start the analysis by stating the conditions on the function under the expectation in problem (4.1).

Assumption A 3 *Function $\tilde{f}(\cdot, \xi)$ is continuous and convex on Ω for any given ξ and there exists a solution $x^*_{\mathcal{N}}$ of problem $\min_{x\in\Omega} f_{\mathcal{N}}(x)$ for any given $\mathcal{N}$.*

The previous assumption also implies that all the sample functions $f_{\mathcal{N}_k}$ are also convex and continuous on Ω. Recall that in Assumption A2 we assumed that the function F is dominated by a P-integrable function on any compact subset of $\mathbb{R}^n$. We state the first main result below. The SAA error measure defined in Chapter 4 is denoted with $h(N_k)$.

Theorem 5.2.1 *Suppose that Assumption A3 holds and that Ω is closed and convex. Then the sequence $\{N_k\}_{k\in\mathbb{N}}$ generated by AN-SPS satisfies*

$$\lim_{k\to\infty} h(N_k) = 0. \tag{5.15}$$

Proof. First, we show that retaining the same sample pushes θ_k to zero. Assume that $N_k = N$ for all $k \geq \tilde{k}$ and some $N < \infty$, $\tilde{k} \in \mathbb{N}$. According to Step S5 of AN-SPS algorithm, this means that $\mathcal{N}_k = \mathcal{N}_{\tilde{k}} := \mathcal{N}$ for all $k \geq \tilde{k}$. Let us show that this implies boundedness of $\{x_k\}_{k\in\mathbb{N}}$. Notice that for all k the step size and the search direction are bounded, more precisely, $\alpha_k \leq \bar{\alpha}_k \leq 1$ and

$$\|p_k\| = \|\zeta_k v_k\| \leq \bar{\zeta}\|v_k\| \leq \bar{\zeta}.$$

Thus, the $\tilde{k}$ initial iterations must be bounded, i.e., there must exist $C_{\tilde{k}}$ such that $\|x_k\| \leq C_{\tilde{k}}$ for all $k = 0, 1, ..., \tilde{k}$. Now, let us observe the remaining sequence of iterates, i.e., $\{x_{\tilde{k}+j}\}_{j\in\mathbb{N}}$. Let $x^*_{\mathcal{N}}$ be an arbitrary solution of the problem $\min_{x\in\Omega} f_{\mathcal{N}}(x)$. Notice that the convexity of $f_{\mathcal{N}}$ and the fact that $\bar{g}_k \in \partial f_{\mathcal{N}}(x_k)$ for all $k \geq \tilde{k}$ and $x \in \mathbb{R}^n$ imply

$$-\overline{g}_k^T (x_k - x) \leq f_{\mathcal{N}}(x) - f_{\mathcal{N}}(x_k).$$

Then, by using the non-expansivity of the projection operator and the fact that $x^*_{\mathcal{N}} \in \Omega$, for all $k \geq \tilde{k}$ we obtain

$$\begin{aligned} ||x_{k+1} - x^*_{\mathcal{N}}||^2 &= ||P_\Omega(z_{k+1}) - P_\Omega(x^*_{\mathcal{N}})||^2 \\ &\leq ||z_{k+1} - x^*_{\mathcal{N}}||^2 = ||x_k - \alpha_k \zeta_k v_k - x^*_{\mathcal{N}}||^2 \\ &= ||x_k - x^*_{\mathcal{N}}||^2 - 2\alpha_k \zeta_k \frac{1}{q_k} \overline{g}_k^T (x_k - x^*_{\mathcal{N}}) + \alpha_k^2 \zeta_k^2 ||v_k||^2 \\ &\leq ||x_k - x^*_{\mathcal{N}}||^2 + 2\alpha_k \frac{\zeta_k}{q_k} (f_{\mathcal{N}_k}(x^*_{\mathcal{N}}) - f_{\mathcal{N}_k}(x_k)) + \alpha_k^2 \bar{\zeta}^2 \\ &\leq ||x_k - x^*_{\mathcal{N}}||^2 + \alpha_k^2 \bar{\zeta}^2. \end{aligned} \tag{5.16}$$

In the last inequality, we use the fact that $\mathcal{N}_k = \mathcal{N}$ for all $k \geq \tilde{k}$. Thus,

$$f_{\mathcal{N}_k}(x^*_{\mathcal{N}}) - f_{\mathcal{N}_k}(x_k) = f_{\mathcal{N}}(x^*_{\mathcal{N}}) - f_{\mathcal{N}}(x_k) \leq 0$$

and since $\alpha_k \zeta_k / q_k > 0$ we obtain the result. Furthermore, by using the induction argument, we obtain that for every $p \in \mathbb{N}$ there holds

$$\begin{aligned} ||x_{\tilde{k}+p} - x^*_{\mathcal{N}}||^2 &\leq ||x_{\tilde{k}} - x^*_{\mathcal{N}}||^2 + \bar{\zeta}^2 \sum_{j=0}^{p-1} \alpha^2_{\tilde{k}+j} \leq ||x_{\tilde{k}} - x^*_{\mathcal{N}}||^2 + \bar{\zeta}^2 \sum_{j=0}^{\infty} \alpha_j^2 \\ &\leq ||x_{\tilde{k}} - x^*_{\mathcal{N}}||^2 + \bar{\zeta}^2 C_2^2 \sum_{j=0}^{\infty} \frac{1}{k^2} := \bar{C}_{\tilde{k}} < \infty. \end{aligned}$$

Thus, we conclude that the sequence of iterates must be bounded, i.e., there exists a compact set $\bar{\Omega} \subseteq \Omega$ such that $\{x_k\}_{k \in \mathbb{N}} \subseteq \bar{\Omega}$. Since the function $f_{\mathcal{N}}$ is convex due to Assumption A3, there follows that $f_{\mathcal{N}}$ is locally Lipschitz continuous. Moreover, it is (globally) Lipschitz continuous on the compact set $\bar{\Omega}$. Let us denote the corresponding Lipschitz constant by $L_{\bar{\Omega}}$. Then, we know that $\|g\| \leq L_{\bar{\Omega}}$ holds for any $g \in \partial f_{\mathcal{N}}(x)$ and any $x \in \bar{\Omega}$.

Now, we prove that

$$\liminf_{k \to \infty} f_{\mathcal{N}}(x_k) = f^*_{\mathcal{N}}, \tag{5.17}$$

where $f_{\mathcal{N}}^* = \min_{x\in\Omega} f_{\mathcal{N}}(x)$. Suppose the contrary, i.e., there exists $\varepsilon_{\mathcal{N}} > 0$ such that for all $k \geq \tilde{k}$ there holds $f_{\mathcal{N}}(x_k) - f_{\mathcal{N}}^* \geq 2\varepsilon_{\mathcal{N}}$. Recall that Assumption A3 implies that $f_{\mathcal{N}}^*$ is finite and that $f_{\mathcal{N}}$ is continuous. Therefore, there exists a sequence $\{y_j^{\mathcal{N}}\}_{j\in\mathbb{N}} \in \Omega$ such that $\lim_{j\to\infty} f_{\mathcal{N}}(y_j^{\mathcal{N}}) = f_{\mathcal{N}}^*$. Moreover, there exists a point $\tilde{y}_{\mathcal{N}} \in \Omega$ such that

$$f_{\mathcal{N}}(\tilde{y}_{\mathcal{N}}) < f_{\mathcal{N}}^* + \varepsilon_{\mathcal{N}}.$$

Therefore, we conclude that for all $k \geq \tilde{k}$ there holds

$$f_{\mathcal{N}}(x_k) \geq f_{\mathcal{N}}^* + 2\varepsilon_{\mathcal{N}} = f_{\mathcal{N}}^* + \varepsilon_{\mathcal{N}} + \varepsilon_{\mathcal{N}} > f_{\mathcal{N}}(\tilde{y}_{\mathcal{N}}) + \varepsilon_{\mathcal{N}},$$

and thus for all $k \geq \tilde{k}$ we have

$$-\overline{g}_k^T (x_k - \tilde{y}_{\mathcal{N}}) \leq f_{\mathcal{N}}(\tilde{y}_{\mathcal{N}}) - f_{\mathcal{N}}(x_k) \leq -\varepsilon_{\mathcal{N}}.$$

Following the same steps as in (5.16) and using the previous inequality, we conclude that for all $k \geq \tilde{k}$ there holds

$$\begin{aligned}
||x_{k+1} - \tilde{y}_{\mathcal{N}}||^2 &\leq ||z_{k+1} - \tilde{y}_{\mathcal{N}}||^2 \\
&\leq ||x_k - \tilde{y}_{\mathcal{N}}||^2 - 2\alpha_k\zeta_k\frac{1}{q_k}\overline{g}_k^T (x_k - \tilde{y}_{\mathcal{N}}) + \alpha_k^2\zeta_k^2||v_k||^2 \\
&\leq ||x_k - \tilde{y}_{\mathcal{N}}||^2 - 2\alpha_k\frac{\zeta_k}{q_k}\varepsilon_{\mathcal{N}} + \alpha_k^2\bar{\zeta}^2 \\
&\leq ||x_k - \tilde{y}_{\mathcal{N}}||^2 - 2\alpha_k\frac{1}{q_k}\underline{\zeta}\varepsilon_{\mathcal{N}} + \alpha_k^2\bar{\zeta}^2.
\end{aligned}$$

Now, using the fact that

$$q_k = \max\{1, \|\bar{g}_k\|\} \leq \max\{1, L_{\bar{\Omega}}\} := q, \tag{5.18}$$

we conclude that for all $k \geq \tilde{k}$ there holds

$$\begin{aligned}
||x_{k+1} - \tilde{y}_{\mathcal{N}}||^2 &\leq ||x_k - \tilde{y}_{\mathcal{N}}||^2 - 2\alpha_k\frac{1}{q}\underline{\zeta}\varepsilon_{\mathcal{N}} + \alpha_k^2\bar{\zeta}^2 \\
&= ||x_k - \tilde{y}_{\mathcal{N}}||^2 - \alpha_k(\frac{2}{q}\underline{\zeta}\varepsilon_{\mathcal{N}} - \alpha_k\bar{\zeta}^2).
\end{aligned}$$

Since $\alpha_k \leq C_2/k$, there holds $\lim_{k\to\infty} \alpha_k = 0$ and thus there must exist $\bar{k} \geq \tilde{k}$ such that $\alpha_k \bar{\zeta}^2 \leq \frac{1}{q}\underline{\zeta}\varepsilon_{\mathcal{N}} := \underline{\varepsilon}_{\mathcal{N}}$.
Therefore, we have

$$||x_{k+1} - \tilde{y}_{\mathcal{N}}||^2 \leq ||x_k - \tilde{y}_{\mathcal{N}}||^2 - \alpha_k \underline{\varepsilon}_{\mathcal{N}}.$$

Moreover, for any $p \in \mathbb{N}$ there holds

$$||x_{\bar{k}+p} - \tilde{y}_{\mathcal{N}}||^2 \leq ||x_{\bar{k}} - \tilde{y}_{\mathcal{N}}||^2 - \underline{\varepsilon}_{\mathcal{N}} \sum_{j=0}^{p-1} \alpha_{\bar{k}+j}$$

and letting $p \to \infty$ we obtain the contradiction since

$$\sum_{k=0}^{\infty} \alpha_k \geq \sum_{k=0}^{\infty} 1/k = \infty.$$

Thus, we conclude that (5.17) must hold. Therefore there exists $K_1 \subseteq \mathbb{N}$ such that

$$\lim_{k\in K_1} f_{\mathcal{N}}(x_k) = f^*_{\mathcal{N}}$$

and since the iterations are bounded, there exists $K_2 \subseteq K_1$ and a solution $\tilde{x}^*_{\mathcal{N}}$ of the problem $\min_{x\in\Omega} f_{\mathcal{N}}(x)$ such that

$$\lim_{k\in K_2} x_k = \tilde{x}^*_{\mathcal{N}}. \tag{5.19}$$

Now, we show that the whole sequence of iterates converges. Let

$$\{x_k\}_{k\in K_2} := \{x_{k_i}\}_{i\in\mathbb{N}}.$$

Following the steps of (5.16) we obtain that the following holds for any $s \in \mathbb{N}$

$$||x_{k_i+s} - \tilde{x}^*_{\mathcal{N}}||^2 \leq ||x_{k_i} - \tilde{x}^*_{\mathcal{N}}||^2 + \bar{\zeta}^2 \sum_{j=0}^{s-1} \alpha^2_{k_i+j} \leq ||x_{k_i} - \tilde{x}^*_{\mathcal{N}}||^2 + \bar{\zeta}^2 \sum_{j=k_i}^{\infty} \alpha_j^2 =: b_i.$$

Thus, for any $s, m \in \mathbb{N}$ there holds

$$||x_{k_i+s} - x_{k_i+m}||^2 \leq 2||x_{k_i+s} - \tilde{x}^*_{\mathcal{N}}||^2 + 2||x_{k_i+m} - \tilde{x}^*_{\mathcal{N}}||^2 \leq 4b_i.$$

Since $\sum_{j=k_i}^{\infty} \alpha_j^2$ is a residual of convergent sum and (5.19) holds, we have

$$\lim_{i\to\infty} b_i = 0.$$

Therefore, for every $\varepsilon > 0$ there exists $k_i \in \mathbb{N}$ such that for all $t, l \geq k_i$ there holds $||x_t - x_l|| \leq \varepsilon$, i.e., the sequence $\{x_k\}_{k\in\mathbb{N}}$ is a Cauchy sequence and thus convergent. This, together with (5.19), implies

$$\lim_{k\to\infty} x_k = \tilde{x}^*_{\mathcal{N}},$$

and the Step S3 of AN-SPS algorithm implies

$$\lim_{k\to\infty} \theta_k = \lim_{k\to\infty} \|s_k\| = \lim_{k\to\infty} \|x_{k+1} - x_k\| = 0.$$

This completes the first part of the proof, i.e., we have just proved that if the sample is kept fixed, the sequence $\{\theta_k\}_{k\in\mathbb{N}}$ tends to zero.

Finally, we prove the main result (5.15), i.e., $\lim_{k\to\infty} h(N_k) = 0$. Assume the contrary. Since the sequence $\{h(N_k)\}_{k\in\mathbb{N}}$ is nonincreasing, this means that we can assume there exists $\bar{h} > 0$ such that

$$h(N_k) \geq \bar{h} \text{ for all } k \in \mathbb{N}.$$

This further implies that there exist $N < N_\infty$ and $\bar{k} \in \mathbb{N}$ such that

$$N_k = N \text{ for all } k \geq \bar{k},$$

where $N_\infty = \infty$ in unbounded sample case and N_∞ coincides with the full sample size in bounded sample (finite-sum) case. Thus, according to S5 of AN-SPS algorithm, there holds that

$$\theta_k \geq h(N_k) = h(N) \geq \bar{h} > 0$$

for all $k \geq \bar{k}$, since we would have an increase of the sample size N otherwise. On the other hand, we have just proved that if the sample size is fixed, then $\lim_{k\to\infty} \theta_k = 0$, which is in contradiction with $\theta_k \geq \bar{h} > 0$. Thus, we conclude that

$$\lim_{k\to\infty} h(N_k) = 0,$$

which completes the proof. ■

Next, we analyze the conditions that provide a sequence of bounded iterates generated by AN-SPS algorithm. Recall that the SAA error sequence $\bar{e}_k$ is defined as before (5.9),

$$\bar{e}_k = |f_{\mathcal{N}_k}(x_k) - f(x_k)| + |f_{\mathcal{N}_k}(x^*) - f(x^*)|.$$

The proof of the following proposition is similar to the proof of Proposition 5.1.1 in Section 5.1.2, but the conditions are relaxed since we have $N_k \to \infty$ as a consequence of the Theorem 5.2.1. Moreover, scaling of the subgradients relaxes the assumption of uniformly bounded $\bar{g}_k$ sequence. Although the modifications are mainly technical, we provide proof for the sake of completeness. Condition (5.20) in the sequel states the sample size growth under which we achieve bounded iterations. For instance, in the cumulative sample case, $N_k = k$ is sufficient to ensure this condition (see Section 4.2.1). Although we believe that the condition $\sum_{k=0}^{\infty} \bar{e}_k/k \leq C_4 < \infty$ is not too strong, it is still an assumption and not the consequence of the algorithm, so this issue remains as an open question for the future work.

Proposition 5.2.1 *Suppose that Ω is closed and convex, Assumption A3 holds and $\{x_k\}_{k\in\mathbb{N}}$ is a sequence generated by Algorithm AN-SPS. Then if*

$$\sum_{k=0}^{\infty} \bar{e}_k/k \leq C_4 < \infty, \tag{5.20}$$

there exists a compact set $\bar{\Omega} \subseteq \Omega$ such that $\{x_k\} \subseteq \bar{\Omega}$.

Proof. Let x^* be an arbitrary solution of the problem (4.1). Following the steps of (5.16) and the definition (5.9) we obtain for all $k = 0, 1, ..$

$$\begin{aligned}
||x_{k+1} - x^*||^2 &= ||P_\Omega(z_{k+1}) - P_\Omega(x^*)||^2 \\
&\leq ||x_k - x^*||^2 + 2\alpha_k \frac{\zeta_k}{q_k}(f_{\mathcal{N}_k}(x^*) - f_{\mathcal{N}_k}(x_k)) + \alpha_k^2 \bar{\zeta}^2 \\
&\leq ||x_k - x^*||^2 + 2\alpha_k \frac{\zeta_k}{q_k}(f(x^*) - f(x_k) + \bar{e}_k) + \alpha_k^2 \bar{\zeta}^2 \\
&\leq ||x_k - x^*||^2 + 2\alpha_k \frac{\zeta_k}{q_k}(f(x^*) - f(x_k)) + 2\alpha_k \frac{\zeta_k}{q_k}\bar{e}_k + \alpha_k^2 \bar{\zeta}^2 \\
&\leq ||x_k - x^*||^2 + 2\alpha_k \bar{\zeta}\bar{e}_k + \alpha_k^2 \bar{\zeta}^2,
\end{aligned}$$

where we use the fact that x_k is feasible and thus $f(x^*) - f(x_k) \leq 0$ and that $q_k \geq 1$. Further, by the induction argument and the fact that $\alpha_k \leq C_2/k$ we obtain

$$||x_k - x^*||^2 \leq ||x_0 - x^*||^2 + 2C_2\bar{\zeta}\sum_{k=0}^{\infty}\frac{\bar{e}_k}{k} + \bar{\zeta}^2\sum_{k=0}^{\infty}\frac{C_2^2}{k^2} \leq C_5 < \infty.$$

This completes the proof. ■

As it can be seen from the proof, $\bar{\Omega}$ stated in the previous proposition depends only on x_0 and given constants, so it can be (theoretically) determined independently of the sample path. However, since we consider unbounded sample in general, we need the following assumption.

Assumption A 4 *For every $x \in \Omega$ there exists a constant L_x such that $F(x, \xi)$ is locally Lipschitz-L_x continuous for any ξ.*

This assumption implies that each SAA function is locally Lipschitz continuous with a constant that depends only on a point x and not

on a random vector ξ. In the bounded sample case this is obviously satisfied under Assumption A3, while in general, it holds for a certain class of functions - when ξ is separable from x for instance. Next, we prove the almost sure convergence of the AN-SPS algorithm under the stated assumptions. Notice that (5.20) does not necessarily imply $\lim_{k\to\infty} \bar{e}_k = 0$. Thus, we need Assumption A2, which is a common assumption in stochastic analysis, in order to ensure a.s. convergence of the sequence $\{\bar{e}_k\}_{k\in\mathbb{N}}$. Under the stated assumptions, ULLN implies (4.6), i.e., $\lim_{N\to\infty} \sup_{x\in S} |f_N(x) - f(x)| = 0$ a.s. for any compact subset $S \subseteq \mathbb{R}^n$. This will further imply the a.s. convergence of the sequence $\bar{e}_k$. Notice that $\lim_{k\to\infty} \bar{e}_k = 0$ is satisfied in the bounded sample case, as well as (5.20) since AN-SPS achieves the full sample eventually. In that case, the Assumptions A2 and A4 are not needed for the convergence result.

Remark: The following theorem states a.s. convergence of the proposed method. Although it follows the same steps, the proof differs from the proof of Theorem 5.1.1 in several places. Under the stated assumptions we prove that the sample size tends to infinity and that the iterations remain within a compact set. After that, the proof follows the steps of the proof in Theorem 5.1.1 completely, except for the scaling of the subgradient in Step S1 of the AN-SPS algorithm. This alters the inequalities, but Assumption A4 implies that q_k can be uniformly bounded from above and below, as it will be shown in the first part of the proof of the following theorem, thus the main flow remains the same. We state the proof for completeness.

Theorem 5.2.2 *Suppose that Assumptions A2-A4 and* (5.20) *hold and that* Ω *is closed and convex. Then the sequence* $\{x_k\}_{k\in\mathbb{N}}$ *generated by AN-SPS converges to a solution of problem* (4.1) *almost surely.*

Proof. First, notice that Theorem 5.2.1 implies that

$$\lim_{k\to\infty} N_k = \infty$$

in ubounded sample case. Moreover, Proposition 5.2.1 implies that $\{x_k\} \subseteq \bar{\Omega}$. Furthermore, Assumption A4 implies that for any $\mathcal{N}$ we have locally Lipschitz-L_x continuous function $f_{\mathcal{N}}(x)$. Thus, there exists a constant L such that $f_{\mathcal{N}}$ is Lipschitz-L continuous on $\bar{\Omega}$ for any $\mathcal{N}$. This further implies that $\|\bar{g}_k\| \leq L$ for each k and

$$1 \leq q_k \leq \max\{1, L\} := \bar{q}. \tag{5.21}$$

Denote by $\mathcal{W}$ the set of all possible sample paths of the AN-SPS algorithm. First we prove that

$$\liminf_{k\to\infty} f(x_k) = f^* \quad \text{a.s.} \tag{5.22}$$

where $f^* = \inf_{x\in\Omega} f(x)$. Suppose that $\liminf_{k\to\infty} f(x_k) = f^*$ does not happen with probability 1. In that case there exists a subset of sample paths $\tilde{\mathcal{W}} \subseteq \mathcal{W}$ such that $P(\tilde{\mathcal{W}}) > 0$ and for every $w \in \tilde{\mathcal{W}}$ there holds

$$\liminf_{k\to\infty} f(x_k(w)) > f^*,$$

i.e., there exists $\varepsilon(w) > 0$ small enough such that

$$f(x_k(w)) - f^* \geq 2\varepsilon(w)$$

for all k. Since f is assumed to be continuous and bounded from below on Ω, f^* is finite and we conclude that there exists a point $\tilde{y}(w) \in \Omega$ such that $f(\tilde{y}(w)) < f^* + \varepsilon(w)$. This further implies

$$f(x_k(w)) - f(\tilde{y}(w)) > f(x_k(w)) - f^* - \varepsilon(w) \geq 2\varepsilon(w) - \varepsilon(w) = \varepsilon(w).$$

Let us take an arbitrary $w \in \tilde{\mathcal{W}}$. Denote

$$z_{k+1}(w) := x_k(w) + \alpha_k(w)p_k(w).$$

Notice that non-expansivity of orthogonal projection and the fact that $\tilde{y} \in \Omega$ together imply

$$||x_{k+1}(w) - \tilde{y}(w)|| = ||P_\Omega(z_{k+1}(w)) - P_\Omega(\tilde{y}(w))|| \leq ||z_{k+1}(w) - \tilde{y}(w)||. \tag{5.23}$$

Furthermore, using (5.21) and the fact that $\bar{g}_k$ is subgradient of convex function $f_{\mathcal{N}_k}$, $\bar{g}_k \in \partial f_{\mathcal{N}_k}(x_k)$, we have

$$f_{\mathcal{N}_k}(x_k) - f_{\mathcal{N}_k}(\tilde{y}) \leq \overline{g}_k^T(x_k - \tilde{y})$$

and dropping w in order to facilitate the reading we obtain

$$\begin{aligned}
||z_{k+1} - \tilde{y}||^2 &= ||x_k + \alpha_k p_k - \tilde{y}||^2 = ||x_k - \alpha_k \zeta_k v_k - \tilde{y}||^2 \\
&= ||x_k - \tilde{y}||^2 - 2\alpha_k \zeta_k \frac{\overline{g}_k^T}{q_k}(x_k - \tilde{y}) + \alpha_k^2 \zeta_k^2 ||v_k||^2 \\
&\leq ||x_k - \tilde{y}||^2 + 2\alpha_k \frac{\zeta_k}{q_k}(f_{\mathcal{N}_k}(\tilde{y}) - f_{\mathcal{N}_k}(x_k)) + \alpha_k^2 \zeta_k^2 \\
&\leq ||x_k - \tilde{y}||^2 + 2\alpha_k \frac{\zeta_k}{q_k}(f(\tilde{y}) - f(x_k) + e_k^+) + \alpha_k^2 \zeta_k^2 \\
&\leq ||x_k - \tilde{y}||^2 - 2\alpha_k \frac{\zeta_k}{q_k}(f(x_k) - f(\tilde{y})) + 2e_k^+ \alpha_k \overline{\zeta} + \alpha_k^2 \overline{\zeta}^2 \\
&\leq ||x_k - \tilde{y}||^2 - 2\alpha_k \frac{\underline{\zeta}}{\bar{q}}\varepsilon + 2e_k^+ \alpha_k \overline{\zeta} + \alpha_k^2 \overline{\zeta}^2 \\
&= ||x_k - \tilde{y}||^2 - \alpha_k \left(2\frac{\underline{\zeta}}{\bar{q}}\varepsilon - 2e_k^+ \overline{\zeta} - \alpha_k \overline{\zeta}^2 \right),
\end{aligned} \tag{5.24}$$

where $e_k^+ = |f_{\mathcal{N}_k}(\tilde{y}) - f(\tilde{y})| + \max_{x \in \bar{\Omega}} |f_{\mathcal{N}_k}(x) - f(x)|$.

Since, $\{x_k\} \subseteq \bar{\Omega}$, ULLN under the stated assumptions implies

$$\lim_{k \to \infty} e_k^+(w) = 0$$

for almost every $w \in \mathcal{W}$. Since $P(\tilde{\mathcal{W}}) > 0$, there must exist a sample path $\tilde{w} \in \tilde{\mathcal{W}}$ such that

$$\lim_{k\to\infty} e_k^+(\tilde{w}) = 0.$$

This further implies the existence of $\tilde{k}(\tilde{w}) \in \mathbb{N}$ such that for all $k \geq \tilde{k}(\tilde{w})$ we have

$$\alpha_k(\tilde{w})\overline{\zeta}^2 + 4e_k^+(\tilde{w})\overline{\zeta} \leq \varepsilon(\tilde{w})\frac{\zeta}{\bar{q}} \tag{5.25}$$

because Step S2 of AN-SPS algorithm implies that $\lim_{k\to\infty} \alpha_k = 0$ for any sample path. Furthermore, since (5.24) holds for all $w \in \tilde{\mathcal{W}}$ and thus for $\tilde{w}$ as well, from (5.23)-(5.25) we obtain

$$\begin{aligned} ||x_{k+1}(\tilde{w}) - \tilde{y}(\tilde{w})||^2 &\leq ||z_{k+1}(\tilde{w}) - \tilde{y}(\tilde{w})||^2 \\ &\leq ||x_k(\tilde{w}) - \tilde{y}(\tilde{w})||^2 - \alpha_k(\tilde{w})\varepsilon(\tilde{w})\frac{\zeta}{\bar{q}} \end{aligned}$$

and

$$||x_{k+s}(\tilde{w}) - \tilde{y}(\tilde{w})||^2 \leq ||x_k(\tilde{w}) - \tilde{y}(\tilde{w})||^2 - \varepsilon(\tilde{w})\frac{\zeta}{\bar{q}}\sum_{j=0}^{s-1}\alpha_j(\tilde{w}).$$

Letting $s \to \infty$ yields a contradiction since $\sum_{k=0}^{\infty}\alpha_k \geq \sum_{k=0}^{\infty} 1/k = \infty$ for any sample path and we conclude that (5.22) holds.

Now, let us prove that

$$\lim_{k\to\infty} x_k = x^* \quad \text{a.s.} \tag{5.26}$$

Notice that (5.20) implies that $\sum_{k=0}^{\infty}\alpha_k\bar{e}_k \leq \sum_{k=0}^{\infty}\frac{C_2}{k}\bar{e}_k \leq C_2C_4 < \infty$ since $\alpha_k \leq C_2/k$. Since (5.22) holds, we know that

$$\liminf_{k\to\infty} f(x_k(w)) = f^*, \tag{5.27}$$

for almost every $w \in \mathcal{W}$. In other words, there exists $\overline{\mathcal{W}} \subseteq \mathcal{W}$ such that $P(\overline{\mathcal{W}}) = 1$ and (5.27) holds for all $w \in \overline{\mathcal{W}}$. Let us consider arbitrary $w \in \overline{\mathcal{W}}$. We will show that $\lim_{k\to\infty} x_k(w) = x^*(w) \in X^*$ which will imply the result (5.26). Once again let us drop w to facilitate the notation. Let $K_1 \subseteq \mathbb{N}$ be a subsequence of iterations such that

$$\lim_{k\in K_1} f(x_k) = f^*.$$

Since $\{x_k\}_{k\in K_1} \subseteq \{x_k\}_{k\in\mathbb{N}}$ and $\{x_k\}_{k\in\mathbb{N}}$ is bounded, there exist $K_2 \subseteq K_1$ and $\tilde{x}$ such that

$$\lim_{k\in K_2} x_k = \tilde{x}. \tag{5.28}$$

Then, we have

$$f^* = \lim_{k\in K_1} f(x_k) = \lim_{k\in K_2} f(x_k) = f(\lim_{k\in K_2} x_k) = f(\tilde{x}).$$

Therefore, $f(\tilde{x}) = f^*$ and we have $\tilde{x} \in X^*$.

Now, we show that the whole sequence of iterates converges. Let $\{x_k\}_{k\in K_2} := \{x_{k_i}\}_{i\in\mathbb{N}}$. Following the steps of (5.24) and using the fact that $f(x_k) \geq f(\tilde{x})$ for all k, we obtain that the following holds for any $s \in \mathbb{N}$

$$\begin{aligned} ||x_{k_i+s} - \tilde{x}||^2 &\leq ||x_{k_i} - \tilde{x}||^2 + 2\overline{\zeta}\sum_{j=0}^{s-1} \bar{e}_{k_i+j}\alpha_{k_i+j} + \overline{\zeta}^2 \sum_{j=0}^{s-1} \alpha_{k_i+j}^2 \\ &\leq ||x_{k_i} - \tilde{x}||^2 + 2\overline{\zeta}\sum_{j=0}^{\infty} \bar{e}_{k_i+j}\alpha_{k_i+j} + \overline{\zeta}^2 \sum_{j=0}^{\infty} \alpha_{k_i+j}^2 \\ &= ||x_{k_i} - \tilde{x}||^2 + 2\overline{\zeta}\sum_{j=k_i}^{\infty} \bar{e}_j\alpha_j + \overline{\zeta}^2 \sum_{j=k_i}^{\infty} \alpha_j^2 =: a_i. \end{aligned} \tag{5.29}$$

Moreover, for any $s, m \in \mathbb{N}$ there holds

$$||x_{k_i+s} - x_{k_i+m}||^2 \leq 2||x_{k_i+s} - \tilde{x}||^2 + 2||x_{k_i+m} - \tilde{x}||^2 \leq 4a_i.$$

Due to the fact that $\sum_{j=k_i}^{\infty} \bar{e}_j \alpha_j$ and $\sum_{j=k_i}^{\infty} \alpha_j^2$ are the residuals of convergent sums, and that (5.28) holds, we conclude that

$$\lim_{i \to \infty} a_i = 0.$$

Thus, we have just proved that $\{x_k\}_{k \in \mathbb{N}}$ is a Cauchy sequence and thus convergent, which together with (5.28) implies that

$$\lim_{k \to \infty} x_k = \tilde{x}. \blacksquare$$

Finally, we state the result for finite-sum problem (4.4) as an important class of (4.1). As we mentioned before, Assumption A2 is redundant in this case as well as (5.20) since $\bar{e}_k = 0$ for all k large enough. Moreover, Assumption A4 is also satisfied due to the fact that there are only finitely many functions f_i. In the end, notice that under Assumption A3, the full sample is eventually achieved and the proof of Theorem 5.2.1 also reveals that the convergence is deterministic. We summarise this in the next theorem.

Theorem 5.2.3 *Suppose that Assumption A3 holds and that Ω is closed and convex. Then the sequence $\{x_k\}_{k \in \mathbb{N}}$ generated by AN-SPS converges to a solution of problem* (4.4).

We also provide the worst-case complexity analysis for the relevant finite-sum problem (4.4).

Theorem 5.2.4 *Suppose that the assumptions of Theorem 5.2.3 hold and that the sample size increases as in* (5.14). *Then, ε-vicinity of an optimal value f^* of problem* (4.4) *is reached after at most*

$$\hat{k} = 2\bar{k} + \left(\frac{q(\bar{c}_1 + ||x_{\bar{k}} - x^*||^2)}{\underline{\zeta}} \right)^{\frac{1}{1-\delta}} \varepsilon^{\frac{1}{\delta - 1}}$$

iterations, where

$$\bar{k} := (\lceil C_2\bar{\zeta}N\rceil + 1)\frac{\log(N/N_0)}{\log(r)}, \quad \bar{c}_1 := \sum_{k=0}^{\infty}\frac{C_2^2\bar{\zeta}^2}{k^2},$$

provided that $\alpha_k \geq k^{-\delta}, \delta \in [0,1)$ *for all* $k \in \{\bar{k}, \bar{k}+1, ..., \hat{k}\}$.

Proof. Let us denote by $N^1 < N^2 < ... < N^d$ all the sample sizes that are used during the optimization process. Then, we have that $N^1 = N_0$, where N_0 is the initial sample size, and $N^d = N$ since we have proved that the full sample is reached eventually. Furthermore, according to (5.14), we know that $N^d \geq r^{d-1}N_0$ and thus we conclude that

$$d - 1 \leq \frac{\log(N/N_0)}{\log(r)}.$$

Furthermore, notice that for any $k \in \mathbb{N}$ there holds

$$\theta_k = \|x_{k+1} - x_k\| = \|P_\Omega(z_{k+1}) - P_\Omega(x_k)\| \leq \|z_{k+1} - x_k\| = \|\alpha_k p_k\| \leq \frac{C_2}{k}\bar{\zeta}.$$

Suppose that we are at iteration k with a sample size $N_k = N^j$, with $j < d$. Then, according to Step S5 of Algorithm 6, the sample size N^j is changed after at most

$$\lceil \frac{C_2\bar{\zeta}}{h(N^j)} \rceil + 1$$

iterations. Moreover, since $N^j \leq N - 1$ for all $j = 1, ..., d-1$, there holds

$$h(N^j) \geq h(N-1) = \frac{N - (N-1)}{N} = \frac{1}{N}$$

for all $j = 1, ..., d-1$ and thus the number of iterations with the same sample size smaller than N is uniformly bounded by $\lceil C_2\bar{\zeta}N\rceil + 1$.

Thus, we conclude that after at most

$$\bar{k} := (\lceil C_2 \bar{\zeta} N \rceil + 1) \frac{\log(N/N_0)}{\log(r)}$$

iterations the full sample size is reached.

Now, let us observe the iterations $k \geq \bar{k}$. Theorem 5.2.3 implies that $\lim_{k\to\infty} f_{\mathcal{N}}(x_k) = f^*$ and thus there exists a finite iteration k such that $f_{\mathcal{N}}(x_k) < f^* + \varepsilon$. Let us denote by $\hat{j}$ the smallest $j \in \mathbb{N}_0$ such that $f_{\mathcal{N}}(x_{\bar{k}+\hat{j}}) < f(x^*) + \varepsilon$, where x^* is a solution of problem (4.4). Using the same arguments as in (5.16), we obtain

$$||x_{\bar{k}+\hat{j}} - x^*||^2 \leq ||x_{\bar{k}} - x^*||^2 - \sum_{j=0}^{\hat{j}-1} 2\alpha_{\bar{k}+j}\zeta_{\bar{k}+j} \frac{1}{q_{\bar{k}+j}} (f_{\mathcal{N}}(x_{\bar{k}+j}) - f(x^*)) + \sum_{j=0}^{\hat{j}-1} \alpha_{\bar{k}+j}^2 \zeta_{\bar{k}+j}^2. \tag{5.30}$$

Notice that

$$\sum_{j=0}^{\hat{j}-1} \alpha_{\bar{k}+j}^2 \zeta_{\bar{k}+j}^2 \leq \sum_{k=0}^{\infty} \frac{C_2^2 \bar{\zeta}^2}{k^2} := \bar{c}_1 < \infty. \tag{5.31}$$

Moreover, using (5.18), (5.31), $\zeta_k \geq \underline{\zeta}$ for all k, and

$$\alpha_{\bar{k}+j} \geq \frac{1}{(\bar{k}+j)^\delta} \geq \frac{1}{(\bar{k}+\hat{j})^\delta}, \quad f_{\mathcal{N}}(x_{\bar{k}+j}) - f(x^*) \geq \varepsilon, \quad j = 0, ..., \hat{j}-1,$$

from (5.30) we obtain

$$0 \leq ||x_{\bar{k}} - x^*||^2 - \frac{2\hat{j}\underline{\zeta}\varepsilon}{q(\bar{k}+\hat{j})^\delta} + \bar{c}_1.$$

Finally, let us observe two cases: 1) $\hat{j} \leq \bar{k}$, and 2) $\hat{j} > \bar{k}$. In the first case, the upper bound on $\hat{j}$ is obvious. In the second case, we have

$$0 \leq ||x_{\bar{k}} - x^*||^2 - \frac{2\hat{j}\underline{\zeta}\varepsilon}{q\hat{j}^{\delta}2^{\delta}} + \bar{c}_1 \leq ||x_{\bar{k}} - x^*||^2 - \frac{\hat{j}^{1-\delta}\underline{\zeta}\varepsilon}{q} + \bar{c}_1,$$

and thus

$$\hat{j} \leq \left(\frac{q(\bar{c}_1 + ||x_{\bar{k}} - x^*||^2)}{\underline{\zeta}}\right)^{\frac{1}{1-\delta}} \varepsilon^{\frac{1}{\delta-1}} =: \bar{c}_2.$$

Combining both cases we conclude that

$$\hat{j} \leq \max\{\bar{c}_2, \bar{k}\} \leq \bar{c}_2 + \bar{k}$$

and thus $\hat{k} \leq \bar{k} + \bar{c}_2 + \bar{k} = 2\bar{k} + \bar{c}_2$, which completes the proof. ■

5.2.3 Numerical Results

Within this section, we test the performance of the AN-SPS algorithm on the well-known binary classification data sets listed in Table 5.1. The problem (5.12) that we consider is a constrained finite-sum problem with L_2-regularized hinge loss local cost functions.

AN-SPS algorithm is implemented with the following parameters: $C_2 = 100, \eta = 10^{-4}, m = 2, N_0 = \lceil 0.1N \rceil$. The initial point x_0 is chosen randomly from Ω. We use Algorithm 3 with $B_k = I$ to find a descent direction $-\overline{g}_k$ which is further scaled as in Step S1 of AN-SPS algorithm, i.e., $p_k = -\zeta_k \overline{g}_k / q_k$. The sample size is updated according to Step S5 of AN-SPS and (5.14). Recall that the sample size is increased only if $\theta_k < h(N_k)$.

We use cumulative samples, i.e., $\mathcal{N}_k \subseteq \mathcal{N}_{k+1}$ and thus, following the conclusions in [9], we calculate the spectral coefficients based on $s_k = x_{k+1} - x_k$ and the subgradient difference $y_k = \tilde{g}_k - \bar{g}_k$, where $\tilde{g}_k \in \partial f_{\mathcal{N}_k}(x_{k+1})$. This choice requires additional costs with respect to

the choice of $\tilde{g}_k = \bar{g}_{k+1}$, but it diminishes the influence of the noise since the difference is calculated on the same approximate function. Furthermore, we test four different choices for the spectral coefficient, BB1, BB2, ABB, and ABBmin, introduced in Section 2.2.1. Recall that

- Barzilai-Borwein 1 (BB1)

$$\lambda_k^{BB1} = \frac{s_k^T s_k}{s_k^T y_k},$$

- Barzilai-Borwein 2 (BB2)

$$\lambda_k^{BB2} = \frac{y_k^T s_k}{y_k^T y_k},$$

- Adaptive Barzilai-Borwein (ABB)

$$\lambda_k := \begin{cases} \lambda_k^{BB2}, & \frac{\lambda_k^{BB2}}{\lambda_k^{BB1}} < 0.8, \\ \lambda_k^{BB1}, & \text{otherwise}, \end{cases}$$

- Adaptive Barzilai-Borwein - minimum (ABBmin)

$$\lambda_k := \begin{cases} \min\{\lambda_j^{BB2} : j = \max\{1, k-5\}, \ldots, k\}, & \frac{\lambda_k^{BB2}}{\lambda_k^{BB1}} < 0.8, \\ \lambda_k^{BB1}, & \text{otherwise}. \end{cases}$$

For all the considered choices we take a safeguard

$$\zeta_k = \min\{\overline{\zeta}, \max\{\underline{\zeta}, \lambda_k\}\}, \quad \underline{\zeta} = 10^{-4}, \quad \overline{\zeta} = 10^4.$$

Since the fixed step size such as $\alpha_k = 1/k$ was already addressed in Section 5.1.3, where the results show that it was clearly outperformed by the line search LS-SPS method, we focus our attention on adaptive step size rules. The value of $\tilde{\alpha}_k^1$ is chosen to be $\tilde{\alpha}_k^1 = \frac{1/k + \bar{\alpha}_k}{2}$, i.e., it is the middle point of the interval $\left[\frac{1}{k}, \bar{\alpha}_k\right]$. Regarding the nonmonotone rule, we also test four choices (see Section 2.3):

- Maximum (MAX), for $c = 5$: $F_k = \max_{i\in[\max\{1,k-5\},k]} f_{\mathcal{N}_i}(x_i)$,
- Convex combination (CCA), for $\eta_k = 0.85$:

$$F_k = \max\{f_{\mathcal{N}_k}(x_k), D_k\},\ D_{k+1} = \frac{\eta_k q_k}{q_{k+1}} D_k + \frac{1}{q_{k+1}} f_{\mathcal{N}_{k+1}}(x_{k+1}),$$

$$D_0 = f_{\mathcal{N}_0}(x_0),\ q_{k+1} = \eta_k q_k + 1,\ q_0 = 1,$$

- Monotone rule (MON): $F_k = f_{\mathcal{N}_k}(x_k)$,
- Additional term (ADA): $F_k = f_{\mathcal{N}_k}(x_k) + \frac{1}{2^k}$.

In order to find the best combination of the strategies proposed above, we track the objective function value and plot it against FEV - the number of scalar products, which serves as a measure of computational cost. All the plots are in the log scale. In the first phase of the experiments, we test AN-SPS with different combinations of spectral coefficients and nonmonotone rules, on four different data sets. The results reveal the benefits of the ADA rule in almost all cases, as it can be seen on representative graphs on MNIST data set (Figure 5.5). In particular, as expected, more "nonmonotonicity" usually yielded better results when combined with the spectral directions.

Furthermore, in order to see the benefits of the proposed adaptive sample size strategy, we compare AN-SPS with:

a) heuristic (HEUR) where the sample size is increased at each iteration by
$$N_{k+1} = \lceil \min\{1.1N_k, N\} \rceil;$$

b) fixed sample strategy (FULL), where $N_k = N$ at each iteration.

A typical behavior of the sequence $\{N_k\}$ for AN-SPS and HEUR is presented in Figure 5.6 on SPLICE data set for BB2 and ADA rule.

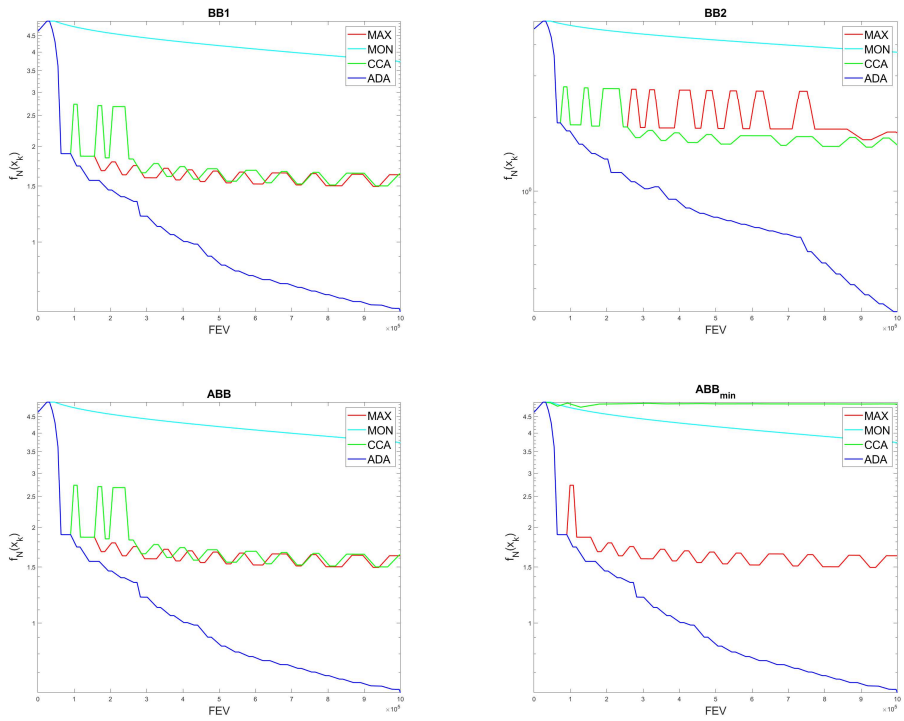

Figure 5.5: AN-SPS algorithm with different nonmonotone rules and spectral coefficients. Objective function value against the computational cost (FEV). MNIST data set.

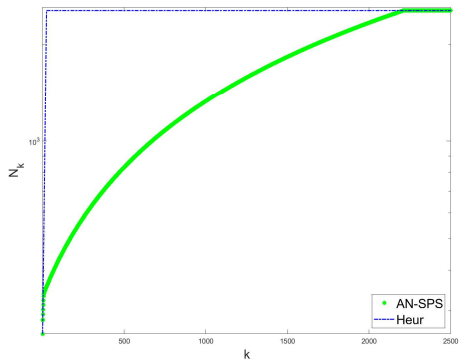

Figure 5.6: AN-SPS sample size versus HEUR sample size sequence. SPLICE data set (BB2 and ADA rule).

We do the same tests for the HEUR and FULL to find the best-performing combinations of BB and line search rules. Finally, we compare the best-performing algorithms of each sample size strategy. The results for all the considered data sets are presented in Figure 5.7 and they show clear advantages of the adaptive sample size strategy in terms of computational costs.

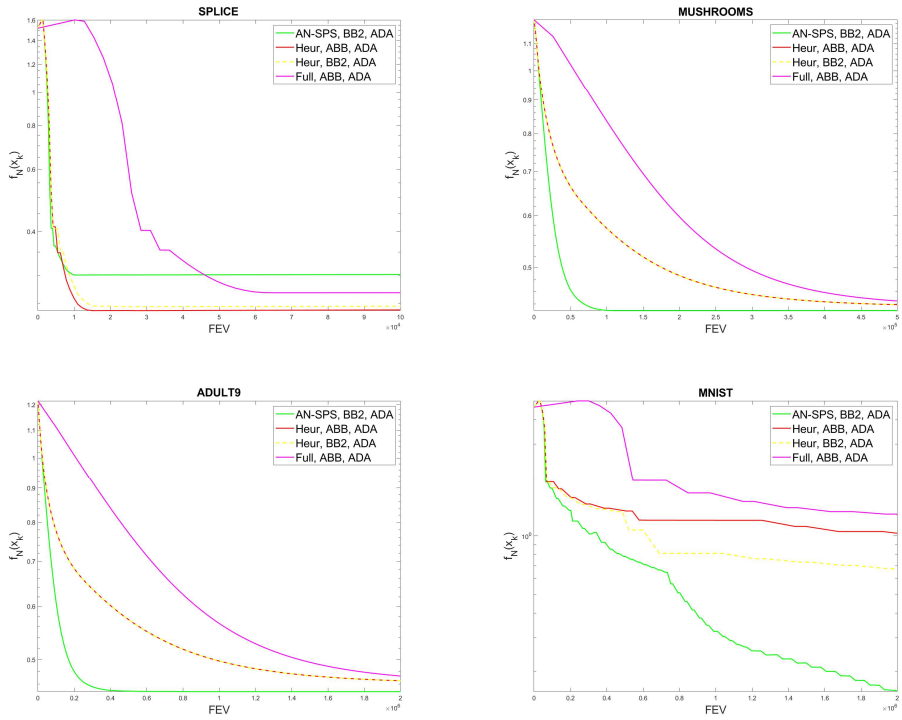

Figure 5.7: Comparison of the best-performing combinations of spectral coefficients and nonmonotone rules of AN-SPS, HEUR, and FULL sample size strategies.

Chapter 6

Nonsmooth Method with Variable Accuracy for Unconstrained Optimization Problems

In this chapter, we consider unconstrained optimization problems in the form (4.3),

$$\min_x f(x) = E\left[F(x,\xi)\right],$$

with nonsmooth and convex objective function in the form of mathematical expectation. The proposed Inexact Restoration - Nonsmooth (IR-NS) algorithm [55], which was developed and implemented within the scope of this thesis, is presented in detail in this chapter. This includes a description of the algorithm, the convergence analysis and an overview of numerical results. The objective function is approximated with a sample average function by using different sample sizes in each iteration. The sample size is chosen in an adaptive manner based on Inexact Restoration. The method uses line search and assumes descent

directions with respect to the current approximate function. We prove the almost sure convergence under the standard assumptions. Numerical results for two types of problems, machine learning hinge loss and stochastic linear complementarity problems, show the efficiency of the proposed scheme.

6.1 Inexact Restoration Nonsmooth Algorithm with Variable Accuracy

The presented method is based on an adaptive variable accuracy and descent directions with respect to the current approximate functions. The sample size is governed by Inexact Restoration (IR) framework introduced by Martinez and Pilota [77] and consists of two phases: the restoration and the optimality phase. The main idea of IR is to treat the phases, restoration and optimality, in a modular way and then to use a merit function, which combines feasibility and optimality and enforces progress towards a feasible optimal point. As IR is a constrained optimization tool, the problem (4.3) is reformulated into a constrained problem as follows

$$\min f_{\mathcal{N}}(x), \text{ s.t. } f_{\mathcal{N}}(x) = f(x), \tag{6.1}$$

where f and $f_{\mathcal{N}}$ are defined in (4.3) and (4.2), respectively.

Notice that (6.1) is equivalent to an unconstrained optimization problem given in form (4.3) if the constraint is satisfied. However if we consider methods that are not strictly feasible, i.e., not all iterations satisfy the constraint, then we can treat $\mathcal{N}$ as an additional variable in the constraint. That is precisely what we will do in the IR approach - in each iteration k of the method we will determine a suitable $N_k = |\mathcal{N}_k|$. There are numerous studies that have confirmed the benefits of using the IR approach in the varying accuracy approximations framework, [8, 60]. The key advantage of this approach is

the fact that feasibility and optimality are kept in balance through merit function. Therefore, the accuracy of the approximate objective function depends on the progress toward optimality in each iteration. So, the accuracy is adaptive, endogenous to the algorithm and there is no need for additional parameters or heuristics in the sample size determination. Furthermore, the sequence of sample sizes is very often nonmonotone, increasing the accuracy (and the computational cost) whenever we approach the solution to ensure good quality of the approximate solution, and decreasing the accuracy (and the costs) when the current iteration is far away from the solution. The approach has been used for variable accuracy approximations for the first time in [60] for the problem of finite-sum minimization coupled with line search descent direction method, based on results from [33]. It is extended to trust region framework and constrained problems, [8, 10, 77]. An approach for solving problems with variable accuracy in both the objective function and constraints is analyzed in [16].

The approach presented here differs in several aspects. First of all, we consider the approximate objective of the form (4.2), $f_{\mathcal{N}}(x) = \frac{1}{N}\sum_{i\in\mathcal{N}} f_i(x)$, and prove that the algorithm introduced here yields $N \to \infty$. In other words we approach the objective function almost surely under some standard conditions. This property of the algorithm is a direct consequence of IR strategy. Furthermore, the conditional expectation of the relevant SAA estimator is equal to the objective function under our settings (for details see the final paragraph of Section 6.2. and the proof of Lemma 6.1.4), and the step size is not directly involved. Another important difference lies in the fact that the objective function and its approximations are not differentiable, and thus the step size analysis is more complicated even in the strongly convex case.

Our contributions are the following. We define IR-NS algorithm for nonsmooth optimization with variable accuracy and prove almost sure convergence of the algorithm under the set of standard assumptions.

By using Inexact Restoration for sample size selection we generalize the results from [102]. More precisely, since IR-NS pushes the SAA error to zero, in the case of finite-sum problems where the objective function is given by (4.2) with the finite full sample size N, the true objective function is reached eventually and the convergence results from [102] hold. IR-NS algorithm also covers a wider class of problems than finite-sums, including infinite-sums. The experiments we perform confirm the intuitive reasoning that working with variable, adaptive sample size is more effective than working with predefined or full sample size as in [102]. To emphasize this fact we present experiments with the same search direction as in [102] - the nonsmooth BFGS descent direction obtained by using Algorithm 3, and demonstrate the advantages of variable sample size approach proposed in IR-NS. In general, an arbitrary descent direction in the sense of Assumption A5 stated below is applicable. From a theoretical point of view, the complexity of order ε^{-2} is proved, which also applies to the method from [102]. The obtained complexity is in line with the results from [15] where the complexity of IR is analyzed. The result in [15] is obtained for smooth constrained problems and is of the form $\varepsilon_{feas}^{-1} + \varepsilon_{opt}^{-2}$, with ε_{feas} being the constant for feasibility and ε_{opt} coincides with the ε that we consider here. Notice that the problems considered in [15] are smooth and deterministic. The complexity results obtained in [10] are not comparable to the complexity results for IR-NS as the methods analyzed in [10] are specialized for smooth problems and problems with regularization. It is important to notice that the choice of sample size we propose here introduces a stochastic iterative sequence which might seem as an unnecessary complication if one is dealing with finite-sum problems. However, we will show that the complexity remains the same and asymptotically we get a.s. convergence, so the stochastic nature does not alter the expected theoretical results. On the other hand, the intrinsic nature of the sample size variation, based on the progress of the iterative process, yields sig-

nificant computational cost savings as demonstrated in the numerical results.

6.1.1 The Algorithm

First, let us recall the Assumption A1, which summarizes the properties of the problem (4.3), where it is assumed that $f_i(x) = F(x, \xi_i)$, $i = 1, 2, \ldots$, are continuous, convex and bounded from below with a constant C for all ξ_i. Following the standard line search method, we assume that a descent direction can be provided for any given function $f_{\mathcal{N}}$ such that condition (3.8) holds.

Assumption A 5 *For any given N, x and B such that $mI \preceq B(x) \preceq MI$, for some positive and bounded constants $m \leq M$ we can compute a direction $p_{\mathcal{N}} \in \mathbb{R}^n$ such that*

$$p_{\mathcal{N}}(x) = -B(x)\bar{g}_{\mathcal{N}}(x) \quad \textit{and} \sup_{g \in \partial f_{\mathcal{N}}(x)} g^T p_{\mathcal{N}}(x) \leq -\frac{m}{2}\|\bar{g}_{\mathcal{N}}(x)\|^2,$$

where $\bar{g}_{\mathcal{N}}(x) \in \partial f_{\mathcal{N}}(x)$.

Let us briefly discuss the plausibility of the above assumption. One possibility to generate such direction is using Algorithm 3 presented in Section 3.5, where B is the BFGS matrix. If an oracle for calculating $\sup_{g \in \partial f_{\mathcal{N}}(x)} g^T p_{\mathcal{N}}(x)$ is available, then we can take the subgradient descent direction. Another approach would be to use gradient subsampling techniques [17]. For directions that satisfy Assumption A5 the following result, as a direct corollary of the Theorem 3.5.1 listed in the Section 3.5, holds. We provide the proof for the sake of completeness.

Lemma 6.1.1 *Let Assumptions A1 and A5 hold. Then there exist $\tau_{\mathcal{N}}(x) > 0$ and $\gamma \in (0, 1)$ such that the subgradient Armijo condition*

$$f_{\mathcal{N}}(x + \alpha p_{\mathcal{N}}(x)) \leq f_{\mathcal{N}}(x) - \gamma\alpha\|p_{\mathcal{N}}(x)\|^2.$$

holds for all $\alpha \in [0, \tau_{\mathcal{N}}(x)]$.

Proof. Let us fix an arbitrary N and an arbitrary $x \in \mathbb{R}^n$. If $\bar{g}_{\mathcal{N}}(x) = 0$ the statement is obviously true. In the case $\bar{g}_{\mathcal{N}}(x) \neq 0$ we can define $\delta(\alpha) := f_{\mathcal{N}}(x + \alpha p_{\mathcal{N}}(x))$, where $p_{\mathcal{N}}(x)$ is a descent direction satisfying Assumption A5. For such $p_{\mathcal{N}}(x)$ there holds

$$\delta'(0) = \sup_{g \in \partial f_N(x)} g^T p_{\mathcal{N}}(x) < 0.$$

Consider

$$l(\alpha) := f_{\mathcal{N}}(x) + \alpha\eta \sup_{g \in \partial f_{\mathcal{N}}(x)} g^T p_{\mathcal{N}}(x),$$

for some $\eta \in (0, 1)$. Given that $\sup_{g \in \partial f_{\mathcal{N}}(x)} g^T p_{\mathcal{N}}(x) < 0$, $f_{\mathcal{N}}$ is bounded from below and convex by Assumption A1, there exists an unique intersection of the functions δ and l on the interval $\alpha \in (0, \infty)$. Let us denote this intersection by $\tau_{\mathcal{N}}(x)$. Then, for all $\alpha \in [0, \tau_{\mathcal{N}}(x)]$ there holds

$$f_{\mathcal{N}}(x + \alpha p_{\mathcal{N}}(x)) \leq f_{\mathcal{N}}(x) + \alpha\eta \sup_{g \in \partial f_{\mathcal{N}}(x)} g^T p_{\mathcal{N}}(x).$$

Furthermore, Assumption A5 implies

$$f_{\mathcal{N}}(x + \alpha p_{\mathcal{N}}(x)) \leq f_{\mathcal{N}}(x) - \alpha\eta \frac{m}{2} \|\bar{g}_{\mathcal{N}}(x)\|^2 \leq f_{\mathcal{N}}(x) - \alpha\eta \frac{m}{2M^2} \|p_{\mathcal{N}}(x)\|^2$$

and the statement holds for $\gamma = \eta m/(2M^2)$. ■

The problem we are solving is defined by (6.1). Clearly the feasibility condition $f_{\mathcal{N}}(x) = f(x)$ cannot be enforced in the general case of expected value as in that case, we should have $N \to \infty$. Furthermore, neither the deviation from feasible condition $|f(x) - f_{\mathcal{N}}(x)|$ can be computed. Thus we introduce an approximate infeasibility measure defined as a function $h(N)$ for arbitrary integer N. Assume that $h : \mathbb{N} \to \mathbb{R}_+ \cup \{0\}$ is monotonically decreasing function such that $\lim_{N \to \infty} h(N) = 0$. In other words, $h(N)$ is a proxy for $|f(x) - f_{\mathcal{N}}(x)|$ as we have seen earlier. Recall that if we are solving a finite-sum problem, i.e., if $f(x) = f_{N_{\max}}(x)$ for a fixed $N_{\max}$ then for arbitrary

$N \leq N_{\max}$ we can use (4.7), $h(N) = (N_{max} - N)/N_{max}$. For the case of unbounded N one possible simple choice is (4.8), $h(N) = 1/N$. The merit function for IR is defined in the usual way

$$\Phi(x, N, \vartheta) := \vartheta f_{\mathcal{N}}(x) + (1 - \vartheta)h(N),$$

where $\vartheta \in (0, 1)$ is the penalty parameter used to give different weights to the objective function and the measure of infeasibility and N is an integer that defines the level of accuracy in the approximate function $f_{\mathcal{N}}$.

At each iteration k we have the accuracy parameter as an integer N_k, the solution estimate x_k, the penalty parameter ϑ_k and the approximate objective function $f_{\mathcal{N}_k}$. The presented algorithm is denoted as Algorithm 7.

Let us briefly discuss the key points of IR-NS algorithm. In Step S1 the feasibility is improved, i.e., a new sample size candidate $\tilde{N}_{k+1}$ is chosen. Additionally, the value $f_{\tilde{\mathcal{N}}_{k+1}}(x_k)$ might increase with respect to $f_{\mathcal{N}_k}(x_k)$ by at most $\beta h(N_k)$. Thus, optimality can deteriorate with respect to the previous iteration but the deterioration is controlled by the function h, i.e., it depends on the accuracy of the objective function. So, for smaller N_k - which means a looser approximation of the true objective function, the deterioration of optimality can be relatively large, as we assume that we are still far away from the solution. Parameter β can be arbitrarily large, but finite. In some applications (ex. finite-sums) one can prove that such β exists under standard conditions. However, in general, since we do not impose differentiability of the objective function nor any other special property, the following assumption is needed.

Assumption A 6 *Suppose that there exists β such that* (6.2) *holds for each k.*

The penalty parameter is updated in such a way that it ensures a decrease of the merit function as stated in Lemma 6.1.2. Moreover,

Algorithm 7: IR-NS (**I**nexact **R**estoration - **N**on**S**mooth)

S0 *Initialization.* Given $x_0 \in \mathbb{R}^n, N_0 \in \mathbb{N}, \vartheta_0, r \in (0,1), \beta, \gamma, \overline{\gamma} > 0$. Set $k = 0$.

S1 *Restoration phase.* Find $\tilde{N}_{k+1} \geq N_k$ such that

$$h(\tilde{N}_{k+1}) \leq r h(N_k),$$

$$f_{\tilde{\mathcal{N}}_{k+1}}(x_k) - f_{\mathcal{N}_k}(x_k) \leq \beta h(N_k). \tag{6.2}$$

S2 *Updating the penalty parameter.*
If

$$\Phi(x_k, \tilde{N}_{k+1}, \vartheta_k) - \Phi(x_k, N_k, \vartheta_k) \leq \frac{1-r}{2}\left(h(\tilde{N}_{k+1}) - h(N_k)\right) \tag{6.3}$$

set $\vartheta_{k+1} = \vartheta_k$.
Else

$$\vartheta_{k+1} := \frac{(1+r)(h(N_k) - h(\tilde{N}_{k+1}))}{2\left[f_{\tilde{\mathcal{N}}_{k+1}}(x_k) - f_{\mathcal{N}_k}(x_k) + h(N_k) - h(\tilde{N}_{k+1})\right]}. \tag{6.4}$$

S3 *Optimization Phase.* Choose $N_{k+1} \leq \tilde{N}_{k+1}$, $p_{\mathcal{N}_{k+1}} \in \mathbb{R}^n$ and $\alpha_k \in (0,1]$ such that

$$f_{\mathcal{N}_{k+1}}(x_k + \alpha_k p_{\mathcal{N}_{k+1}}(x_k)) - f_{\tilde{\mathcal{N}}_{k+1}}(x_k) \leq -\gamma\alpha_k ||p_{\mathcal{N}_{k+1}}(x_k)||^2, \tag{6.5}$$

$$h(N_{k+1}) \leq h(\tilde{N}_{k+1}) + \bar{\gamma}\alpha_k^2 ||p_{\mathcal{N}_{k+1}}(x_k)||^2, \tag{6.6}$$

$$\begin{aligned} &\Phi(x_k + \alpha_k p_{\mathcal{N}_{k+1}}(x_k), N_{k+1}, \vartheta_{k+1}) - \Phi(x_k, N_k, \vartheta_{k+1}) \\ &\leq \frac{1-r}{2}\big(h(\tilde{N}_{k+1}) - h(N_k)\big). \end{aligned} \tag{6.7}$$

S4 Set $p_k = p_{\mathcal{N}_{k+1}}(x_k)$, $x_{k+1} = x_k + \alpha_k p_k$, $k := k+1$ and go to **S1**.

it can also be shown that the sequence of ϑ_k is non-increasing and bounded away from zero which prevents the optimality part to vanish from the merit function. The proof of Lemma 6.1.2 is fundamentally the same as in [60, Lemma 2.1], but we provide it for the sake of completeness.

Lemma 6.1.2 *[60] Let Assumptions A1, A5 and A6 hold. Then the sequence $\{\vartheta_k\}$ generated by Algorithm IR-NS is positive and non-increasing, the inequality*

$$\Phi(x_k, \tilde{N}_{k+1}, \vartheta_{k+1}) - \Phi(x_k, N_k, \vartheta_{k+1}) \leq \frac{1-r}{2}\left(h(\tilde{N}_{k+1}) - h(N_k)\right)$$

holds and there exists $\vartheta^* > 0$ such that $\lim_{k\to\infty} \vartheta_k = \vartheta^*$.

Proof. First, let us show that the sequence $\{\vartheta_k\}$ is non-increasing. If (6.3) holds, from Step S2 we have $\vartheta_{k+1} = \vartheta_k$.
Otherwise, since (6.3) does not hold , it follows

$$\Phi(x_k, \tilde{N}_{k+1}, \vartheta_k) - \Phi(x_k, N_k, \vartheta_k) > \frac{1-r}{2}\big(h(\tilde{N}_{k+1}) - h(N_k)\big),$$

i.e.,

$$\begin{aligned}&\vartheta_k\left(f_{\tilde{\mathcal{N}}_{k+1}}(x_k) - f_{\mathcal{N}_k}(x_k)\right) + (1-\vartheta_{k+1})\left(h(\tilde{N}_{k+1}) - h(N_k)\right)\\ &> \frac{1-r}{2}\big(h(\tilde{N}_{k+1}) - h(N_k)\big).\end{aligned}$$

Furthermore,

$$\begin{aligned}&\vartheta_k\left(f_{\tilde{\mathcal{N}}_{k+1}}(x_k) - f_{\mathcal{N}_k}(x_k) - h(\tilde{N}_{k+1}) + h(N_k)\right)\\ &> \frac{1-r}{2}\big(h(\tilde{N}_{k+1}) - h(N_k)\big) - h(\tilde{N}_{k+1}) + h(N_k).\end{aligned}$$

Therefore, we have

$$\vartheta_k > \frac{1+r}{2}\frac{h(N_k)-h(\tilde{N}_{k+1})}{f_{\tilde{\mathcal{N}}_{k+1}}(x_k)-f_{\mathcal{N}_k}(x_k)+h(N_k)-h(\tilde{N}_{k+1})} := \vartheta_{k+1},$$

and it follows that $\{\vartheta_k\}$ is a non-increasing sequence.

Now, we want to show that ϑ_{k+1} given by (6.4) is bounded away from zero. It holds

$$\begin{aligned}
&\Phi(x_k,\tilde{N}_{k+1},\vartheta_{k+1})-\Phi(x_k,N_k,\vartheta_{k+1})\\
&=\vartheta_{k+1}\left[f_{\tilde{\mathcal{N}}_{k+1}}(x_k)-f_{\mathcal{N}_k}(x_k)+h(N_k)-h(\tilde{N}_{k+1})\right]+h(\tilde{N}_{k+1})-h(N_k)\\
&=\frac{(1+r)(h(N_k)-h(\tilde{N}_{k+1}))}{2}+h(\tilde{N}_{k+1})-h(N_k)\\
&=\frac{h(\tilde{N}_{k+1})-h(N_k)-r(h(\tilde{N}_{k+1})-h(N_k))}{2}\\
&=\frac{1-r}{2}\Big(h(\tilde{N}_{k+1})-h(N_k)\Big).
\end{aligned}$$

Moreover,

$$\begin{aligned}
\frac{1}{\vartheta_{k+1}} &= \frac{2}{1+r}\frac{f_{\tilde{\mathcal{N}}_{k+1}}(x_k)-f_{\mathcal{N}_k}(x_k)+h(N_k)-h(\tilde{N}_{k+1})}{h(N_k)-h(\tilde{N}_{k+1})}\\
&= \frac{2}{1+r}\Big(\frac{f_{\tilde{N}_{k+1}}(x_k)-f_{N_k}(x_k)}{h(N_k)-h(\tilde{N}_{k+1})}+1\Big)\\
&\le \frac{2}{1+r}\Big(\frac{\beta h(N_k)}{h(N_k)-rh(N_k)}+1\Big)\\
&= \frac{2}{1+r}\Big(\frac{\beta}{1-r}+1\Big) := \frac{1}{\tilde{\vartheta}},
\end{aligned}$$

Therefore,

$$\vartheta_{k+1} \ge \tilde{\vartheta} > 0.$$

Since $\{\vartheta_k\}$ is a non-increasing sequence bounded from below, it follows that there exists ϑ^* such that

$$\lim_{k\to\infty} \vartheta_k = \vartheta^*,\ \vartheta^* \geq \tilde{\vartheta}. \blacksquare$$

In Step S3 we chose the sample size to be used in the subsequent iteration. Notice that one possible choice is $N_{k+1} = \tilde{N}_{k+1}$ since (6.5)-(6.6) are satisfied due to Lemma 6.1.1 and, as we will prove in Lemma 6.1.3, there exists α_k which satisfies inequality (6.7) in that case as well. On the other hand, in order to decrease the overall costs, we try to decrease the sample size if it still provides the decrease in the merit function (6.7). The resulting sample size N_{k+1} can be larger, equal, or smaller than N_k. Our numerical study shows that allowing the decrease of sample size is beneficial in terms of overall function evaluations. In practical implementations, we estimate the sample size lower bound N_{k+1}^{trial} derived from (6.7) and let

$$N_{k+1} \in \{N_{k+1}^{trial}, \lceil (N_{k+1}^{trial} + \tilde{N}_{k+1})/2 \rceil, \tilde{N}_{k+1}\}.$$

We use the backtracking technique for finding α_k, but at each backtracking step we try all three candidate values for N_{k+1}. This is just one possible approach and the optimal strategy remains an open question, probably problem-dependent.

Lemma 6.1.3 *Let Assumptions A1, A5 and A6 hold. Then, there exists $\gamma > 0$ such that Step S3 of Algorithm IR-NS is well-defined.*

Proof. The algorithm is well defined if there exists a choice of $N_{k+1} \leq \tilde{N}_{k+1}$ and a descent direction p_k such that (6.5) - (6.7) hold for some $\alpha_k > 0$ and a suitable $\gamma > 0$ for each k. Let us take $N_{k+1} = \tilde{N}_{k+1}$ and retain the same sample so that $f_{\mathcal{N}_{k+1}} = f_{\tilde{\mathcal{N}}_{k+1}}$. In that case Lemma 6.1.1 implies the existence of $\tau_k := \tau_{\mathcal{N}_{k+1}}(x_k) > 0$ such that the inequality (6.5) holds for all $\alpha \in [0, \tau_k]$. Since (6.6) is trivially

satisfied for this choice of N_{k+1}, it remains to prove the existence of $\alpha_k \in [0, \tau_k]$ such that (6.7) holds. By (6.5), (6.6) and Lemma 6.1.2, for all $\alpha \in [0, \tau_k]$,

$$\begin{aligned}
&\Phi(x_k + \alpha p_k, N_{k+1}, \vartheta_{k+1}) - \Phi(x_k, N_k, \vartheta_{k+1}) \\
&= \Phi(x_k + \alpha p_k, N_{k+1}, \vartheta_{k+1}) - \Phi(x_k, \tilde{N}_{k+1}, \vartheta_{k+1}) \\
&+ \Phi(x_k, \tilde{N}_{k+1}, \vartheta_{k+1}) - \Phi(x_k, N_k, \vartheta_{k+1}) \\
&\leq \Phi(x_k + \alpha p_k, N_{k+1}, \vartheta_{k+1}) - \Phi(x_k, \tilde{N}_{k+1}, \vartheta_{k+1}) \\
&+ \frac{1-r}{2}\left(h(\tilde{N}_{k+1}) - h(N_k)\right) \\
&= \vartheta_{k+1}\left(f_{N_{k+1}}(x_k + \alpha p_k) - f_{\tilde{N}_{k+1}}(x_k)\right) \\
&+ \frac{1-r}{2}\left(h(\tilde{N}_{k+1}) - h(N_k)\right) \\
&\leq -\vartheta_{k+1}\gamma\alpha||p_k||^2 + \frac{1-r}{2}\left(h(\tilde{N}_{k+1}) - h(N_k)\right) \\
&\leq \frac{1-r}{2}\left(h(\tilde{N}_{k+1}) - h(N_k)\right).
\end{aligned}$$

Therefore, (6.7) holds for all $\alpha \in [0, \tau_k]$. ■

Notice that in the above Lemma, we proved only that the algorithm is well defined, i.e., we can always take $N_{k+1} = \tilde{N}_{k+1}$ and the $(k+1)$th iteration is well defined. However, other possibilities for N_{k+1} exist and we discuss some of them in Section 5.2.3. Since the sample size sequence is not monotonically increasing in general, it is not obvious that N_k tends to infinity. Nevertheless, using essentially the same proof as in [60, Theorem 2.1], we conclude that the infeasibility measure tends to zero yielding the result of $\lim_{k\to\infty} N_k = \infty$. Specially, for the finite-sum problem we conclude that the full sample is reached after a finite number of iterations. The proof of Theorem

2.1 in [60] contains an important relation stated below

$$\sum_{k=0}^{\infty} h(N_k) \leq C_6 < \infty, \tag{6.8}$$

where $C_6 > 0$ is a constant, that we will use in further convergence analysis presented in the next section.

Let us now provide more insights regarding the stochastic concept of the proposed algorithm. IR-NS yields a stochastic sequence of iterates x_k. The stochastic nature comes from the sequence of random variables N_k that determine the samples to be used for the SAA functions. Assume that we are at iteration k and x_k is known. Denote by $\mathcal{F}_k$ the σ-algebra generated by $x_0, ..., x_k$, i.e., by random variables that determine $f_{\tilde{\mathcal{N}}_j}, j = 1, ..., k$ and $f_{\mathcal{N}_j}, j = 0, ..., k$. Since the samples are assumed to be i.i.d., we have conditionally unbiased estimators. More precisely, at the beginning of Step S1 of the algorithm a new sample size $\tilde{N}_{k+1}$ is chosen and a random sample is generated to obtain $f_{\tilde{\mathcal{N}}_{k+1}}$. Thus, since x_k is $\mathcal{F}_k$-measurable (i.e., known at that point of the algorithmic procedure), there holds

$$E\left[f_{\tilde{\mathcal{N}}_{k+1}}(x_k)|\mathcal{F}_k\right] = f(x_k), \tag{6.9}$$

where $E\left[\cdot|\mathcal{F}_k\right]$ denotes the conditional expectation with respect to $\mathcal{F}_k$ [35]. Also $E\left[f_{\mathcal{N}_{k+1}}(x_k)|\mathcal{F}_k\right] = f(x_k)$. However, $E\left[f_{\mathcal{N}_{k+1}}(x_{k+1})|\mathcal{F}_k\right]$ is not equal to $f(x_{k+1})$ in general because x_{k+1} is dependent on N_{k+1}. More precisely, the second round of stochastic influence within iteration k comes at Step S3 where we choose N_{k+1} which may yield totally different sample for $f_{\mathcal{N}_{k+1}}$ with respect to $f_{\tilde{\mathcal{N}}_{k+1}}$ in general (each trial sample size may yield different sample). Moreover, the direction $p_{\mathcal{N}_{k+1}}(x_{k+1})$ and the step size α_k directly depend on the generated samples and thus we lose the martingale property. This is a common situation in stochastic line search (see [14] for instance). In Step S4, we set the next iteration and return to Step S1, repeating the procedure.

6.1.2 Convergence Analysis

The convergence analysis is performed under the set of standard assumptions for stochastic problems. We analyze conditions needed for a.s. convergence of IR-NS and provide complexity result at the end of this section.

Assumption A 7 *The objective function f has bounded level sets.*

This assumption holds if the objective function is strongly convex for example, and we have the following result.

Lemma 6.1.4 *Let Assumptions A1 and A5-A7 hold. Suppose that there exists a constant C_0 such that $F(x_0, \xi) \leq C_0$ for any ξ. Then $f(x_k) \leq C_2$ holds for all k, i.e., $\{x_k\}_{k\in\mathbb{N}} \subseteq D$, where*

$$D = \{x \in \mathbb{R}^n \mid f(x) \leq C_7\}$$

and $C_7 = C_0 + 2\beta C_6$.

Proof. The set D is compact by Assumption A7. Using inequalities (6.2)-(6.5), for all k we obtain

$$f_{\mathcal{N}_{k+1}}(x_{k+1}) \leq f_{\tilde{\mathcal{N}}_{k+1}}(x_k) - \gamma\alpha_k ||p_{\mathcal{N}_{k+1}}(x_k)||^2 \leq f_{\mathcal{N}_k}(x_k) + \beta h(N_k).$$

Furthermore, using the induction argument and (6.8) we get

$$f_{\mathcal{N}_{k+1}}(x_{k+1}) \leq f_{\mathcal{N}_0}(x_0) + \beta \sum_{j=0}^{k} h(N_j) \leq f_{\mathcal{N}_0}(x_0) + \beta C_6,$$

for all $k = 0, 1, ...$ Obviously, the assumption of uniformly bounded F at the initial point x_0 implies that $f_{\mathcal{N}_0}(x_0) \leq C_0$ and we obtain

$$f_{\mathcal{N}_k}(x_k) \leq C_0 + \beta C_6, \tag{6.10}$$

for all $k = 1, 2, ...$ Finally, by (6.9) and inequalities (6.2) and (6.10) we get

$$f(x_k) = E\left[f_{\tilde{\mathcal{N}}_{k+1}}(x_k)|\mathcal{F}_k\right] \leq E\left[f_{\mathcal{N}_k}(x_k) + \beta h(N_k)|\mathcal{F}_k\right] \leq C_0 + 2\beta C_6 := C_7,$$

which completes the proof. ■

Recall that in Assumption A2 we assume that the function F is dominated by a P-integrable function on any compact subset of $\mathbb{R}^n$. Let now the function F be dominated by an integrable function on a bounded open set $\tilde{D}^0$ such that $D \subset \tilde{D}^0$. Define

$$\tilde{e}_k := \max_{x,y \in \tilde{D}} \{|f(x) - f_{\mathcal{N}_{k+1}}(x)| + |f(y) - f_{\tilde{\mathcal{N}}_{k+1}}(y)|\}, \tag{6.11}$$

where $\tilde{D}$ is a compact enlargement of D, i.e., $\tilde{D}$ is the closure of an open set $\tilde{D}^0 \supset D$. Therefore, both D and $\tilde{D}$ are compact sets and $D \subsetneq \tilde{D}$.

Notice that ULLN and the fact $h(N_k) \to 0$ imply that $\tilde{e}_k \to 0$ a.s. if $N_k \to \infty$. Let us analyze the convergence depending on properties of the step size sequence $\{\alpha_k\}$ and the error sequence $\{\tilde{e}_k\}$.

Theorem 6.1.1 *Let Assumptions A1-A2 and A5-A7 hold and $\{x_k\}$ be a sequence generated by Algorithm IR-NS. If $\alpha_k \geq \overline{\alpha} > 0$ for all $k \in \mathbb{N}$ then there exists an accumulation point x^* of $\{x_k\}$ which is a solution of problem* (4.3) *a.s.*

Proof. Denote $\bar{g}_k = \bar{g}_{\mathcal{N}_k}(x_k)$. Then Assumption A5 and (6.5) imply

$$f_{\mathcal{N}_{k+1}}(x_{k+1}) \leq f_{\tilde{\mathcal{N}}_{k+1}}(x_k) - \gamma\alpha_k||p_k||^2 \leq f_{\tilde{\mathcal{N}}_{k+1}}(x_k) - \eta\alpha_k||\overline{g}_k||^2,$$

where $\eta = \gamma m^2$. Furthermore,

$$\begin{aligned} f(x_{k+1}) &\leq f_{\tilde{\mathcal{N}}_{k+1}}(x_k) - \eta\alpha_k||\overline{g}_k||^2 + f(x_{k+1}) - f_{\mathcal{N}_{k+1}}(x_{k+1}) \\ &\leq f(x_k) - \eta\alpha_k||\overline{g}_k||^2 + |f(x_{k+1}) - f_{\mathcal{N}_{k+1}}(x_{k+1})| \\ &\quad + |f_{\tilde{\mathcal{N}}_{k+1}}(x_k) - f(x_k)|. \end{aligned}$$

From the definition of $\tilde{e}_k$ (6.11), we obtain

$$f(x_{k+1}) \leq f(x_k) - \eta\bar{\alpha}||\overline{g}_k||^2 + \tilde{e}_k. \tag{6.12}$$

We will show that $\liminf_{k\to\infty} ||\overline{g}_k||^2 = 0$. Assume the contrary, i.e., that $||\overline{g}_k||^2 \geq \varrho > 0$ for some $\varrho > 0$ and all k. Then

$$\eta\overline{\alpha}||\overline{g}_k||^2 \geq \eta\overline{\alpha}\varrho > 0.$$

Since $\tilde{e}_k \to 0$ a.s., there exists $\overline{k}$ such that for all $k \geq \overline{k}$ there holds $\tilde{e}_k \leq \frac{1}{2}\eta\overline{\alpha}||\overline{g}_k||^2$ a.s. and thus (6.12) implies

$$f(x_{k+1}) \leq f(x_k) - \eta\overline{\alpha}/2 \text{ a.s.}$$

Equivalently, for all $s \in \mathbb{N}$ we have

$$f(x_{\overline{k}+s}) \leq f(x_{\overline{k}}) - \frac{s}{2}\eta\overline{\alpha}\varrho \quad \text{a.s.} \tag{6.13}$$

Letting $s \to \infty$ yields a contradiction with the Assumption A1 which implies that f is bounded from below. Therefore, we conclude that there there exists $K \subseteq \mathbb{N}$ such that

$$\lim_{k\in K} \overline{g}_k = 0 \text{ a.s.}$$

Since $\{x_k\} \subset D$ and D is compact there follows that there exist $K_1 \subseteq K$ and $x^* \in D$ such that

$$x^* = \lim_{k\in K_1} x_k.$$

Now, using the fact that $\overline{g}_k \in \partial f_{\mathcal{N}_{k+1}}(x_k)$, for all $x \in \mathbb{R}^n$ we have

$$f_{\mathcal{N}_{k+1}}(x) \geq f_{\mathcal{N}_{k+1}}(x_k) + \overline{g}_k^T(x - x_k).$$

Thus, for arbitrary $x \in \tilde{D}$ we have

$$\begin{aligned} f(x) &\geq f_{\mathcal{N}_{k+1}}(x_k) + \overline{g}_k^T(x - x_k) + f(x) - f_{\mathcal{N}_{k+1}}(x) \\ &= f(x_k) + \overline{g}_k^T(x - x_k) - \left(f_{\mathcal{N}_{k+1}}(x) - f(x) + f(x_k) - f_{\mathcal{N}_{k+1}}(x_k)\right) \\ &\geq f(x_k) + \overline{g}_k^T(x - x_k) - \left(|f(x) - f_{\mathcal{N}_{k+1}}(x)| + |f(x_k) - f_{\mathcal{N}_{k+1}}(x_k)|\right). \end{aligned}$$

Therefore,

$$f(x) \geq f(x_k) - ||\overline{g}_k||||x - x_k|| - 2\tilde{e}_k.$$

Taking the limit over K_1 and using the fact that $||x - x_k||$ is bounded, we obtain that for every $x \in \tilde{D}$ there holds

$$f(x) \geq f(x^*), \text{ a.s.} \tag{6.14}$$

Recall that $x^* \in D$ and $\tilde{D}$ is a compact enlargement of D so x^* cannot be on the boundary of $\tilde{D}$, so there exists $\epsilon > 0$ such that $\mathcal{B}(x^*, \epsilon) \subset \tilde{D}$. Thus, x^* is a local minimizer of f a.s. Since f is assumed to be convex, we conclude that $x^* \in X^*$ a.s. ■

It can be also proved that every strictly strong accumulation point [101] is a solution a.s. Recall Definition 1.1.8, we say that a point x^* is a strictly strong accumulation point of the sequence $\{x_k\}_{k\in\mathbb{N}}$ if there exists a subsequence $K \subseteq \mathbb{N}$ and a constant $b \in \mathbb{N}$ such that $\lim_{k_i \in K} x_{k_i} = x^*$ and $k_{i+1} - k_i \leq b$ for any two consecutive elements $k_i, k_{i+1} \in K$. According to the available literature, [92, 99], and up to the best of our knowledge, a stronger statement in a.s. sense is not possible without some additional assumptions on the rate of increase of N_k.

Theorem 6.1.2 *Assume that the conditions of Theorem 6.1.1 hold. Then every strictly strong accumulation point of the sequence* $\{x_k\}$ *is a solution of problem* (4.3) *a.s.*

Proof.

Let x^* be an arbitrary strictly strong accumulation point of the sequence $\{x_k\}$, i.e., $x^* = \lim_{i\to\infty} x_{k_i}$ and $s_i := k_{i+1} - k_i \leq b$ for every $i \in \mathbb{N}$. Since (6.12) holds for each $k \in \mathbb{N}$, we obtain

$$f(x_{k_{i+1}}) \leq f(x_{k_i}) - \eta\overline{\alpha} \sum_{j=0}^{s_i-1} ||\overline{g}_{k_i+j}||^2 + \sum_{j=0}^{s_i-1} t_{k_i+j} \leq f(x_{k_i}) - \eta\overline{\alpha}||\overline{g}_{k_i}||^2 + \omega_i,$$

where $\omega_i = \sum_{j=0}^{b-1} t_{k_i+j}$. Notice that $\omega_i \to 0, i \to \infty$ a.s. We want to show that

$$\liminf_{i\to\infty} ||\overline{g}_{k_i}||^2 = 0 \text{ a.s.} \tag{6.15}$$

Assume the contrary, i.e., for all $i \in \mathbb{N}$ there holds $||\overline{g}_{k_i}||^2 \geq \varrho > 0$ for some $\varrho > 0$. Then,

$$\eta\overline{\alpha}||\overline{g}_{k_i}||^2 \geq \eta\overline{\alpha}\varrho > 0$$

for all $i \in \mathbb{N}$. Therefore, there exists $\bar{i}$ such that for all $i \geq \bar{i}$ there holds $\omega_i \leq \frac{1}{2}\eta\overline{\alpha}\varrho$ a.s. and thus

$$f(x_{k_{i+1}}) \leq f(x_{k_i}) - \frac{1}{2}\eta\overline{\alpha}\varrho \text{ a.s.}$$

Letting $i \to \infty$ in the last inequality we obtain

$$f(x^*) \leq f(x^*) - \frac{1}{2}\eta\overline{\alpha}\varrho < f(x^*),$$

which is a contradiction. So, (6.15) holds and repeating the steps (6.13)-(6.14) from the proof of Theorem 6.1.1, we obtain the result, i.e., $x^* \in X^*$ a.s. ■

Next, we show that the convergence result as in Theorem 6.1.1 can be obtained under weaker assumptions on the step size sequence, but assuming that the increase of sample size N_k is eventually fast enough, i.e., $\sum_{k=0}^{\infty} \tilde{e}_k < \infty$. For instance, if the sample is cumulative, the log bound given by (4.10) in Section 4.2.1 holds and $\sum_{k=0}^{\infty} \tilde{e}_k < \infty$ is true

if $N_k \geq e^k$. Therefore, one can switch to exponential growth after a certain number of iterations of the IR-NS algorithm, taking advantage of cheap iterations in the early stages and theoretically proved convergence for the fast increase of sample size sequence in the later stages of the algorithm. The switching point is an interesting problem but beyond this thesis's scope.

Theorem 6.1.3 *Let Assumptions A1-A2 and A5-A7 hold and* $\{x_k\}$ *be a sequence generated by Algorithm IR-NS. If* $\sum_{k=0}^{\infty} \alpha_k = \infty$ *and* $\sum_{k=0}^{\infty} \tilde{e}_k < \infty$ *then there exists an accumulation point* x^* *of* $\{x_k\}$ *which is a solution of problem* (4.3).

Proof. Following the steps of the proof of Theorem 6.1.1 we obtain

$$f(x_{k+1}) \leq f(x_k) - \eta \alpha_k ||\overline{g}_k||^2 + \tilde{e}_k$$

for every k and thus

$$f(x_{k+1}) \leq f(x_0) - \eta \sum_{i=0}^{k} \alpha_i ||\overline{g}_i||^2 + \sum_{i=0}^{k} t_i.$$

The function f is bounded from below and $\sum_{k=0}^{\infty} \tilde{e}_k < \infty$, so we conclude

$$\sum_{k=0}^{\infty} \alpha_k ||\overline{g}_k||^2 < \infty. \tag{6.16}$$

Furthermore, the assumption $\sum_{k=0}^{\infty} \alpha_k = \infty$ implies the existence of a subset K_1 such that $\lim_{k \in K_1} \overline{g}_k = 0$. Indeed, if we assume the contrary, i.e., that there exists $\varepsilon > 0$ such that $\|\overline{g}_k\| \geq \varepsilon > 0$ for k large enough, then we obtain

$$\sum_{k=0}^{\infty} \alpha_k ||\overline{g}_k||^2 \geq \sum_{k=0}^{\infty} \alpha_k \varepsilon^2 = \varepsilon^2 \sum_{k=0}^{\infty} \alpha_k = \infty,$$

which is in contradiction with (6.16). Since the whole sequence $\{x_k\}_{k\in\mathbb{N}}$ is bounded due to Lemma 6.1.4, there exist $K_2 \subseteq K_1$ and $x^* \in D$ such that

$$\lim_{k\in K_2} x_k = x^*.$$

Now, repeating the proof of Theorem 6.1.1 - the part after (6.13), we conclude that $x^* \in X^*$. ■

The following result is based on considerations in [7] and [38] and essentially yields worst-case complexity analysis with respect to the expected objective function value.

Theorem 6.1.4 *Let Assumptions A1-A2 and A5-A7 hold, $\varepsilon > 0$ and $\{x_k\}$ be a sequence generated by Algorithm IR-NS. Furthermore, assume that $\alpha_k \geq \overline{\alpha} > 0$ for all $k \in \mathbb{N}$ and $\sum_{k=0}^{\infty} \tilde{e}_k \leq \overline{t} < \infty$. Then, after at most*

$$\overline{k} = \lceil \frac{R^2(\overline{t} + f(x_0) - f^*)}{\eta\overline{\alpha}} \varepsilon^{-2} \rceil$$

iterations, we have

$$E\left[f(x_{\overline{k}}) - f^*\right] \leq \varepsilon,$$

where R is the diameter of D.

Proof. First, notice that (6.16) holds and since $\alpha_k \geq \overline{\alpha}$ we obtain

$$\lim_{k\to\infty} ||\overline{g}_k||^2 = 0.$$

Take arbitrary $\varepsilon > 0$ and define $\varepsilon_1 = \varepsilon/R$. Since $\overline{g}_k$ tends to zero, there exists $\overline{k}$ such that $||\overline{g}_{\overline{k}}|| \leq \varepsilon_1$. Let $\overline{k}$ be the first such iteration. Then for $k = 0, 1, \ldots, \overline{k} - 1$ we have $||\overline{g}_k|| > \varepsilon_1$. Moreover, from (6.12) we get

$$\tilde{e}_k + f(x_k) - f(x_{k+1}) \geq \eta\overline{\alpha}\varepsilon_1^2$$

for $k = 0, 1, \ldots, \overline{k} - 1$ and by summing up both sides of this inequality and using $\sum_{k=0}^{\infty} \tilde{e}_k \leq \overline{t} < \infty$ we obtain

$$\eta\overline{\alpha}\varepsilon_1^2\overline{k} \leq \overline{t} + f(x_0) - f(x_{\overline{k}}) \leq \overline{t} + f(x_0) - f^*,$$

i.e.,

$$\overline{k} \leq \frac{\overline{t} + f(x_0) - f^*}{\varepsilon_1^2 \eta \overline{\alpha}} = \varepsilon^{-2} \frac{R^2(\overline{t} + f(x_0) - f^*)}{\eta \overline{\alpha}}.$$

Since $f_{\mathcal{N}_{k+1}}$ is convex and $\overline{g}_k \in \partial f_{\mathcal{N}_{k+1}}(x_k)$ there holds

$$f_{\mathcal{N}_{\overline{k}+1}}(x^*) \geq f_{\mathcal{N}_{\overline{k}+1}}(x_{\overline{k}}) + \overline{g}_{\overline{k}}^T (x^* - x_{\overline{k}}),$$

i.e.,

$$f_{\mathcal{N}_{\overline{k}+1}}(x_{\overline{k}}) - f_{\mathcal{N}_{\overline{k}+1}}(x^*) \leq \overline{g}_{\overline{k}}^T (x_{\overline{k}} - x^*) \leq ||\overline{g}_{\overline{k}}|| ||x^* - x_{\overline{k}}|| \leq \varepsilon_1 R = \varepsilon.$$

Denote by $\mathcal{F}_{\bar{k}}$ the σ-algebra generated by $x_0, \ldots, x_{\bar{k}}$. Since the sample is assumed to be i.i.d. and the approximate functions are computed as sample average, we obtain

$$E\left[(f(x_{\overline{k}}) - f(x^*)\right] = E\left[E\left[f_{\mathcal{N}_{\overline{k}+1}}(x_{\overline{k}}) - f_{\mathcal{N}_{\overline{k}+1}}(x^*) | \mathcal{F}_{\bar{k}}\right]\right] \leq \varepsilon. \blacksquare$$

Let us conclude this section by considering the finite-sum case which falls into the IR-NS framework. Recall that $h(N_k) \to 0$. So, in the case of finite-sum we have $N_k = N_{\max}$ for all $k \geq k_0$ where k_0 is random but finite. Moreover, $\tilde{e}_k$ becomes zero eventually, so the summability of $\tilde{e}_k$ holds. Furthermore, (6.12) reveals that $f(x_{k+1}) \leq f(x_k)$ for all $k \geq k_0$ and thus the iterations remain in the level set $\mathcal{L} = \{x | f(x) \leq f(x_{k_0})\}$. If the level set is compact then the Assumption A7 is obviously satisfied. Finally, notice that Assumption A2 is needed only to ensure that $\tilde{e}_k$ tends to zero a.s. which is obviously true in the finite-sum case. Also, notice that in the strongly convex finite-sum case, there exists C such that all f_i functions are bounded from below by C. Therefore the following result holds.

Corollary 6.1.1 *Let Assumptions A5 and A6 hold and assume* $\sum_k \alpha_k = \infty$. *If* $f = f_{\mathcal{N}_{max}}$ *and* $f_i, i = 1, ..., N_{max}$ *are continuous and strongly convex, then there exists an accumulation point* x^* *of* $\{x_k\}$ *which is a solution of problem* (4.3). *Moreover, if* $\alpha_k \geq \overline{\alpha} > 0$ *for all* $k \in \mathbb{N}$, *then the worst-case complexity is of order* $\mathcal{O}(\varepsilon^{-2})$.

6.1.3 Numerical Results

In this subsection, we test IR-NS variable sample size scheme on two classes of nonsmooth convex problems:

a) Finite-Sums (FS), i.e., bounded sample size with real-world data,

b) Expected Residual Minimization (ERM) reformulation of Stochastic Linear Complementarity Problems (SLCP) with unbounded sample size and simulated data.

The first class belongs to the machine learning framework and considers L_2-regularized binary hinge loss functions for binary classification as in the previous chapter. The considered data sets are given in Table 6.1 and the unconstrained optimization problem is of the form

$$\min_{x\in\mathbb{R}^n} f(x) := \frac{\lambda}{2}||x||^2 + \frac{1}{N_{max}}\sum_{i=1}^{N_{max}} \max(0, 1 - z_i x^T w_i),$$

where $\lambda = 10^{-5}$ is a regularization constant, $w_i \in \mathbb{R}^n$ are the input features, $z_i \in \{\pm 1\}$ the corresponding labels, N_{max} is the size of the relevant data set (testing or training).

	Data set	N	n	N_{train}	N_{test}	Max_{FEV}
1	SPLICE [105]	3175	60	2540	635	10^6
2	MUSHROOMS [70]	8124	112	6500	1624	10^6
3	ADULT9 [105]	32561	123	26049	6512	10^7
4	MNIST(binary) [106]	70000	784	60000	10000	10^7

Table 6.1: Properties of the data sets used in the experiments.

SLCP consists of finding a vector $x \in \mathbb{R}^n$ such that

$$x \geq 0, M(\xi)x + q(\xi) \geq 0, x^T(M(\xi)x + q(\xi)) = 0, \xi \in \Omega,$$

where Ω is the underlying sample space, $M(\xi) \in \mathbb{R}^{n,n}$ is a random matrix and $q(\xi) \in \mathbb{R}^n$ is a random vector. ERM reformulation (see [57] for example) is defined as follows

$$\min f(x) = E\left[||\tilde{F}(x,\xi)||^2\right], \quad \text{s. t.} \quad x \geq 0,$$

where $\tilde{F}(x,\xi) : \mathbb{R}^n \times \Omega \to \mathbb{R}^n$, $\tilde{F}(x,\xi) = \phi(x, M(\xi)x + q(\xi))$ and $\phi : \mathbb{R}^2 \to \mathbb{R}$ is the NCP function defined as $\phi(a,b) = \min\{a,b\}$.

The SAA approximate objective function (4.2) is defined as

$$f_{\mathcal{N}_k}(x) = \frac{1}{N_k}\sum_{j=1}^{N_k} f_j(x)$$

with $f_j(x) = \|\tilde{F}(x,\xi_j)\|^2 = \sum_{l=1}^{n}\left(\min\{x_l, [M(\xi_j)x]_l + [q(\xi_j)]_l\}\right)^2$.

Since numerical results for deterministic (full sample) problem provided in [102] reveal the advantages of BFGS-type methods in nonsmooth optimization, we chose to use the method proposed therein for finding a descent direction satisfying Assumption A5. The functions in consecutive iterations differ in general, and y_k needed for BFGS update is the difference of subgradients of different SAA functions, a safeguard is needed to ensure that the resulting matrices are uniformly positive definite. Thus we start with the identity matrix and skip the BFGS update if $y_k(x_{k+1} - x_k) < 10^{-4}\|y_k\|^2$. Both types of tested problems, FS and ERM allow us to calculate $\sup_{g\in\partial f_{\mathcal{N}}(x)} p^T g$ which is crucial for finding the descent BFGS direction. We denote the proposed algorithm by IRBFGS to emphasize the fact that the BFGS directions are used.

The parameters of IRBGFS algorithm are $\vartheta_0 = 0.9$, $r = 0.95, \overline{\gamma} = 1$ and $\gamma = 10^{-4}$. We use the function $h(N_k) = \frac{N-N_k}{N}$ for FS and $h(N_k) = \frac{1}{N_k}$ for ERM problem. Thus, we have

$$\tilde{N}_{k+1} = \min\{N, \lceil N - r(N - N_k)\rceil\}$$

for bounded and

$$\tilde{N}_{k+1} = \left\lceil \frac{N_k}{r} \right\rceil$$

for unbounded sample case. $N_0 = \lceil 0.1N \rceil$ for FS, while for ERM problems we take $N_0 = 1000$. Step S3 is performed as already stated: we estimate the sample size lower bound N_{k+1}^{trial} derived from (6.7) and let

$$N_{k+1} \in \{N_{k+1}^{trial}, \lceil (N_{k+1}^{trial} + \tilde{N}_{k+1})/2 \rceil, \tilde{N}_{k+1}\}.$$

The backtracking technique for finding $\alpha_k = 0.5^j$ is used, but at each backtracking step, we try all three candidate values for N_{k+1}. We use cumulative samples, although other approaches are feasible as well. The value N_{k+1}^{trial} is calculated as follows:

a) for FS

$$N_{k+1}^{trial} := N_k + \frac{1-r}{2} \cdot \frac{\tilde{N}_{k+1} - N_k}{1 - \vartheta_{k+1}} - \hat{\vartheta}_{k+1} \left(\gamma\alpha ||p_{k-1}||^2 - f_{\tilde{\mathcal{N}}_{k+1}}(x_k) + f_{\mathcal{N}_k}(x_k) \right),$$

where $\hat{\vartheta}_{k+1} = N \cdot \frac{\vartheta_{k+1}}{1-\vartheta_{k+1}}$;

b) for ERM

$$N_{k+1}^{trial} := \frac{1 - \vartheta_{k+1}}{\frac{1-r}{2} \cdot \frac{N_k - \tilde{N}_{k+1}}{\tilde{N}_{k+1} N_k} + \frac{1-\vartheta_{k+1}}{N_k} + \vartheta_{k+1} \left(\gamma\alpha ||p_{k-1}||^2 - f_{\tilde{\mathcal{N}}_{k+1}}(x_k) + f_{\mathcal{N}_k}(x_k) \right)}.$$

The motivation for these choices comes from condition (6.7) from Step S3. The merit function at the new point should be decreased for at least $\frac{1-r}{2}(h(\tilde{N}_{k+1}) - h(N_k))$. Therefore, approximating $||p_k||$ with $||p_{k-1}||$ and using (6.5) and (6.6) from Step S3, we obtain the lower bound N_{k+1}^{trial} for N_{k+1}. If this value falls below N_0, we simply take $N_{k+1}^{trial} = N_0$.

Our numerical study has two goals:

i) to investigate if the variable sample size approach is beneficial in terms of overall optimization costs;

ii) to investigate if the potential decrease of the sample size coming from S3 is beneficial.

This is why we compare the proposed IRBFGS method to:

a) FBFGS which takes the full sample (when applicable) at each iteration, i.e., in FS problems $N_k = N_{max}$ for each k;

b) HBFGS which takes $N_{k+1} = \tilde{N}_{k+1}$ for each k.

The criterion for comparison is the number of scalar products denoted by FEV. We report the average values of 10 independent runs. The algorithms are stopped when the maximum number of scalar products, Max_{FEV} is reached. In the FS case, we track the value of the (full sample) objective function, while in the ERM case, we track the Euclidean difference between x_k and the solution x^* since the objective function is not computable while the solution is known in advance.

Figure 6.1 shows the results on FS problems with uniform random x_0. Since training and testing errors behave similarly, we report only the testing error. The y-axes are in logarithmic scale. The plots demonstrate the computational savings obtained by IRBFGS in almost all cases. In fact, both subsampled methods, IRBFGS and HBFGS use smaller FEV to obtain solutions of the same quality as the full BFGS - FBFGS. Comparing IRBFGS and HBFGS, one can see that IRBFGS is more efficient, and an occasional decrease of N_k in Step S3 is beneficial in terms of computational effort measured by FEV. The typical behavior of the sample size sequence is plotted in Figure 6.3 (left).

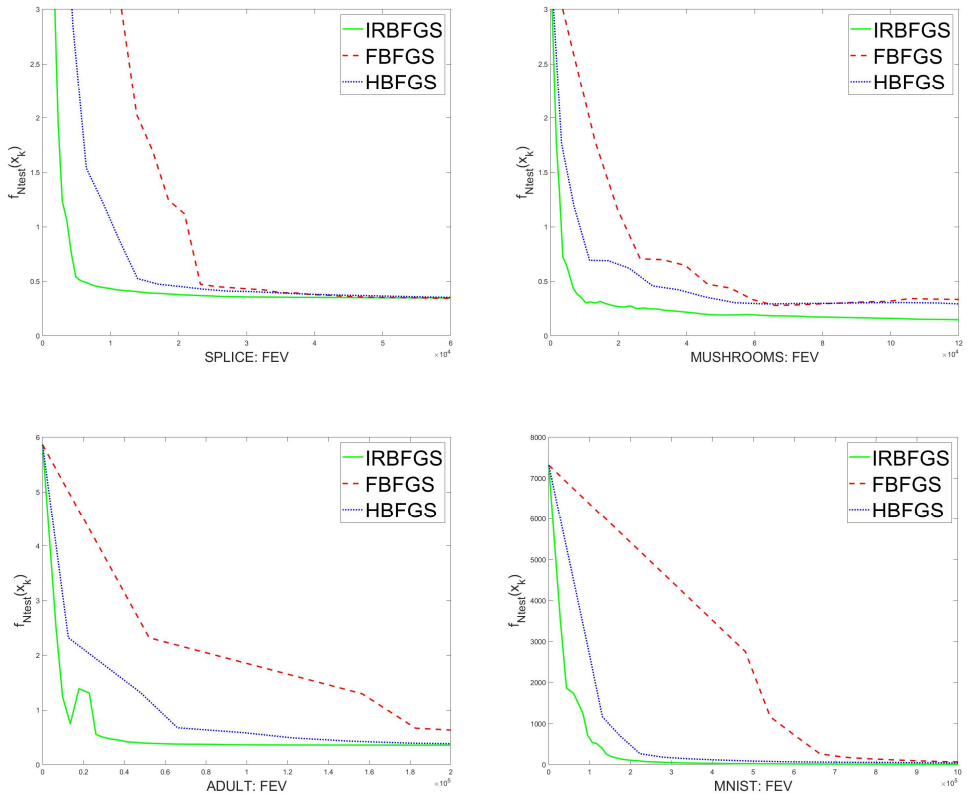

Figure 6.1: FS Problem. Testing loss versus function evaluations.

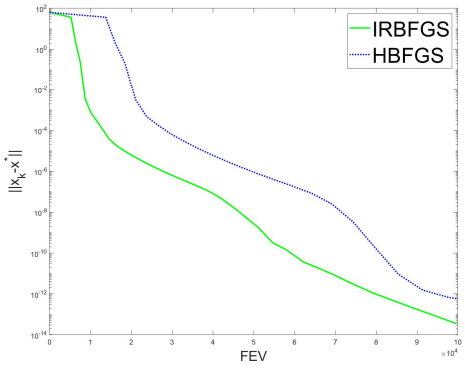

Figure 6.2: ERM Problem. The error $\|x_k - x^*\|$ versus function evaluations

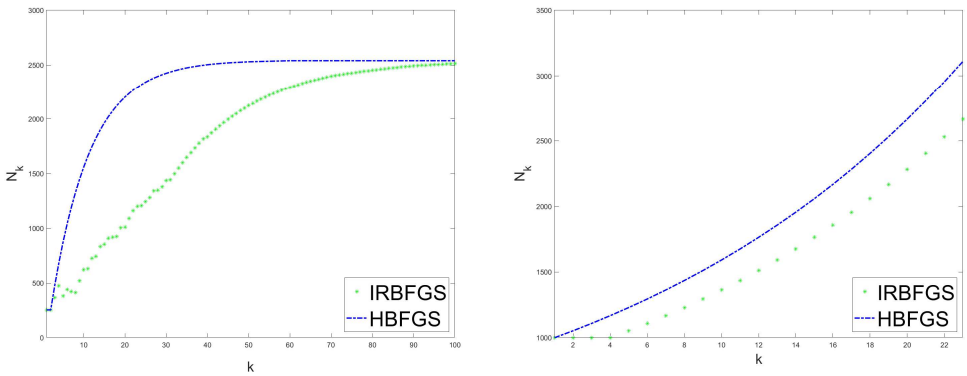

Figure 6.3: IRBFGS sample size versus HBFGS sample size sequence: FS Problem - SPLICE data set (left) and ERM Problem (right).

ERM problems are formed as in [21, 68, 57] where the first-order methods were tested. Here we proceed with the nonsmooth BFGS direction. We report the results for the problem with $n = 100$ and volatility measure $\sigma = 10$. Max_{FEV} is set to 10^5 and the average ending sample size is 4714 for IRBFGS and 3110 for HBFGS. The results and typical behavior of the sample size sequence are presented in Figures 6.2 and 6.3 (right), respectively. As we can see, IRBFGS algorithm significantly outperforms the heuristic scheme HBFGS.

Chapter 7

Conclusions

Through extensive research and analysis, several key findings and contributions have been made, which have implications for both theory and practice. The research conducted in this thesis has deepened our understanding of numerical methods and optimization techniques in the context of nonsmooth problems. The developed algorithms (Algorithm 4, 5, 6, 7) have demonstrated their effectiveness in handling nonsmooth objective functions given in the form of mathematical expectation, providing valuable insights into the behavior and convergence properties of these methods. As an exact evaluation of the expected value is either impossible or prohibitively expensive, subsampling is employed to bypass this difficulty. For both types of problems, constrained and unconstrained, in each iteration, all proposed procedures use a SAA function instead of the mathematical expectation function, and employ the advantages of the variable sample size method based on adaptive updating of the sample size. Two iterative procedures (SPS and AN-SPS) are proposed for solving a constrained optimization problem, while for an unconstrained case, the proposed algorithm is referred to as IR-NS. In the following, we provide a concise summary of the original contributions made in this thesis:

I Spectral Projected Subgradient method

i) The stochastic spectral projected gradient method is adapted to the nonsmooth framework;

ii) The a.s. convergence of the proposed SPS method is proved under the standard assumptions;

iii) The SPS is further upgraded by introducing a specific line search technique resulting in LS-SPS;

iv) Numerical results on machine learning problems show the efficiency of the proposed method, especially LS-SPS.

II Adaptive sample size nonmonotone line search spectral projected subgradient method

i) An adaptive sample size strategy is proposed and we prove that this strategy pushes the sample size to infinity (or to the maximal sample size for finite-sum case);

ii) We show that the scaling can relax the boundedness assumptions on subgradients, iterations, and feasible set;

iii) For finite-sum problems, we provide the worst-case complexity analysis of the proposed method;

iv) The LS-SPS is generalized in a sense that we allow different nonmonotone line search rules. Although important for practical behavior of the algorithm, this change does not effect the convergence analysis and it is investigated mainly through numerical experiments;

v) Considering the spectral coefficient, we investigate different strategies for its formulation in a stochastic framework. Different combinations of spectral coefficients and nonmonotone rules are evaluated within numerical experiments conducted on machine learning Hinge loss problems.

III Inexact restoration nonsmooth method with variable accuracy

i) The general algorithm is defined within Inexact Restoration approach, using a suitable approximate function computed as the sample average approximation in each iteration;

ii) The sample size is determined adaptively, taking into account the progress toward the stationary point and thus balancing the computational cost and accuracy in endogenous way without heuristic elements;

iii) It is proved, using the standard IR methodology, that the sample size tends to infinity or attains the fixed maximal value;

iv) The theoretical analysis reveals a.s. convergence towards stationary points under the set of standard assumptions;

v) The numerical experiments are based on the BFGS direction adapted to the nonsmooth environment [102]. The oracle for computing the direction is taken from literature for the hinge loss problems and Expected Residual Minimization of Stochastic Linear Complementarity Problem. The obtained numerical results are in line with the theoretical considerations and confirm the efficiency of the algorithm.

Bibliography

[1] P. Apkarian, D. Noll, L. Ravanbod, Nonsmooth bundle trust-region algorithm with applications to robust stability, *Set-Valued and Variational Analysis 24(1) (2016), pp. 115 – 148. https://doi.org/10.1007/s11228-015-0352-5*

[2] A. Asl, M. L. Overton, Analysis of limited-memory BFGS on a class of nonsmooth convex functions, *IMA Journal of Numerical Analysis 41(1) (2021), pp. 1-27, https://doi.org/10.1093/imanum/drz052.*

[3] A. Asl, M. L. Overton, Analysis of the gradient method with an Armijo–Wolfe line search on a class of non-smooth convex functions, *Optimization methods and software 35(2) (2020), pp. 223-242, https://doi.org/10.1080/10556788.2019.1673388.*

[4] A. Bagirov, N.Karmitsa, M. Mäkelä, Introduction to Nonsmooth Optimization, *Springer, (2014), https://doi.org/10.1007/978-3-319-08114-4.*

[5] J. Barzilai, J. M. Borwein, Two-point step size gradient method, *IMA J. Numerical Analysis, 8(1) (1988), pp. 141–148, https://doi.org/10.1093/imanum/8.1.141*

[6] F. Bastin, C. Cirillo, P.L. Toint, An adaptive Monte Carlo algorithm for computing mixed logit estimators, *Comput. Manag. Sci. 3(1) (2006), pp. 55-79. https://doi.org/10.1007/s10287-005-0044-y*

[7] S. Bellavia, N. Krejić, N. Krklec Jerinkić, Subsampled Inexact Newton methods for minimizing large sums of convex function, *IMA Journal of Numerical Analysis 40(4) (2018), pp. 2309-2341. https://doi.org/10.1093/imanum/drz027*

[8] S. Bellavia, N. Krejić, B. Morini, Inexact restoration with subsampled trust-region methods for finite-sum minimization, *Computational Optimization and Applications 76 (2020), pp. 701–736, https://doi.org/10.1007/s10589-020-00196-w*

[9] S. Bellavia, N. Krklec Jerinkić, G. Malaspina, Subsampled nonmonotone spectral gradient methods, *Communications in Applied and Industrial Mathematics, 11(1) (2020), pp. 19-34. https://doi.org/10.2478/caim-2020-0002*

[10] E. G. Birgin, N. Krejić, J. M. Martinez, Iteration and evaluation complexity on the minimization of functions whose computation is intrinsically inexact, *Mathematics of Computation 89 (2020), pp. 253-278, https://doi.org/10.1090/mcom/3445.*

[11] E. G. Birgin, N. Krejić, J. M. Martinez, Globaly convergent inexact quasi-Newton methods for solving nonlinear systems, *Numer. Algorithms 32 (2003), pp. 249-260.*

[12] E. G. Birgin, J. M. Martínez, M. Raydan, Nonmonotone spectral projected gradients on convex sets, *SIAM Journal on Optimization. 10 (2000) pp. 1196-1211. https://doi.org/10.1137/S1052623497330963*

[13] S. Boyd, A. Mutapcic, Stochastic subgradient methods, *Lecture Notes for EE364b, Stanford University, (2008).*

[14] S. Bubeck, Convex optimization: Algorithms and complexity, *Found. Trends Mach. Learn. 8(3-4) (2015), pp. 231-357, https://doi.org/10.1561/2200000050.*

[15] L. F. Bueno, J. M. Martinez, On the complexity of an inexact restoration method for constrained optimization, *SIAM Journal on Optimization 30(1) (2020), pp. 80-101, https://doi.org/10.1137/18M1216146*

[16] L. F. Bueno, J. M. Martinez, Inexact restoration for minimization with inexact evaluation both of the objective function and the constraints, *arXiv preprint arXiv:2201.01162, (2022).*

[17] J. V. Burke, A. S. Lewis, M. L. Overton, Approximating subdifferentials by random sampling of gradients, *Mathematics of Operations Research 27(3) (2002), pp. 567-584, https://doi.org/10.1287/moor.27.3.567.317.*

[18] J. V. Burke, A. S. Lewis, M. L. Overton, A robust gradient sampling algorithm for nonsmooth, nonconvex optimization, *SIAM Journal on Optimization 15(3) (2005), pp. 751-779, https://doi.org/10.1137/030601296.*

[19] P. J. Carrington, J. Scott, S. Wasserman, eds., Models and Methods in Social Network Analysis, *Structural Analysis in the Social Sciences, Cambridge University Press (2005), https://doi.org/10.1017/CBO9780511811395.*

[20] V. Cevher, S. Becker, M. Schmidt, Convex Optimization for Big Data: Scalable, randomized, and parallel algorithms

for big data analytics, *IEEE Signal Processing Magazine 31(5) (2014), pp. 32-43, DOI: 10.1109/MSP.2014.2329397.*

[21] X. Chen, C. Zhang, M. Fukushima, Robust solution of monotone stochastic linear complementarity problems, *Springer, Math. Program. 117 (2009), pp. 51-80, https://doi.org/10.1007/s10107-007-0163-z*

[22] F. H. Clarke, Optimization and Nonsmooth Analysis, *John Wiley & Sons, 1983.*

[23] A. R. Conn, N. I. M. Gould, Ph. L. Toint, Trust-region Methods, *SIAM, 2000.*

[24] F. E. Curtis, M. L. Overton, A sequential quadratic programming method for nonconvex, nonsmooth constrained optimization, *SIAM Journal on Optimization 22(2) (2012), pp. 474-500.*

[25] F. E. Curtis, X. Que, An adaptive gradient sampling algorithm for non-smooth optimization, *Optimization Methods and Software 28(6) (2013), pp.1302-1324, DOI: 10.1080/10556788.2012.714781.*

[26] F. E. Curtis, X. Que, A quasi-newton algorithm for nonconvex, nonsmooth optimization with global convergence guarantees, *Mathematical Programming Computation 17(4) (2015), pp. 399-428.*

[27] D. di Serafino, N. Krejić, N. Krklec Jerinkić, M. Viola, LSOS: Line-search Second-Order Stochastic optimization methods for nonconvex finite sums, *arXiv:2007.15966v2 (2021).*

[28] D. Di Serafino, V. Ruggiero, G. Toraldo, L. Zanni, On the steplength selection in gradient methods for unconstrained optimization, *Applied Mathematics and Computation 318 (2018), pp. 176-195, https://doi.org/10.1016/j.amc.2017.07.037.*

[29] Y. H. Dai, On the nonmonotone line search, *Journal of Optimization Theory and Applications 112 (2002), pp. 315–330.*

[30] E. D. Dolan, J. J. Moré, Benchmarking optimization software with performance profiles, *Math. Program., Ser. A 91 (2002), pp. 201-213, https://doi.org/10.1007/s101070100263.*

[31] J. C. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, *J. Mach. Learn. Res. 12 (2011), pp. 2121-2159.*

[32] R. Durrett, Probability: theory and examples, *Vol. 49 (2009), Cambridge university press.*

[33] A. Fischer, A. Friedlander, A new line search inexact restoration approach for nonlinear programming, *Computational Optimization and Applications 46(2) (2010), pp. 333-346, https://doi.org/10.1007/s10589-009-9267-0*

[34] G. Frassoldati , L. Zanni , G. Zanghirati, New adaptive stepsize selections in gradient methods, *J. Ind. Manag. Optim. 4 (2) (2008), pp. 299–312, DOI: 10.3934/jimo.2008.4.299.*

[35] M. P. Friedlander, M. Schmidt, Hybrid deterministic-stochastic methods for data fitting, *SIAM Journal on Scientific Computing 34(3) (2012), pp. 1380-1405, https://doi.org/10.1137/110830629.*

[36] A. P. George, W. B. Powell, Adaptive stepsizes for recursive estimation with applications in approximate dynamic programming, *Mach. Learn. 65 (2006), pp. 167-198, https://doi.org/10.1007/s10994-006-8365-9.*

[37] M. A. Gomes-Ruggiero, J. M. Martínez, S. A. Santos, Spectral projected gradient method with inexact restoration for minimization with nonconvex constraints, *SIAM Journal on Scientific Computing, 31(3) (2009), pp. 1628-1652. https://doi.org/10.1137/070707828*

[38] G. N. Grapiglia, E. W. Sachs, On the worst-case evaluation complexity of non-monotone line search algorithms, *Computational Optimization and applications 68(3) (2017), pp. 555-577, DOI: 10.1007/s10589-017-9928-3.*

[39] L. Grippo, F. Lampariello, S. Lucidi, A nonono-tone line search technique for Newton's method, *SIAM Journal on Numerical Analysis 23(4) (1986), pp. 707-716, https://doi.org/10.1137/0723046.*

[40] L. Grippo, F. Lampariello, S. Lucidi, A truncated Newton method with nonmonotone line search for unconstrained optimization, *J. Optim. Theory Appl. 60 (1989), pp. 401–419.*

[41] J. Guo, A. Lewis, Nonsmooth variants of Powell's BFGS convergence theorem, *SIAM Journal on Optimization 28(2) (2018), pp. 1301-1311.*

[42] W. Hare, C. Sagastizábal, M. Solodov, A proximal bundle method for nonsmooth nonconvex functions with inexact information, *Computational Optimization and Applications, 63(1) (2016), pp. 1-28. https://doi.org/10.1007/s10589-015-9762-4*

[43] P. Hennig, Fast probabilistic optimization from noisy gradients, *in Proceedings of the 30th International Conference on Machine Learning (2013), pp. 62-70.*

[44] T. Homem-de-Mello, Variable-Sample Methods for Stochastic Optimization, *ACM Trans. Model. Comput. Simul. 13(2) (2003), pp. 108–133. https://doi.org/10.1145/858481.858483*

[45] A. N. Iusem, A. Jofré, R. I. Oliveira, P. Thompson, Variance-based extragradient methods with line search for stochastic variational inequalities, *SIAM Journal on Optimization 29(1) (2019), pp. 175–206, https://doi.org/10.1137/17M1144799.*

[46] A. Jalilzadeh, A. Nedić, U. V. Shanbhag, F. Yousefian, A Variable Sample-Size Stochastic Quasi-Newton Method for Smooth and Nonsmooth Stochastic Convex Optimization, *IEEE Conference on Decision and Control (CDC), Miami Beach, FL, (2018), pp. 4097-4102, doi: 10.1109/CDC.2018.8619209.*

[47] Q. Jiang, Proximal Bundle Methods and Nonlinear Acceleration: An Exploration, Senior Thesis, (2021).

[48] J.E. Kelly, The cutting-plane method for solving convex programs, *Journal of the Society for Industrial and Applied Mathematics 8(4) (1960), pp. 703-712.*

[49] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, *in Proceedings of the International Conference on Learning Representations (ICLR) (2015).*

[50] K. C. Kiwiel, An aggregate subgradient method for nonsmooth convex minimization, *Mathematical Programming 27(3) (1983), pp. 320–341.*

[51] K. C. Kiwiel, Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization, *SIAM Journal on Optimization 18(2) (2007), pp. 379-388, https://doi.org/10.1137/050639673.*

[52] N. Krejić, N. Krklec, Line search methods with variable sample size for unconstrained optimization, *Journal of Computational and Applied Mathematics 245 (2013), pp. 213-231, https://doi.org/10.1016/j.cam.2012.12.020.*

[53] N. Krejić, N. Krklec Jerinkić, Nonmonotone line search methods with variable sample size, *Numer. Algorithms 68(4) (2015), pp. 711-739, https://doi.org/10.1007/s11075-014-9869-1.*

[54] N. Krejić, N. Krklec Jerinkić, Spectral projected gradient method for stochastic optimization, *Journal of Global Optimization, 73 (2018), pp. 59–81. https://doi.org/10.1007/s10898-018-0682-6*

[55] N. Krejić, N. Krklec Jerinkić, T. Ostojić, An inexact restoration-nonsmooth algorithm with variable accuracy for stochastic nonsmooth convex optimization problems in machine learning and stochastic linear complementarity problems, *Journal of Computational and Applied Mathematics (2023), 423, 114943.*

[56] N. Krejić, N. Krklec Jerinkić, T. Ostojić, Spectral projected subgradient method for nonsmooth convex optimization problems, *Numerical Algorithms (2022), pp. 1-19.*

[57] N. Krejić, N. Krklec Jerinkić, S. Rapajić, Barzilai-Borwein method with variable sample size for stochastic linear

complementarity problems, *Optimization 65(2) (2016), pp. 479-499, https://doi.org/10.1080/02331934.2015.1062008.*

[58] N. Krejić, N. Krklec Jerinkić, A. Rožnjik, Variable sample size method for equality constrained optimization problems, *Optimization Letters, 12(3) (2018), pp. 485–497, https://doi.org/10.1007/s11590-017-1143-8.*

[59] N. Krejić, Z. Lužanin, Z. Ovcin, I. Stojkovska, Descent direction method with line search for unconstrained optimization in noisy environment, *Optimization Methods and Software 30(6) 2015 , pp. 1164-1184, https://doi.org/10.1080/10556788.2015.1025403.*

[60] N. Krejić, J. M. Martinez, Inexact Restoration approach for minimization with inexact evaluation of the objective function, *Mathematics of Computation 85 (2016), pp. 1775-1791, https://doi.org/10.1090/mcom/3025.*

[61] N. Krklec Jerinkić, Line search methods with variable sample size, *Doctoral dissertation, University of Novi Sad, Serbia, (2014).*

[62] N. Krklec Jerinkić, T. Ostojić, AN-SPS: Adaptive Sample Size Nonmonotone Line Search Spectral Projected Subgradient Method for Convex Constrained Optimization Problems, *arXiv preprint arXiv:2208.10616, (2022).*

[63] N. Krklec Jerinkić, A. Rožnjik, Penalty variable sample size method for solving optimization problems with equality constraints in a form of mathematical expectation, *Numerical Algorithms 83(2) (2020), pp. 701-718, https://doi.org/10.1007/s11075-019-00699-6.*

[64] C. Lemarechal, C. Sagastizabal, Variable metric bundle methods: From conceptual to implementable forms, *Mathematical Programming 76(3) (1997), pp. 393 – 410. https://doi.org/10.1007/BF02614390*

[65] A. Lewis, M. Overton, Nonsmooth optimization via BFGS, *SIAM J. Optimiz. (2009), pp. 1-35.*

[66] A. Lewis, M. Overton, Nonsmooth optimization via quasi-Newton methods, *Mathematical Programming 141 (2013), pp. 135–163.*

[67] A. Lewis, S. Zhang, Nonsmoothness and a variable metric method, *Journal of Optimization Theory and Applications 165 (2015), pp. 151–171.*

[68] X. Li, H. Liu, X. Sun, Feasible smooth method based on Barzilai-Borwein method for stochastic linear complementarity problem, *Numer. Algorithms 57 (2011), pp. 207-215, https://doi.org/10.1007/s11075-010-9424-7*

[69] D. H. Li, M. Fukushima, A derivative-free line search and global convergence of Broyden-like method for nonlinear equations, *Opt. Methods Software 13 (2000), pp. 181-201, DOI:10.1080/10556780008805782.*

[70] M. Lichman, UCI machine learning repository, *https://archive.ics. uci.edu/ml/index.php, (2013).*

[71] M. Loreto, A. Crema, Convergence analysis for the modified spectral projected subgradient method,*Optimization Letters, 9(5) (2015), pp. 915-929, https://doi.org/10.1007/s11590-014-0792-0.*

[72] M. Loreto, Y. Xu, D. Kotval, A numerical study of applying spectral-step subgradient method for solving nonsmooth unconstrained optimization problems, *Computers & Operations Research, 104 (2019), pp. 90-97, https://doi.org/10.1016/j.cor.2018.12.006.*

[73] L. Lukšan, J. Vlček, A bundle-Newton method for nonsmooth unconstrained minimization, *Mathematical Programming 83(1) (1998), pp. 373 – 391. https://doi.org/10.1007/BF02680566*

[74] M. Mäkelä, Survey of bundle methods for nonsmooth optimization, *Optimization methods and software, 17(1) (2002), pp. 1-29. https://doi.org/10.1080/10556780290027828*

[75] M. Mäkelä, P. Neittaanmaki, Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control, *World Scientific Publishing Co. (1992). https://doi.org/10.1142/1493*

[76] K. Marti, Stochastic optimization methods, *Springer, Heidelberg, third ed. (2015), Applications in engineering and operations research, https://doi.org/10.1007/978-3-662-46214-0.*

[77] J. M. Martinez, E. A. Pilotta, Inexact restoration algorithms for constrained optimization, *Journal of Optimization Theory and Applications 104 (2000), pp. 135-163, https://doi.org/10.1023/A:1004632923654.*

[78] L. Martinez, R. Andrade, E. G. Birgin, J. M. Martinez, Packmol: A package for building initial configurations for molecular dynamics simulations, *Journal of Computational Chemistry, 30 (2009), pp. 2157-2164, https://doi.org/10.1002/jcc.21224.*

[79] K. Miettinen, Nonlinear Multiobjective Optimization, *Springer, (1998). https://doi.org/10.1007/978-1-4615-5563-6*

[80] R. Mifflin, A modification and an extension of Lemarechal's algorithm for nonsmooth optimization, *Mathematical Programming Studies 17 (1982), pp. 77 – 90. https://doi.org/10.1007/BFb0120960*

[81] A. S. Nemirovsky, D. B. Yudin, Problem complexity and method efficiency in optimization, *Wiley, New York (1983), https://doi.org/10.1137/1027074.*

[82] J. Nocedal, S. J. Wright, Numerical Optimization, 2nd ed., *Springer Ser. Oper. Res. Financ. Eng., Springer, New York, (2006).*

[83] W. D. Oliveira, C. Sagastizábal, Bundle methods in the XXIst century: A bird's-eye view, *Pesquisa Operacional, 34 (2014), pp. 647-670. https://doi.org/10.1590/0101-7438.2014.034.03.0647*

[84] C. Paquette, K. Scheinberg, A stochastic line search method with expected complexity analysis, *SIAM Journal on Optimization 30(1) (2020), pp. 349-376, https://doi.org/10.1137/18M1216250.*

[85] B. Polyak, Introduction to Optimization, *Optimization Software, Inc., (1987).*

[86] D. Rajter-Ćirić, Verovatnoća, *Univerzitet u Novom Sadu, Prirodno-matematički fakultet, 2009.*

[87] H. Robbins, D. Siegmund, A convergence theorem for non negative almost supermartingales and some applications, *In Op-*

timizing methods in statistics (1971), pp. 233-257, Academic Press, https://doi.org/10.1016/B978-0-12-604550-5.50015-8.

[88] R. Tyrrell Rockafellar, Convex Analysis, *Princton University Press, 1970.*

[89] C. Sagastizabal, M. Solodov, An infeasible bundle method for nonsmooth convex constrained optimization without penalty function or a filter, *SIAM Journal on Optimization, 140(1) (2005), pp. 146 – 169. https://doi.org/10.1137/040603875*

[90] T. Schaul, S. Zhang, Y. LeCun, No more pesky learning rates, *In International Conference on Machine Learning PMLR (2013, May), pp. 343-351.*

[91] S. Shalev-Shwartz, Y. Singer, N. Srebro, A. Cotter, Pegasos: primal estimated sub-gradient solver for SVM, *Mathematical programming, 127(1) (2011), pp. 3-30. https://doi.org/10.1007/s10107-010-0420-4*

[92] A. Shapiro, D. Dentcheva, A. Ruszczynski, Lectures on Stochastic Programming: Modeling and Theory, *MPS-SIAM Series on Optimization (2009). https://doi.org/10.1137/1.9780898718751*

[93] N. Shor, Minimization Methods for Non-differentiable Functions, *Springer Series in Computational Mathematics, Springer, (1985).*

[94] J. C. Spall, Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control, *John Wiley & Sons (2005).*

[95] C. Tan, S. Ma, Y. H. Dai, Y.Qian, Barzilai-borwein step size for stochastic gradient descent, *Advances*

in neural information processing systems, 29, (2016). https://doi.org/10.48550/arXiv.1605.04131

[96] R. Tavakoli, H. Zhang, A nonmonotone spectral projected gradient method for large-scale topology optimization problems, *Numerical Algebra, Control and Optimization Vol. 2, No. 2 (2012), pp. 395-412.*

[97] J. B. H. Urruty, C. Lemarechal, Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods, *Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 1996.*

[98] D. Vicari, A. Okada, G. Ragozini, C. Weihs, eds., Analysis and Modeling of Complex Data in Behavioral and Social Sciences, *Springer, Cham, (2014), https://doi.org/10.1007/978-3-319-06692-9.*

[99] Y. Wardi, Stochastic algorithms with Armijo stepsizes for minimization of functions, *Journal of Optimization Theory and Applications 64 (1990), pp. 399–417, https://doi.org/10.1007/BF00939456.*

[100] S. Wright, B. Recht, Nonsmooth Optimization Methods, *In Optimization for Data Analysis, Cambridge: Cambridge University Press (2022), pp. 153-169, doi:10.1017/9781009004282.010*

[101] D. Yan, H. Mukai, Optimization Algorithm with Probabilistic Estimation, *Journal of Optimization Theory and Applications 64, 79(2) (1993), pp. 345-371, https://doi.org/10.1007/BF00940585.*

[102] J. Yu, S. Vishwanathan, S. Guenter, N. Schraudolph, A Quasi-Newton Approach to Nonsmooth Convex Optimization

Problems in Machine Learning, *Journal of Machine Learning Research 11 (2010), pp. 1145-1200.*

[103] H. Zhang, W. W. Hager, A nonmonotone line search technique and its application to unconstrained optimization, *SIAM Journal on Optimization 4 (2004), pp. 1043-1056, https://doi.org/10.1137/S1052623403428208.*

[104] B. Zhou, L. Gao, Y.H. Dai, Gradient methods with adaptive step-sizes, *Comput. Optim. Appl. 35 (1) (2006), pp.69–86, https://doi.org/10.1007/s10589-006-6446-0.*

[105] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/binary.html

[106] http://yann.lecun.com/exdb/mnist/

[107] https://github.com/ritchie-xl/Bundle-Method-Matlab/blob/master/bundle_method.m

Short Biography

Tijana Ostojić (Stojančević) was born on February 27, 1991, in Sombor, Serbia. She attended the elementary school "Dositej Obradović" and the high school "Veljko Petrović". In 2009, she enrolled in the Bachelor program Applied Mathematics at the Faculty of Sciences, University of Novi Sad, which she completed in 2012, with the average grade 10.00. In 2014, she received her Master's degree in mathematics (average grade 10.00) and started her PhD studies in the field of numerical mathematics at the same university. Since 2014, Tijana has held a position as a teaching assistant at the Faculty of Technical Sciences, Department of Fundamental Sciences, Chair for Mathematics. In this role, she has been involved in teaching various undergraduate mathematics courses across different engineering programs. During her PhD studies, Tijana actively participated in the research project titled *"Numerical Methods, Simulations, and Applications"*. Additionally, she has been engaged in two bilateral projects, one with Croatia titled as "*Calculus of Variations, Optimization, and Applications*" and another with Italy focused on "*Second-Order Methods for Optimization Methods in Machine Learning*". These projects received support from the Ministry of Education, Science, and Technological Development, Republic of Serbia. She has also attended several international conferences, further enriching her academic and research experience.

Овај Образац чини саставни део докторске дисертације, односно докторског уметничког пројекта који се брани на Универзитету у Новом Саду. Попуњен Образац укоричити иза текста докторске дисертације, односно докторског уметничког пројекта.

План третмана података

Назив пројекта/истраживања
Модификације метода Њутновог типа за решавање семи-глатких проблема стохастичке оптимизације Modifications of Newton-type methods for solving semi-smooth stochastic optimization problems
Назив институције/институција у оквиру којих се спроводи истраживање
Универзитет у Новом Саду Природно-математички факултет
Назив програма у оквиру ког се реализује истраживање
Докторске студије математике
1. Опис података
У овој студији нису прикупљани подаци
2. Прикупљање података
3. Третман података и пратећа документација
4. Безбедност података и заштита поверљивих информација
5. Доступност података
6. Улоге и одговорност