



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



Additional Sampling Methods for Finite Sum Constrained Problems

- PhD thesis -

Metode dodatnog uzorkovanja za probleme konačnih suma sa ograničenjima

- Doktorska disertacija -

Supervisor:
Prof. Dr. Nataša Krklec
Jerinkić

Candidate:
Nemanja Vučićević

Novi Sad, 2026

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА¹

Врста рада:	Докторска дисертација
Име и презиме аутора:	Немања Вучићевић
Ментор (титула, име, презиме, звање, институција):	Др Наташа Крклец Јеринкић, редовни професор, Природно-математички факултет, Универзитет у Новом Саду
Наслов рада:	Методе додатног узорковања за проблеме коначних сума са ограничењима
Језик и писмо рада:	Енглески језик
Физички опис рада:	Унети број: Страница 203 Поглавља 8 Референци 76 Табела 1 Слика 0 Графикана 14 Прилога 0
Научна област:	Математика
Ужа научна област (научна дисциплина):	Нумеричка математика
Кључне речи / предметна одредница:	Нелинеарна оптимизација, оптимизација са ограничењима, методи линијског претраживања, немонотон линијско претраживање, апроксимација узорачком средином, додатно узорковање, стохастичка оптимизација
Резиме на српском језику:	Развој и имплементација савремених модела у областима вештачке интелигенције, анализе великих података, машинског учења, инжењерског рачунања, природно имплицирају решавање комплексних оптимизационих проблема великих димензија. У овим применама, оптимизациони процеси су директно условљени специфичним окружењем и великим обимом података, што у пракси генерише високе рачунске захтеве и ограничава ефикасност стандардних конвенционалних метода. Поред саме сложености функције циља, реални модели захтевају увођење експлицитних ограничења на променљиве, као што су нелинеарне једнакости или ограничења са доњом и горњом границом. У оваквим условима, израчунавање тачних вредности функције и њених извода на целом узорку података у свакој итерацији постаје рачунски неодрживо, што намеће потребу за развојем ефикасних стохастичких метода и стратегија адаптивног узорковања које

¹ Аутор докторске дисертације потписао је и приложио следеће Обрасце:

5б – Изјава о ауторству;

5в – Изјава о истоветности штампане и електронске верзије докторског рада и дозвола за објављивање личних података;

5г – Изјава о коришћењу.

Ове Изјаве се чувају у институцији у штампаном и електронском облику и не кориче се са радом.

	балансирају између тачности апроксимације и рачунских трошкова. Главни циљ ове дисертације јесте формулација, теоријска анализа и нумеричка верификација нове класе стохастичких алгоритама првог реда заснованих на механизму додатног узорковања за решавање оптимизационих проблема са нелинеарним једнакостима и ограничењима са доњом и горњом границом. Предложени методолошки оквир успешно превазилази велики број израчунавања у раним фазама оптимизације коришћењем мањих подузорака, док истовремено гарантује асимптотску тачност и конвергенцију. У оквиру проблема са нелинеарним ограничењима једнакости, развијен је алгоритам ASPEN (Additional Sampling Penalty method for Equality Nonlinear constraints). Метод комбинује приступ квадратне казнене функције и немонотону линијску претрагу Армижовог типа са иновативним тестом квалитета на независном, помоћном узорку. Доказана је скоро сигурна конвергенција генерисаног низа тачака ка Каруш-Кун-Такеровим тачкама уз претпоставку услова линеарне независности ограничења. Такође, успешно је изведено проширење овог алгоритма на оквир тежинских коначних сума. За класу проблема са ограничењима са доњом и горњом границом, формулисани су алгоритми AS-NC (за ненегативна ограничења) и AS-BOX (за општа ограничења са доњом и горњом границом). Ове методе користе пројектоване градијентне правце и уводе оригиналан приступ праћењу структуралне конзистентности активног скупа. Поред доказа скоро сигурне конвергенције ка стационарним тачкама, изведена је оцена очекиване рачунске сложености реда, док је под претпоставком строге конвексности доказана скоро сигурна конвергенција ка јединственом решењу. Свеобухватна практична ефикасност предложених алгоритама верификована је кроз детаљне нумеричке експерименте на реалним и синтетичким скуповима података високих димензија из колекција LIBSVM и CUTEst. Испитивани рачунски оквир обухвата решавање сложених конвексних и неконвексних проблема, у распону од логистичке регресије под нелинеарним и ограничењима са доњом и горњом границом, преко обучавања вештачких неуронских мрежа, до анализе осетљивости алгоритама на ниво хетерогености у подацима. Рачунски учинак, евалуиран преко кумулативног броја скаларних производа, удаљености од решења и пројектоване мере стационарности, недвосмислено потврђује да развијени механизми адаптивног и додатног узорковања остварују драстичне уштеде ресурса. Показано је да предложени алгоритми обезбеђују изузетну робусност и супериорне перформансе у поређењу са детерминистичким методама, базичним хеуристичким стохастичким стратегијама, као и савременим, специјализованим алгоритмима унутрашњих тачака и стохастичког секвенцијалног квадратног програмирања.
Датум прихватања теме од стране надлежног већа:	26.2.2026.
Датум одбране: (Попуњава накнадно институција)	
Чланови комисије: (титула, име, презиме, звање, институција)	Председник: др Наташа Крејић, редовни професор, Природно-математички факултет, Универзитет у Новом Саду Члан: др Наташа Крклец Јеринкић, редовни професор, Природно-математички факултет, Универзитет у Новом Саду Члан: др Сања Рапајић, редовни професор, Природно-математички факултет, Универзитет у Новом Саду

	Члан: др Андреа Рожњик, ванредни професор, Грађевински факултет Суботица, Универзитет у Новом Саду
Напомена:	

**UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCES**

KEY WORD DOCUMENTATION²

Document type:	Doctoral dissertation
Author:	Nemanja Vučićević
Supervisor (title, first name, last name, position, institution)	Nataša Krklec Jerinkić, PhD, Full Professor, Faculty of Sciences, University of Novi Sad
Thesis title in English:	Additional Sampling Methods for Finite Sum Constrained Problems
Language and script:	English
Physical description:	Number of: Pages 203 Chapters 8 References 76 Tables 1 Illustrations 0 Graphs 14 Appendices 0
Scientific field:	Mathematics
Scientific subfield (scientific discipline):	Numerical mathematics
Subject, Key words:	Nonlinear optimization, constrained optimization, line search methods, nonmonotone line search, sample average approximation, additional sampling, stochastic optimization
Abstract in English:	The development and implementation of modern models in the fields of artificial intelligence, big data analysis, machine learning, and engineering computing inherently imply solving complex, large-scale optimization problems. In these applications, optimization processes are directly conditioned by the specific environment and the massive volume of data, which in practice generates high computational demands and limits the efficiency of standard conventional methods. In addition to the inherent complexity of the objective function, real-world models require the introduction of explicit constraints on variables, such as nonlinear equalities or bound constraints. Under such conditions, evaluating the exact values of the function and its derivatives over the entire data sample at each iteration becomes computationally unsustainable, necessitating the development of efficient stochastic methods and adaptive sampling strategies that balance approximation accuracy against computational costs. The main objective of this dissertation is the formulation, theoretical analysis,

² The author of the doctoral dissertation has signed the following Statements:

56 – Statement on the authorship,

5B – Statement that the printed and e-version of the doctoral dissertation are identical and authorization to use personal data,

5r – Copyright statement.

The paper and e-versions of Statements are held at the institution and are not included into the printed thesis.

	and numerical verification of a new class of first-order stochastic algorithms based on an additional sampling mechanism for solving optimization problems with nonlinear equalities and bound constraints. The proposed methodological framework successfully alleviates a high computational cost in the early stages of optimization by utilizing smaller subsamples, while simultaneously guaranteeing asymptotic accuracy and convergence. Within the framework of problems with nonlinear equality constraints, the ASPEN (Additional Sampling Penalty method for Equality Nonlinear constraints) algorithm is developed. The method combines a quadratic penalty approach and a nonmonotone Armijo-type line search with an innovative quality test on an independent, auxiliary sample. Almost sure convergence of the generated sequence of iterates to Karush–Kuhn–Tucker points is proved under the Linear Independence Constraint Qualification. Furthermore, an extension of this algorithm to a weighted finite sum framework is successfully derived. For the class of problems with bound constraints, the AS-NC (for nonnegativity constraints) and AS-BOX (for general bound constraints) algorithms are formulated. These methods utilize projected gradient directions and introduce an original approach to monitoring the structural consistency of the active set. In addition to proving almost sure convergence to stationary points, an expected computational complexity bound is derived, while almost sure convergence to the unique solution is established under the assumption of strong convexity. The comprehensive practical efficiency of the proposed algorithms is verified through detailed numerical experiments on real and synthetic large-scale datasets from the LIBSVM and CUTEst repositories. The investigated computational framework encompasses solving complex convex and nonconvex problems, ranging from logistic regression under nonlinear and box constraints, to artificial neural network training, and the sensitivity analysis of the algorithms with respect to data heterogeneity. The computational performance, evaluated via the cumulative number of scalar products, the distance to the solution, and the projected stationarity measure, unequivocally confirms that the developed adaptive and additional sampling mechanisms achieve drastic resource savings. It is shown that the proposed algorithms provide exceptional robustness and superior performance compared to deterministic methods, baseline heuristic stochastic strategies, as well as state-of-the-art, specialized interior-point and stochastic sequential quadratic programming algorithms.
Date of endorsement by the scientific board:	26.2.2026.
Date of defence: (Filled in by the institution)	
Thesis defence board: (title, first name, last name, position, institution)	Chair: Nataša Krejić, PhD, Full Professor, Faculty of Sciences, University of Novi Sad Member: Nataša Krklec Jerinkić, PhD, Full Professor, Faculty of Sciences, University of Novi Sad Member: Sanja Rapajić, PhD, Full Professor, Faculty of Sciences, University of Novi Sad Member: Andrea Rožnjik, PhD, Associate Professor, Faculty of Civil Engineering Subotica, University of Novi Sad
Note:	

Abstract

The development and implementation of modern models in the fields of artificial intelligence, big data analysis, machine learning, and engineering computing inherently imply solving complex, large-scale optimization problems. In these applications, optimization processes are directly conditioned by the specific environment and the massive volume of data, which in practice generates high computational demands and limits the efficiency of standard conventional methods. In addition to the inherent complexity of the objective function, real-world models require the introduction of explicit constraints on variables, such as nonlinear equalities or bound constraints. Under such conditions, evaluating the exact values of the function and its derivatives over the entire data sample at each iteration becomes computationally unsustainable, necessitating the development of efficient stochastic methods and adaptive sampling strategies that balance approximation accuracy against computational costs.

The main objective of this dissertation is the formulation, theoretical analysis, and numerical verification of a new class of first-order stochastic algorithms based on an additional sampling mechanism for solving optimization problems with nonlinear equalities and bound constraints. The proposed methodological framework successfully alleviates a high computational cost in the early stages of optimization by utilizing smaller subsamples, while simultaneously guaranteeing asymptotic accuracy and convergence.

Within the framework of problems with nonlinear equality constraints, the ASPEN (Additional Sampling Penalty method for Equality Nonlinear constraints) algorithm is developed. The method combines a quadratic penalty approach and a nonmonotone Armijo-type line search with an innovative quality test on an independent, aux-

iliary sample. Almost sure convergence of the generated sequence of iterates to Karush–Kuhn–Tucker points is proved under the Linear Independence Constraint Qualification. Furthermore, an extension of this algorithm to a weighted finite sum framework is successfully derived. For the class of problems with bound constraints, the AS-NC (for nonnegativity constraints) and AS-BOX (for general bound constraints) algorithms are formulated. These methods utilize projected gradient directions and introduce an original approach to monitoring the structural consistency of the active set. In addition to proving almost sure convergence to stationary points, an expected computational complexity bound is derived, while almost sure convergence to the unique solution is established under the assumption of strong convexity.

The comprehensive practical efficiency of the proposed algorithms is verified through detailed numerical experiments on real and synthetic large-scale datasets from the LIBSVM and CUTEst repositories. The investigated computational framework encompasses solving complex convex and nonconvex problems, ranging from logistic regression under nonlinear and box constraints, to artificial neural network training, and the sensitivity analysis of the algorithms with respect to data heterogeneity. The computational performance, evaluated via the cumulative number of scalar products, the distance to the solution, and the projected stationarity measure, unequivocally confirms that the developed adaptive and additional sampling mechanisms achieve drastic resource savings. It is shown that the proposed algorithms provide exceptional robustness and superior performance compared to deterministic methods, baseline heuristic stochastic strategies, as well as state-of-the-art, specialized interior-point and stochastic sequential quadratic programming algorithms.

Apstrakt

Razvoj i implementacija savremenih modela u oblastima veštačke inteligencije, analize velikih podataka, mašinskog učenja, inženjerskog računanja, prirodno impliciraju rešavanje kompleksnih optimizacionih problema velikih dimenzija. U ovim primenama, optimizacioni procesi su direktno uslovljeni specifičnim okruženjem i velikim obimom podataka, što u praksi generiše visoke računске zahteve i ograničava efikasnost standardnih konvencionalnih metoda. Pored same složenosti funkcije cilja, realni modeli zahtevaju uvođenje eksplicitnih ograničenja na promenljive, kao što su nelinearne jednakosti ili ograničenja sa donjom i gornjom granicom. U ovakvim uslovima, izračunavanje tačnih vrednosti funkcije i njenih izvoda na celom uzorku podataka u svakoj iteraciji postaje računski neodrživo, što nameće potrebu za razvojem efikasnih stohastičkih metoda i strategija adaptivnog uzorkovanja koje balansiraju između tačnosti aproksimacije i računskih troškova.

Glavni cilj ove disertacije jeste formulacija, teorijska analiza i numerička verifikacija nove klase stohastičkih algoritama prvog reda zasnovanih na mehanizmu dodatnog uzorkovanja za rešavanje optimizacionih problema sa nelinearnim jednakostima i ograničenjima sa donjom i gornjom granicom. Predloženi metodološki okvir uspešno prevazilazi veliki broj izračunavanja u ranim fazama optimizacije korišćenjem manjih poduzoraka, dok istovremeno garantuje asimptotsku tačnost i konvergenciju.

U okviru problema sa nelinearnim ograničenjima jednakosti, razvijen je algoritam ASPEN (Additional Sampling Penalty method for Equality Nonlinear constraints). Metod kombinuje pristup kvadratne kaznene funkcije i nemonotonu linijsku pretragu Armizovog tipa sa inovativnim testom kvaliteta na nezavisnom, pomoćnom uzorku. Dokazana je skoro sigurna konvergencija generisanog niza tačaka ka Karuš-

Kun-Takеровим tačkama uz pretpostavku uslova linearne nezavisnosti ograničenja. Takođe, uspešno je izvedeno proširenje ovog algoritma na okvir težinskih konačnih suma. Za klasu problema sa ograničenjima sa donjom i gornjom granicom, formulisani su algoritmi AS-NC (za nenegativna ograničenja) i AS-BOX (za opšta ograničenja sa donjom i gornjom granicom). Ove metode koriste projektovane gradijentne pravce i uvode originalan pristup praćenju strukturalne konzistentnosti aktivnog skupa. Pored dokaza skoro sigurne konvergencije ka stacionarnim tačkama, izvedena je ocena očekivane računске složenosti reda, dok je pod pretpostavkom stroge konveksnosti dokazana skoro sigurna konvergencija ka jedinstvenom rešenju.

Sveobuhvatna praktična efikasnost predloženih algoritama verifikovana je kroz detaljne numeričke eksperimente na realnim i sintetičkim skupovima podataka visokih dimenzija iz kolekcija LIBSVM i CUTEst. Ispitivani računski okvir obuhvata rešavanje složenih konveksnih i nekonveksnih problema, u rasponu od logističke regresije pod nelinearnim i ograničenjima sa donjom i gornjom granicom, preko obučavanja veštačkih neuronskih mreža, do analize osetljivosti algoritama na nivo heterogenosti u podacima. Računski učinak, evaluiran preko kumulativnog broja skalarnih proizvoda, udaljenosti od rešenja i projektovane mere stacionarnosti, nedvosmisleno potvrđuje da razvijeni mehanizmi adaptivnog i dodatnog uzorkovanja ostvaruju drastične uštede resursa. Pokazano je da predloženi algoritmi obezbeđuju izuzetnu robusnost i superiorne performanse u poređenju sa determinističkim metodama, bazičnim heurističkim stohastičkim strategijama, kao i savremenim, specijalizovanim algoritmima unutrašnjih tačaka i stohastičkog sekvencijalnog kvadratnog programiranja.

Acknowledgments

At the end of this long and significant journey, I would like to express my sincere gratitude to all those who, through their knowledge, support, understanding, and trust, helped me bring this doctoral dissertation to completion.

First and foremost, I owe special gratitude to my supervisor, Prof. Dr. Nataša Krklec Jerinkić, for her expert guidance, dedication, patience, and support throughout the preparation of this doctoral dissertation. Her advice, experience, and clear guidance were of exceptional importance in shaping and completing this research. I am especially grateful for her openness to discussion, exchange of ideas, and readiness to provide support and direction whenever needed.

I also owe great gratitude to Prof. Dr. Nataša Krejić for her trust, support, and valuable advice over the past years. Thanks to her guidance, I had the opportunity to enter a field in which I had not previously worked, to become part of a research group, and to acquire important knowledge and scientific experience through collaborative work. I particularly appreciate her availability, openness to discussion, and willingness to contribute to my scientific development through her experience and ideas.

I would like to thank the members of the committee, Prof. Dr. Sanja Rapajić and Prof. Dr. Andrea Rožnjik, for carefully reading the dissertation, for their useful suggestions and comments, and for the time they devoted to evaluating my work. Their remarks and recommendations contributed to the quality of the final version of the dissertation.

I owe special gratitude to Prof. Dr. Nenad Stojanović, Dr. Aleksandar Milenković, and Prof. Dr. Marina Svičević for their sincere support, understanding, advice, help, and friendship over the past years. Their closeness, collegiality, and constant willingness to be a source of support have meant a great deal to me, both professionally

and personally. I also sincerely thank Prof. Dr. Marija Stanić, who made the initial contact and directed me to PhD, thereby providing me with important support in my further academic and research work. I am also grateful to all my colleagues at the Institute of Mathematics and Informatics, University of Kragujevac, among whom I have many dear friends, as well as to the professors whose advice, support, and stimulating working atmosphere significantly eased the period of my doctoral studies and scientific research.

My deepest gratitude goes to my family, my parents, sisters, grandmothers, and grandfathers, especially to my grandmother Olga, for her support in my education from my earliest days. I thank them for their patience, faith in me, and understanding in the moments when working on the dissertation required a great deal of time, energy, and sacrifice.

A special thank you goes to my wife Milica, for her love, support, patience, and understanding. Her faith in me, encouragement, and presence were my support in the most demanding moments and one of the most important sources of strength that helped me persevere to the end.

To everyone who contributed in any way to the creation of this dissertation, I offer my sincere thanks.

Nemanja Vučićević
Novi Sad, June, 2026

Contents

Abstract	8
Acknowledgments	12
Introduction	18
1 Mathematical Preliminaries	29
1.1 Elements of Linear Algebra and Functional Analysis . .	29
1.2 Probabilistic Foundations for Stochastic Optimization .	38
2 Nonlinear Optimization	47
2.1 Optimization Models and Basic Concepts	48
2.2 Unconstrained Optimization and Descent Directions . .	49
2.3 Line Search Globalization	52
2.4 Nonmonotone Line Search Techniques	54
2.5 Constrained Nonlinear Optimization	56
2.6 Optimality Conditions, Constraint Qualifications, and the KKT System	59
2.7 Sequential Quadratic Programming	62
2.8 Projected Gradient Methods	64

3	Stochastic Optimization	67
3.1	Stochastic Optimization Models	68
3.2	Sample Average Approximation and finite sum Reformulations	70
3.3	Sampling, Mini-Batches, and Estimator Quality	74
3.4	Stochastic First-Order Schemes and Variance Control .	76
3.5	Adaptive Sample Size, Additional Sampling, and Stochastic Step Acceptance	77
3.6	Constrained Stochastic Optimization	81
3.7	Remarks Specific to the Dissertation Setting	85
4	Problem Formulation and Research Motivation	87
4.1	Research Motivation and Problem Classes	88
4.2	Application Settings and Modeling Motivation	91
4.3	Sampled Models, and Standing Assumptions	95
5	Additional Sampling Method for Equality Constraints	99
5.1	Problem formulation	100
5.2	ASPEN algorithm	101
5.3	Convergence analysis	106
5.4	Extension to the weighted finite sum setting	121
6	Additional Sampling Method for Bound-Constrained Problems	127
6.1	Methods for Nonnegativity-Constrained Problems . . .	128
6.1.1	Problem formulation	128
6.1.2	AS-NC algorithm	132
6.1.3	Convergence analysis	136
6.2	Box Constraints	149
6.2.1	Problem formulation	149
6.2.2	AS-BOX algorithm	151
6.2.3	Convergence analysis	153

7	Numerical Experiments	162
7.1	Equality-Constrained Problems: ASPEN	162
7.1.1	Computational setting and benchmark suite . .	163
7.1.2	Comparative Performance of ASPEN	166
7.1.3	Comparison with Stochastic SQP-Type Methods	172
7.1.4	Noise sensitivity and adaptive sample size growth	175
7.2	Box-Constrained Problems: AS-BOX	177
7.2.1	Experimental setting and benchmark methods .	178
7.2.2	Logistic regression under box constraints	180
7.2.3	Adaptive subsample growth	182
7.2.4	Neural network training under box constraints .	184
8	Conclusion	188
	Bibliography	190
	Biography	202

Introduction

Optimization problems with large-scale objective functions and explicit constraints arise naturally in many contemporary applications, including machine learning, data analysis, inverse problems, engineering design, and decision-making under uncertainty. In such settings, the objective function is rarely available as a single inexpensive closed-form expression. Much more often, it is defined as a finite sum of a very large number of component functions associated with individual observations, scenarios, or local models, a structure that naturally encompasses models based on mathematical expectations. At the same time, the decision variables are frequently subject to structural restrictions such as equality relations, lower and upper bounds, or nonnegativity requirements. The simultaneous presence of large-scale objective structure and constraints is therefore one of the central mathematical features of modern optimization.

This dissertation is devoted primarily to constrained finite sum optimization problems of this type, alongside their related stochastic frameworks. The focus is on methods that remain computationally efficient even when the number of local objective terms is very large, while still retaining enough accuracy to make reliable decisions about descent, feasibility, and stationarity. More precisely, the dissertation studies optimization models in which the objective function is primarily represented as a large finite sum

$$f(x) = \frac{1}{N} \sum_{i=1}^N f_i(x),$$

which often serves as a practical realization of an underlying expected-

value model

$$F(x) = \mathbb{E}[f(x; \xi)],$$

and where the variables are constrained either by nonlinear equalities or by simple convex restrictions such as box and nonnegativity constraints.

A major difficulty in these problems is that exact objective and gradient evaluations may be prohibitively expensive if they require the use of all N local functions at every iteration. In large-scale settings, this makes classical full sample algorithms unnecessarily costly, especially during the initial phase of the optimization process, when only coarse information is needed to identify a productive search direction. On the other hand, working with overly small samples may lead to inaccurate models, unreliable descent tests, and poor control of feasibility. Hence, the central algorithmic challenge is not only how to construct a step, but also how to decide how much information is actually needed at a given stage of the method.

This balance between computational cost and approximation quality motivates the use of adaptive sample size strategies. In such methods, the amount of data used at each iteration is allowed to change during the run, rather than being fixed in advance. Small samples can significantly reduce the cost of early iterations, while larger samples become valuable near stationary points, where more accurate information is needed to distinguish genuine progress from sampling artefacts. In the constrained setting, however, this issue becomes even more delicate, because the algorithm must simultaneously account for objective decrease, feasibility improvement, and the reliability of the sampled information on which these decisions are based.

The main methodological idea developed in this dissertation is that adaptive sampling should be supported by an additional independent verification mechanism. More precisely, a candidate point is generated by using one sample of the objective function, while an additional

sample is then employed to assess whether the achieved progress is sufficiently representative of the full problem. This principle, referred to as *additional sampling*, plays a dual role. First, it reduces the risk of accepting a step merely because it performs well on the sample that generated it. Second, it provides a natural criterion for deciding whether the current sample size is adequate or should be enlarged. In this way, additional sampling becomes not only a safeguard, but also a structural ingredient in the control of sample size growth.

The dissertation develops this idea in two constrained optimization settings. The first concerns finite sum problems with nonlinear equality constraints. For this class, feasibility is handled through a quadratic penalty framework, and the proposed method combines penalty techniques, nonmonotone line search, adaptive sample size control, and an additional-sampling acceptance test. This leads to the method ASPEN (*Additional Sampling Penalty method for Equality Nonlinear constraints*), which is designed for large-scale equality-constrained problems in which repeated full sample evaluations are too expensive.

The second setting concerns finite sum problems with constraints, with particular emphasis on box and nonnegativity constraints. In this case, feasibility is naturally handled by projection, and the proposed methods are based on projected sampled directions combined with nonmonotone acceptance rules and additional sampling. This gives rise to the methods AS-NC and AS-BOX, which extend the additional-sampling philosophy to the projected-gradient framework and allow one to treat bound-constrained problems efficiently while preserving a clear stationarity structure.

The dissertation is centered on first-order methods; it is situated within the broader landscape of constrained optimization. Classical deterministic frameworks such as penalty methods, projected gradient methods, interior-point methods, and sequential quadratic programming (SQP) provide important conceptual and computational

benchmarks. In particular, SQP-type methods serve as a natural point of comparison in the equality-constrained setting, while projected and interior-point stochastic methods play the same role in the box-constrained case. However, the primary goal here is not to develop new SQP or interior-point algorithms, but rather to investigate how adaptive sampling and additional-sampling principles can be integrated into penalty and projection frameworks in order to obtain methods that are both theoretically justified and computationally attractive in large-scale problems.

The original contribution of this dissertation lies in the construction, analysis, and numerical investigation of such methods. In particular, the dissertation:

1. develops the ASPEN method for nonlinear equality-constrained finite sum optimization by combining quadratic penalization, adaptive sample size control, nonmonotone line search, and additional sampling;
2. develops additional-sampling projection-based methods for simple bound-constrained problems, with separate treatment of non-negativity constraints and general box constraints;
3. establishes almost sure convergence results and corresponding stationarity properties for the proposed methods under suitable regularity assumptions;
4. provides numerical evidence on benchmark machine learning problems showing that the proposed adaptive approaches can achieve favorable performance relative to deterministic, heuristic, and stochastic benchmark methods.

The main original results presented in the dissertation (Chapters 5-7) are based on the works [41, 42]. These papers provide the foundation for the equality-constrained and box-constrained developments,

while the dissertation places them within a broader and coherent theoretical framework, complements them with background material, and presents the results in a unified form.

Organization of the dissertation.

The remaining chapters of the dissertation are organized as follows.

The chapter *Mathematical Preliminaries* introduces the notation and the basic tools from linear algebra, analysis, convexity, projection theory, and probability that are used throughout the dissertation. Its purpose is to provide a self-contained background for the later algorithmic and convergence analysis.

The chapter *Nonlinear Optimization* reviews the deterministic optimization framework relevant for the dissertation. It recalls basic notions of unconstrained and constrained optimization, first- and second-order optimality conditions, line search globalization techniques, non-monotone acceptance rules, and projected-gradient ideas for simple convex sets.

The chapter *Stochastic Optimization* introduces the stochastic and finite sum viewpoint that underlies the later developments. It discusses expected-value models, Sample Average Approximation, mini-batching, estimator quality, adaptive sample size strategies, and the role of additional sampling in stochastic step acceptance.

The chapter *Problem Formulation and Research Motivation* establishes the common modeling framework of the dissertation and explains the reasons for focusing on nonlinear equality constraints and simple bound constraints. It also introduces the notation and standing assumptions used in the later theoretical and numerical sections.

The chapter *Additional Sampling Method for Equality Constraints* develops the ASPEN method for finite sum optimization problems with nonlinear equality constraints. The penalty formulation is introduced, the algorithm is stated, and its convergence properties are analyzed. The weighted finite sum extension is also considered.

The chapter *Additional Sampling Method for Bound-Constrained*

Problems treats finite sum optimization problems with nonnegativity and box constraints. It first introduces the additional-sampling method for nonnegativity constraints and then extends the framework to the general box-constrained case, together with the corresponding convergence and complexity results.

Finally, the chapter *Numerical Experiments* presents the computational study of the proposed methods. For the equality-constrained setting, ASPEN is compared with deterministic and stochastic benchmark approaches on several machine learning datasets, and its adaptive behavior is further examined on a controlled synthetic problem. For the box-constrained setting, AS-BOX is tested on convex and non-convex learning models and compared with projected-gradient and interior-point-type stochastic methods.

In this way, the dissertation moves from general mathematical and algorithmic background toward two specific classes of constrained finite sum problems, and culminates in original methods whose design, analysis, and numerical behavior are all centered around the same guiding principle: reliable optimization with sampled information requires not only cheaper approximations, but also principled mechanisms for deciding when those approximations are sufficiently informative.

List of Figures

7.1	<i>MNIST</i> dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).	167
7.2	<i>Mushrooms</i> dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).	168
7.3	<i>Splice</i> dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).	168
7.4	<i>a9a</i> dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).	169
7.5	<i>Heart</i> dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).	169
7.6	<i>Australian</i> dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).	170
7.7	Relative sample size N_k/N per iteration for ASPEN, Heur, and Full on the considered binary classification data sets.	171

7.8	ASPEN vs. STO-SQP and two variants of SVR-SQP. Distance to the optimal solution versus the number of scalar products (FEV): a) <i>Mushrooms dataset</i> ; b) <i>a9a dataset</i> ; c) <i>Australian dataset</i> ; d) <i>Heart dataset</i> ; e) <i>Splice dataset</i> ; f) <i>MNIST dataset</i>	173
7.9	Influence of dissimilarity of local cost functions (modeled by $\sigma \in \{0.1, 0.5, 1, 2\}$) on the behavior of ASPEN algorithm applied on problem (7.2): a) optimality gap; b) sample size increase.	177
7.10	Distance to the solution $\ x_k - x^*\ $ versus FEV_k for logistic regression on the <i>Mushrooms dataset</i>	181
7.11	Distance to the solution $\ x_k - x^*\ $ versus FEV_k for logistic regression on the <i>IJCNN1 dataset</i>	182
7.12	AS-BOX: evolution of the subsample size N_k as a function of FEV_k . Part a): N_k versus FEV_k for the <i>Mushrooms dataset</i> . Part b): N_k versus FEV_k for the <i>IJCNN1 dataset</i>	183
7.13	Part a): Cross-entropy loss versus FEV_k for the <i>Mushrooms dataset</i> . Part b): Stationarity measure $\ d(x_k)\ $ versus FEV_k for the <i>Mushrooms dataset</i>	186
7.14	Part a): Cross-entropy loss versus FEV_k for the <i>IJCNN1 dataset</i> . Part b): Stationarity measure $\ d(x_k)\ $ versus FEV_k for the <i>IJCNN1 dataset</i>	187

List of Tables

7.1 Binary classification data set details [14]. 165

List of Algorithms

1	Generic SQP framework	64
2	ASPEN : Additional Sampling Penalty Method for Equality Nonlinear Constraints	107
3	AS-NC : Additional Sampling for Nonnegativity Constraints	134
4	AS-BOX : Additional Sampling for BOX Constraints	152

Chapter 1

Mathematical Preliminaries

This chapter introduces the notation and the basic results from analysis, linear algebra, and probability theory that will be used in the formulation and analysis of algorithms for constrained finite sum and stochastic optimization problems. The aim is not to present a complete theoretical treatment, but rather to provide a self-contained foundation: clear definitions, standard inequalities, and a few key lemmas that will later be invoked implicitly, for example in estimates of merit function decrease, stability properties of projections, and the control of sampling-induced noise.

1.1 Elements of Linear Algebra and Functional Analysis

This section fixes the notation and recalls several concepts from linear algebra, finite-dimensional functional analysis, and smooth optimization that will be used throughout the dissertation. The main emphasis is on vector and matrix norms, operator bounds, spectral properties of symmetric matrices, convexity, Lipschitz continuity, and projec-

tion operators onto closed convex sets. These tools are fundamental in the analysis of gradient-based and projection-based algorithms, in particular for deriving descent estimates, controlling curvature, and preserving feasibility under constraints.

Notation. Vectors are denoted by lowercase letters, such as $x, y \in \mathbb{R}^d$, while matrices are denoted by uppercase letters, such as $A \in \mathbb{R}^{m \times n}$. By $A^\top$, we use the notation the transpose of the matrix A . The zero vector or zero matrix is denoted by 0 , whenever the dimension is clear from the context. For a set $\mathcal{S} \subseteq \mathbb{R}^d$, its closure is denoted by $\text{cl}(\mathcal{S})$.

The standard Euclidean inner product is defined by

$$\langle x, y \rangle := x^\top y,$$

and the corresponding Euclidean norm is

$$\|x\|_2 := \sqrt{\langle x, x \rangle}. \quad (1.1)$$

Definition 1.1.1 For $x \in \mathbb{R}^d$ and $p \in [1, \infty)$, the ℓ_p -norm is defined by

$$\|x\|_p := \left(\sum_{i=1}^d |x_i|^p \right)^{1/p}, \quad (1.2)$$

while the ℓ_∞ -norm is given by

$$\|x\|_\infty := \max_{1 \leq i \leq d} |x_i|. \quad (1.3)$$

Unless stated otherwise, $\|\cdot\|$ denotes the Euclidean norm for vectors and the spectral norm for matrices.

Lemma 1.1.1 For all $x, y \in \mathbb{R}^d$, the Cauchy–Schwarz inequality holds:

$$|\langle x, y \rangle| \leq \|x\|_2 \|y\|_2. \quad (1.4)$$

All norms are equivalent in finite-dimensional spaces. Moreover, for every $x \in \mathbb{R}^d$,

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq \sqrt{d} \|x\|_2 \leq d \|x\|_\infty.$$

Matrix norms and spectral properties. The behavior of linear operators is measured by induced matrix norms, which are particularly important in stability estimates and in the analysis of smooth functions through Hessian bounds ([67]).

Definition 1.1.2 For a matrix $A \in \mathbb{R}^{m \times n}$ and a vector norm $\|\cdot\|_p$, the operator norm induced by $\|\cdot\|_p$ has the form

$$\|A\|_p = \|A\|_{p \rightarrow p} := \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}. \quad (1.5)$$

Lemma 1.1.2 For all matrices of compatible dimensions and all vectors x ,

$$\|Ax\|_2 \leq \|A\|_2 \|x\|_2, \quad (1.6)$$

and

$$\|AB\|_2 \leq \|A\|_2 \|B\|_2. \quad (1.7)$$

Definition 1.1.3 If $A \in \mathbb{R}^{d \times d}$ is square matrix, its trace is

$$\text{tr}(A) := \sum_{i=1}^d A_{ii}. \quad (1.8)$$

For a matrix $A \in \mathbb{R}^{m \times n}$, the Frobenius norm is defined by

$$\|A\|_F := \left(\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2 \right)^{1/2} = \sqrt{\text{tr}(A^\top A)}. \quad (1.9)$$

Lemma 1.1.3 *For matrices of compatible dimensions,*

$$\mathrm{tr}(A^\top B) = \langle A, B \rangle_F,$$

where

$$\langle A, B \rangle_F := \sum_{i,j} A_{ij} B_{ij},$$

and

$$\mathrm{tr}(AB) = \mathrm{tr}(BA). \quad (1.10)$$

For symmetric matrices, eigenvalues and quadratic forms provide a particularly useful description of curvature. Characteristic roots for symmetric matrices are real ([31]).

Definition 1.1.4 *Let $A = A^\top \in \mathbb{R}^{d \times d}$. Its eigenvalues are denoted by*

$$\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_d(A),$$

and we write

$$\lambda_{\max}(A) := \lambda_1(A), \quad \lambda_{\min}(A) := \lambda_d(A).$$

The matrix A is called positive semidefinite, written $A \succeq 0$, if

$$x^\top A x \geq 0 \quad \text{for all } x \in \mathbb{R}^d. \quad (1.11)$$

If the inequality is strict for every $x \neq 0$, then A is positive definite, written $A \succ 0$.

Lemma 1.1.4 *If $A = A^\top$, then*

$$\lambda_{\min}(A) = \min_{\|x\|_2=1} x^\top A x, \quad \lambda_{\max}(A) = \max_{\|x\|_2=1} x^\top A x,$$

and

$$\|A\|_2 = \max\{|\lambda_{\min}(A)|, |\lambda_{\max}(A)|\}. \quad (1.12)$$

Moreover, the following are equivalent:

1. $A \succeq 0$, (i.e., $A \succeq 0$ means that A is positive semidefinite),
2. $\lambda_{\min}(A) \geq 0$,
3. There exists a matrix B such that $A = B^\top B$.

Convexity and smoothness. The concepts of convexity and Lipschitz continuity of the gradient form the analytical basis for descent estimates in deterministic and stochastic optimization methods. Throughout the dissertation, $\nabla f(x)$ represents the gradient of f at x , and $\nabla^2 f(x)$ represents its Hessian matrix.

Definition 1.1.5 A set $S \subseteq \mathbb{R}^d$ is called convex if

$$\theta x + (1 - \theta)y \in S \quad \text{for all } x, y \in S, \theta \in [0, 1]. \quad (1.13)$$

A function $f : S \rightarrow \mathbb{R}$ is said to be convex if

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) \quad \text{for all } x, y \in S, \theta \in [0, 1]. \quad (1.14)$$

If the inequality is strict whenever $x \neq y$ and $\theta \in (0, 1)$, then f is strictly convex.

Definition 1.1.6 A differentiable function $f : S \rightarrow \mathbb{R}$, defined on an open convex set S , is called μ -strongly convex, with $\mu > 0$, if

$$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2 \quad \text{for all } x, y \in S. \quad (1.15)$$

Lemma 1.1.5 Let S be an open convex set and suppose that f is differentiable on S . Then f is convex on S if and only if

$$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle \quad \text{for all } x, y \in S. \quad (1.16)$$

If f is twice continuously differentiable, then convexity is equivalent to

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in S, \quad (1.17)$$

and μ -strong convexity is equivalent to

$$\nabla^2 f(x) \succeq \mu I_d \quad \text{for all } x \in S. \quad (1.18)$$

Definition 1.1.7 (Lipschitz continuity and Lipschitz gradient)

A function $f : S \rightarrow \mathbb{R}$ is said to be L -Lipschitz continuous if there exists $L \geq 0$ for which

$$|f(x) - f(y)| \leq L\|x - y\|_2 \quad \text{for all } x, y \in S. \quad (1.19)$$

A differentiable function $f : S \rightarrow \mathbb{R}$ is said to have an L -Lipschitz continuous gradient, or to be L -smooth, if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \text{for all } x, y \in S. \quad (1.20)$$

If $f \in C^2$, then (1.20) is equivalent to the uniform bound

$$\|\nabla^2 f(x)\|_2 \leq L \quad \text{for all } x \in S. \quad (1.21)$$

Lemma 1.1.6 (Descent lemma) Let $f : S \rightarrow \mathbb{R}$ be differentiable on a convex set S and assume that ∇f is L -Lipschitz continuous on S . Then, for all $x, y \in S$,

$$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2. \quad (1.22)$$

In particular, for an unconstrained gradient step $y = x - \alpha \nabla f(x)$,

$$f(x - \alpha \nabla f(x)) \leq f(x) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla f(x)\|_2^2. \quad (1.23)$$

Projection operators. Since constrained optimization methods rely heavily on projections, we now recall the main properties of projection operators that will be used in the sequel. In finite-dimensional Hilbert spaces, projections represent one of the fundamental approaches to ensuring feasibility while preserving a structure suitable for descent analysis ([25]).

Definition 1.1.8 *A matrix $P \in \mathbb{R}^{d \times d}$ is called a projector if*

$$P^2 = P. \quad (1.24)$$

It is called an orthogonal projector if, in addition,

$$P^\top = P. \quad (1.25)$$

If $\mathcal{M} \subseteq \mathbb{R}^d$ is a linear subspace, then the orthogonal projector onto $\mathcal{M}$ is the unique linear operator $P_{\mathcal{M}}$ such that

$$z = P_{\mathcal{M}}z + (I - P_{\mathcal{M}})z, \quad P_{\mathcal{M}}z \in \mathcal{M}, \quad (I - P_{\mathcal{M}})z \in \mathcal{M}^\perp. \quad (1.26)$$

Definition 1.1.9 *Let $S \subseteq \mathbb{R}^d$ be nonempty, closed, and convex. The Euclidean projection of a point $z \in \mathbb{R}^d$ onto S is given by*

$$\pi_S(z) := \arg \min_{x \in S} \|x - z\|_2. \quad (1.27)$$

In finite dimensions, the minimizer exists and is unique. Thus, $\pi_S : \mathbb{R}^d \rightarrow S$ is a well-defined single-valued mapping.

The projection operator π_S plays an important role in constrained optimization. It provides the closest feasible point to a given vector, and therefore serves as the natural mechanism for restoring feasibility after an unconstrained trial step. In the case of affine subspaces, it coincides with the usual orthogonal projector ([31]). In the case of box or nonnegativity constraints, it is explicitly computable componentwise, which makes projected methods particularly efficient in such settings.

Lemma 1.1.7 *Let $S \subseteq \mathbb{R}^d$ be nonempty, closed, and convex, and let $p = \pi_S(z)$. Then*

$$\langle z - p, x - p \rangle \leq 0 \quad \text{for all } x \in S.$$

Equivalently,

$$z - p \in N_S(p), \tag{1.28}$$

where $N_S(p)$ represents the normal cone of S at p .

Lemma 1.1.8 (Projection onto a box) *Let*

$$S := \{x \in \mathbb{R}^d : \ell \leq x \leq u\}, \tag{1.29}$$

where the inequalities are understood componentwise. Then the Euclidean projection onto S is given componentwise by

$$[\pi_S(z)]_i = \min\{u_i, \max\{\ell_i, z_i\}\}, \quad i = 1, \dots, d. \tag{1.30}$$

In particular, if $S = \mathbb{R}_+^d$, then

$$[\pi_{\mathbb{R}_+^d}(z)]_i = \max\{0, z_i\}. \tag{1.31}$$

The following proposition summarizes two fundamental nonexpansive properties of the projection operator.

Proposition 1.1.1 *Let $S \subset \mathbb{R}^n$ be a closed convex set, $x, y \in \mathbb{R}^n$, and $z \in S$. Then the following properties hold:*

- a) $\|\pi_S(y) - z\|_2 \leq \|y - z\|_2$;
- b) $\|\pi_S(x) - \pi_S(y)\|_2 \leq \|x - y\|_2$.

These formulas explain why projection-based algorithms are especially attractive for box-constrained and nonnegativity-constrained problems: feasibility can be enforced exactly and at negligible computational cost.

Definition 1.1.10 *Let $S \subseteq \mathbb{R}^d$ be convex, closed, and nonempty, let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be differentiable, and let $\alpha > 0$. The projected gradient mapping is obtained by*

$$G_\alpha(x) := \frac{1}{\alpha} (x - \pi_S(x - \alpha \nabla f(x))). \quad (1.32)$$

Equivalently, the corresponding projected step is

$$d_\alpha(x) := \pi_S(x - \alpha \nabla f(x)) - x = -\alpha G_\alpha(x).$$

A point $x \in S$ is stationary for the constrained problem $\min_{x \in S} f(x)$, if and only if

$$G_\alpha(x) = 0, \text{ or, equivalently, } x = \pi_S(x - \alpha \nabla f(x)). \quad (1.33)$$

Lemma 1.1.9 *Suppose that $S \subseteq \mathbb{R}^d$ be nonempty, closed, and convex, and let*

$$d_\alpha(x) = \pi_S(x - \alpha \nabla f(x)) - x.$$

Then

$$\langle \nabla f(x), d_\alpha(x) \rangle \leq -\frac{1}{\alpha} \|d_\alpha(x)\|_2^2. \quad (1.34)$$

If, in addition, ∇f is L -Lipschitz continuous, then

$$f(x + d_\alpha(x)) \leq f(x) - \left(\frac{1}{\alpha} - \frac{L}{2} \right) \|d_\alpha(x)\|_2^2. \quad (1.35)$$

Hence, whenever $\alpha \in (0, 2/L)$, the projected step yields a descent estimate unless $d_\alpha(x) = 0$.

The projection framework described above will be used repeatedly in the constrained algorithms studied later. In particular, the nonexpansiveness and variational characterization of the projection operator allow one to transfer geometric information about the feasible set into quantitative descent estimates, while the explicit form of the projector for simple sets such as boxes and orthants leads to computationally efficient implementations.

1.2 Probabilistic Foundations for Stochastic Optimization

This section collects the probabilistic concepts that will be used later in the analysis of stochastic, sample-based, and mini-batch optimization methods. The purpose is not to develop probability theory in full generality, but rather to introduce the language and results that repeatedly appear when one studies random sampling, stochastic gradients, sample average approximations, and acceptance tests based on subsampled information. The discussion therefore emphasizes expectation, variance, conditional expectation, modes of stochastic convergence, and concentration inequalities, together with their interpretation in the context of optimization algorithms.

Before introducing expectation and stochastic estimators, we briefly recall the basic measurable framework on which probability is built. In stochastic optimization, this level of formalism is useful because random samples, random indices, mini-batches, and iterates all need to be interpreted as measurable objects. Moreover, conditional expectation and filtrations, which play a significant role in iterative stochastic algorithms, are conveniently formulated in terms of sub- σ -algebras ([63]).

Definition 1.2.1 *Let Ω be a nonempty set. A collection $\mathcal{F} \subseteq 2^\Omega$ is called a σ -algebra if it satisfies the following properties:*

(S1) $\Omega \in \mathcal{F}$.

(S2) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, where A^c represents the complement of the set A .

(S3) If $A_1, A_2, \dots \in \mathcal{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Thus, a σ -algebra is a family of events that is stable under complements and countable unions, and therefore also under countable intersections.

Borel sigma algebra serves as a foundational concept in measure theory and topology. It is defined as the collection of all sets that can be formed from open sets in a topological space using countable unions, countable intersections, and complements. In particular, the Borel sigma algebra includes all open intervals, closed intervals, and any set derived from them through the mentioned operations (for real numbers).

Definition 1.2.2 *A pair $(\Omega, \mathcal{F})$ consisting of a set Ω and a σ -algebra $\mathcal{F}$ on Ω is said to be a measurable space, if $(S, \mathcal{S})$ is another measurable space, a mapping $X : \Omega \rightarrow S$ is said to be measurable if*

$$X^{-1}(B) \in \mathcal{F} \quad \text{for every } B \in \mathcal{S}. \quad (1.36)$$

In particular, a real-valued random variable is a measurable mapping from $(\Omega, \mathcal{F})$ into $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, where $\mathcal{B}(\mathbb{R})$ denotes the Borel σ -algebra.

With this notation in place, we introduce the standard probabilistic framework. A probability space is specified as $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω denotes the set of elementary outcomes, $\mathcal{F}$ is a σ -algebra of events, and $\mathbb{P}$ is a probability measure satisfying $\mathbb{P}(\Omega) = 1$ and countable additivity. A collection $\mathcal{F} \subseteq 2^\Omega$ is named a σ -algebra if it contains Ω , is closed under complements, and remains closed with respect to countable unions. A random variable is a measurable mapping from $(\Omega, \mathcal{F})$ into a measurable space, typically $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$. Its distribution is the probability measure induced on $\mathbb{R}^m$ by the rule

$$\mathcal{L}(\zeta)(B) := \mathbb{P}(\zeta \in B), \quad B \in \mathcal{B}(\mathbb{R}^m). \quad (1.37)$$

In stochastic optimization, such random objects may represent data samples, noise realizations, random indices, or random mini-batches selected during an iterative scheme.

Definition 1.2.3 Let X be an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Its expectation is given by

$$\mathbb{E}[X] := \int_{\Omega} X(\omega) d\mathbb{P}(\omega), \quad (1.38)$$

whenever the integral exists. For a random vector $G = (G_1, \dots, G_d)^\top \in \mathbb{R}^d$ such that $\mathbb{E}[\|G\|_2] < \infty$, the expectation is defined componentwise:

$$\mathbb{E}[G] := (\mathbb{E}[G_1], \dots, \mathbb{E}[G_d])^\top. \quad (1.39)$$

Lemma 1.2.1 If X and Y are integrable random variables and $a, b \in \mathbb{R}$, then

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]. \quad (1.40)$$

The same property holds componentwise for integrable random vectors.

Expectation describes the mean behavior of a random quantity, but for algorithmic purposes this is rarely sufficient on its own. In stochastic optimization one also needs to quantify how strongly a random estimator fluctuates around its mean.

Definition 1.2.4 Let X be a real-valued random variable with finite second moment. Its variance is

$$\text{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]. \quad (1.41)$$

If X and Y are square-integrable, their covariance is

$$\text{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]. \quad (1.42)$$

For a random vector $G \in \mathbb{R}^d$ with finite second moment, the covariance matrix is

$$\text{Cov}(G) := \mathbb{E}[(G - \mathbb{E}[G])(G - \mathbb{E}[G])^\top]. \quad (1.43)$$

Lemma 1.2.2 *If $\mathbb{E}[X^2] < \infty$, then*

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2, \quad (1.44)$$

and for all $a, b \in \mathbb{R}$,

$$\text{Var}(aX + b) = a^2 \text{Var}(X). \quad (1.45)$$

For a random vector $G \in \mathbb{R}^d$,

$$\mathbb{E}[\|G - \mathbb{E}[G]\|_2^2] = \text{tr}(\text{Cov}(G)). \quad (1.46)$$

The last identity is especially useful in optimization, since it translates the fluctuation of a random vector into a scalar measure that can be incorporated into descent estimates and complexity bounds.

A central concept in the analysis of iterative stochastic methods is conditional expectation. At iteration k , the algorithm has access only to the information generated up to that point; this information is naturally encoded by a sub- σ -algebra or, more generally, by a filtration.

In iterative stochastic methods, the information available up to iteration k is typically represented by a sub- σ -algebra $\mathcal{F}_k \subseteq \mathcal{F}$. A sequence $\{\mathcal{F}_k\}_{k \geq 0}$ such that

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F} \quad (1.47)$$

is called a filtration. Intuitively, $\mathcal{F}_k$ contains all information accumulated up to step k , including previously selected samples, previously computed iterates, and any other random quantities revealed by the algorithm. This notation is especially useful when expressing conditional unbiasedness and conditional variance bounds for stochastic estimators ([60, 63]). In the following definition, we use the indicator function of the set A , denoted by $\mathbf{1}_A$, which is equal to 1 on A and 0 outside A .

Definition 1.2.5 Let $\mathcal{G}$ be a sub- σ -algebra of $\mathcal{F}$ and let X be an integrable random variable. The conditional expectation of X given $\mathcal{G}$, denoted by $\mathbb{E}[X \mid \mathcal{G}]$, is a $\mathcal{G}$ -measurable random variable that satisfies the following property:

$$\mathbb{E}[X \mathbf{1}_A] = \int_A \mathbb{E}[X \mid \mathcal{G}] dP \quad \text{for all } A \in \mathcal{G}. \quad (1.48)$$

Such a random variable always exists and is uniquely determined up to a set of probability zero (i.e., P -almost surely).

When $X \in L^2$, conditional expectation may also be viewed as the orthogonal projection of X onto the closed subspace of $\mathcal{G}$ -measurable square-integrable random variables. This interpretation is conceptually important: it shows that conditional expectation extracts the best mean-square approximation of a random quantity using only the information contained in $\mathcal{G}$.

Lemma 1.2.3 If $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$ and X is integrable, then

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}] = \mathbb{E}[X \mid \mathcal{H}], \quad (1.49)$$

and in particular

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[X]. \quad (1.50)$$

Lemma 1.2.4 If $\mathbb{E}[X^2] < \infty$, then

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X \mid \mathcal{G})] + \text{Var}(\mathbb{E}[X \mid \mathcal{G}]). \quad (1.51)$$

These identities appear repeatedly in stochastic optimization. They allow one to separate the effect of the current random sampling from the effect of the randomness already accumulated through previous iterations. This is particularly useful when proving descent in expectation, bounding the variance of stochastic gradients, or analyzing adaptive sampling rules.

The quality of a random estimator is commonly described through unbiasedness and mean-square error.

Definition 1.2.6 *An estimator $\widehat{\theta}$ of a quantity θ is called unbiased if*

$$\mathbb{E}[\widehat{\theta}] = \theta. \quad (1.52)$$

Its mean-square error is defined by

$$\text{MSE}(\widehat{\theta}) := \mathbb{E}[(\widehat{\theta} - \theta)^2]. \quad (1.53)$$

For an unbiased estimator, the mean-square error corresponds to the variance. Thus, variance becomes a direct quantitative measure of estimator quality. In gradient-based stochastic optimization, this principle explains why variance reduction is desirable even when unbiasedness is preserved.

Another group of tools needed later concerns different modes of stochastic convergence. These notions formalize what it means for random approximations to approach their targets.

Definition 1.2.7 *A sequence $\{X_n\}$ converges almost surely to X , written*

$$X_n \xrightarrow{\text{a.s.}} X, \quad (1.54)$$

if

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\right\}\right) = 1. \quad (1.55)$$

Definition 1.2.8 *A sequence $\{X_n\}$ converges in probability to X , written*

$$X_n \xrightarrow{p} X, \quad (1.56)$$

if for every $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| > \varepsilon) \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (1.57)$$

Definition 1.2.9 For $p \geq 1$, one says that X_n converges to X in L^p , written

$$X_n \xrightarrow{L^p} X, \quad (1.58)$$

if

$$\mathbb{E}[|X_n - X|^p] \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (1.59)$$

Definition 1.2.10 The sequence (X_n) converges weakly to the random variable X , denoted by

$$X_n \xrightarrow{p} X,$$

if and only if the distribution functions $F_{X_n}(x)$ of X_n converge to the distribution function $F_X(x)$ of X at every point x where F_X is continuous. Formally, this can be expressed as:

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x) \quad \text{for all } x \in \mathbb{R} \text{ such that } F_X \text{ is continuous.}$$

Remark 1.2.1 The implications

$$X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{p} X, \quad X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X, \quad (1.60)$$

are standard. Moreover, convergence in probability implies convergence in distribution. These relations will be used implicitly when passing from moment estimates to probabilistic consistency statements.

For sampling-based methods, the laws of large numbers specify the basic justification for replacing expectations by empirical averages.

Theorem 1.2.1 Let $\{\xi_i\}_{i \geq 1}$ be independent and identically distributed random variables such that $\mathbb{E}[|\xi_1|] < \infty$, and let $\mu = \mathbb{E}[\xi_1]$. Then the sample average

$$\bar{\xi}_n := \frac{1}{n} \sum_{i=1}^n \xi_i \quad (1.61)$$

converges to μ in probability and almost surely:

$$\bar{\xi}_n \xrightarrow{p} \mu, \quad \bar{\xi}_n \xrightarrow{a.s.} \mu. \quad (1.62)$$

Theorem 1.2.2 *If $\{\xi_i\}_{i \geq 1}$ are independent and identically distributed with mean μ and variance $\sigma^2 < \infty$, then*

$$\sqrt{n}(\bar{\xi}_n - \mu) \longrightarrow \mathcal{N}(0, \sigma^2). \quad (1.63)$$

While laws of large numbers describe asymptotic behavior, concentration inequalities provide finite-sample control. These bounds are often the key ingredient when quantifying the approximation quality of sample averages and mini-batch estimators.

Lemma 1.2.5 (Markov inequality) *If $Y \geq 0$ almost surely and $a > 0$, then*

$$\mathbb{P}(Y \geq a) \leq \frac{\mathbb{E}[Y]}{a}. \quad (1.64)$$

Lemma 1.2.6 (Chebyshev inequality) *If $\mathbb{E}[X^2] < \infty$ and $\varepsilon > 0$, then*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}. \quad (1.65)$$

The inequalities above serve different purposes. Markov and Chebyshev inequality provide very general but coarse estimates based only on moments.

The relevance of these tools to stochastic optimization is immediate. Suppose that one wants to approximate an expected objective

$$F(x) = \mathbb{E}[f(x; \xi)] \quad (1.66)$$

by a sample average approximation

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n f(x; \xi_i), \quad (1.67)$$

or to approximate the full gradient by a mini-batch estimator

$$g_k = \frac{1}{m_k} \sum_{i \in \mathcal{N}_k} \nabla f_i(x_k). \quad (1.68)$$

In such situations, one typically studies whether the estimator is conditionally unbiased,

$$\mathbb{E}[g_k \mid \mathcal{F}_k] = \nabla F(x_k), \quad (1.69)$$

and whether its conditional fluctuation can be controlled, for example through a bound of the form

$$\mathbb{E}[\|g_k - \nabla F(x_k)\|_2^2 \mid \mathcal{F}_k] \leq \sigma_k^2. \quad (1.70)$$

Once such relations are available, total expectation and total variance identities make it possible to derive recursive bounds for the expected progress of the algorithm. At the same time, concentration inequalities provide finite-sample guarantees of the form

$$\mathbb{P}(|F_n(x) - F(x)| \geq \varepsilon) \leq \delta$$

or

$$\mathbb{P}(\|g_k - \nabla F(x_k)\|_2 \geq \varepsilon \mid \mathcal{F}_k) \leq \delta, \quad (1.71)$$

which directly motivate adaptive sample size rules, variance reduction strategies, and additional sampling procedures in step acceptance tests.

Throughout the later analysis, we will usually assume that the relevant random quantities admit at least finite second moments, and in some results stronger tail conditions such as boundedness, sub-Gaussianity, or sub-exponential behavior will be imposed when concentration arguments are needed. The exact assumptions will be stated locally in the theorems where they are used, but the concepts introduced in this section form the common probabilistic framework for all subsequent developments.

Chapter 2

Nonlinear Optimization

This chapter introduces deterministic nonlinear optimization framework that underlies the methods studied later in the dissertation. The discussion begins with general optimization models and the local notions of optimality that guide the construction of the algorithms. After that, attention is restricted to smooth unconstrained problems, where descent directions, step-size rules, and line search globalization provide the basic iterative machinery. The chapter then turns to constrained nonlinear optimization, including the geometry of the feasible set, the Karush–Kuhn–Tucker system, and the role of constraint qualifications. Building on that material, we briefly describe sequential quadratic programming methods, which combine local quadratic modeling with linearized constraints. Since the algorithms considered later rely strongly on step acceptance mechanisms, particular emphasis is placed on nonmonotone line search techniques, which relax the classical requirement of monotone decrease and are often more effective for curved, ill-scaled, or inexact models. The chapter closes with projected gradient methods for simple convex constraint sets, especially box-type constraints, which are computationally attractive and closely connected to the algorithmic developments in the sequel.

2.1 Optimization Models and Basic Concepts

We consider optimization problems of the form

$$\min_{x \in S} f(x), \quad (2.1)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the objective function and $S \subseteq \mathbb{R}^n$ is the feasible set. In the unconstrained case one has $S = \mathbb{R}^n$, while in the constrained setting S is determined by equality, inequality, or geometric restrictions ([58]). Throughout the chapter, the objective function is typically assumed to be continuously differentiable, and in some results twice continuously differentiable.

The distinction between local and global solutions is fundamental. In nonlinear optimization, global minimizers are often difficult to characterize or compute, especially when the objective is nonconvex. For that reason, most deterministic methods are designed to locate points satisfying local optimality conditions.

Definition 2.1.1 (Global and local minimizers) *Any point $x^* \in S$ is said to be a global minimizer of (2.1) if*

$$f(x^*) \leq f(x) \quad \text{for all } x \in S. \quad (2.2)$$

It is referred to as a local minimizer if there exists $\delta > 0$ for which

$$f(x^*) \leq f(x) \quad \text{for all } x \in S \cap B(x^*, \delta), \quad (2.3)$$

where $B(x^*, \delta)$ represents the open Euclidean ball with center at x^* and with a radius δ . If the inequality is strict for all $x \neq x^*$ in the neighborhood, then x^* is a strict local minimizer.

For the convex case, every local minimizer is immediately global. In the general nonlinear case, however, one typically settles for stationary points and local minimizers.

2.2 Unconstrained Optimization and Descent Directions

We begin with the unconstrained problem

$$\min_{x \in \mathbb{R}^n} f(x), \quad (2.4)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is assumed to be bounded from below and sufficiently smooth. The first step in the analysis of such problems is the characterization of local minimizers through derivative information.

Definition 2.2.1 (Stationary point) *A point $x \in \mathbb{R}^n$ is called a stationary point of f if*

$$\nabla f(x) = 0. \quad (2.5)$$

Theorem 2.2.1 (First-order necessary condition) *Let $f \in C^1$ on a neighborhood of x^* . If x^* is a local minimizer of (2.4), then*

$$\nabla f(x^*) = 0. \quad (2.6)$$

Every local minimizer is a stationary point, although the converse is not true in general. To distinguish minima from saddle points or maxima, one uses second-order information.

Theorem 2.2.2 (Second-order necessary condition) *Let $f \in C^2$ on a neighborhood of x^* . If x^* is a local minimizer of (2.4), then*

$$\nabla f(x^*) = 0 \quad (2.7)$$

and

$$\nabla^2 f(x^*) \succeq 0. \quad (2.8)$$

Theorem 2.2.3 (Second-order sufficient condition) *Let $f \in C^2$ on a neighborhood of x^* . If*

$$\nabla f(x^*) = 0 \quad (2.9)$$

and

$$\nabla^2 f(x^*) \succ 0, \quad (2.10)$$

then x^* is a strict local minimizer of (2.4).

These conditions motivate iterative methods that attempt to reduce the objective value by repeatedly moving from the current iterate x_k in a suitable direction p_k and with a suitable step length α_k :

$$x_{k+1} = x_k + \alpha_k p_k. \quad (2.11)$$

Definition 2.2.2 (Descent direction) *Let x_k be a point satisfying $\nabla f(x_k) \neq 0$. A vector $p_k \in \mathbb{R}^n$ is called a descent direction for f at x_k if*

$$\nabla f(x_k)^\top p_k < 0. \quad (2.12)$$

The negative gradient is the most basic descent direction:

$$p_k = -\nabla f(x_k). \quad (2.13)$$

This leads to the classical gradient method, which is simple and inexpensive but may be slow when the level sets of the objective are elongated or poorly scaled ([11]).

The Newton direction at the iterate x_k is the vector p_k defined as the solution of the (linear) Newton system

$$\nabla^2 f(x_k) p_k = -\nabla f(x_k). \quad (2.14)$$

A more informative direction is obtained by incorporating second-order curvature. If the Hessian matrix is nonsingular, the following relation defines the Newton direction

$$p_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k). \quad (2.15)$$

Near a nondegenerate local minimizer, Newton's method can achieve quadratic local convergence. However, it may be unstable or undefined far from a solution, and the exact Hessian may be too costly to compute or factorize. This motivates *quasi-Newton* methods, where the Hessian is replaced by a symmetric approximation B_k , yielding

$$B_k p_k = -\nabla f(x_k). \quad (2.16)$$

The matrix B_k is updated from first-order information, typically using the secant relation

$$B_{k+1} s_k = y_k, \quad s_k = x_{k+1} - x_k, \quad y_k = \nabla f(x_{k+1}) - \nabla f(x_k). \quad (2.17)$$

Among the best-known updates are the BFGS, DFP, and SR1 formulas. BFGS is especially important in practice because it preserves positive definiteness under the curvature condition

$$s_k^\top y_k > 0, \quad (2.18)$$

provided that the previous approximation is also positive definite. SR1, on the other hand, can produce indefinite approximations and hence search directions that are not necessarily descent directions. This observation is one of the reasons why nonmonotone globalization techniques become important in later sections. A related class is formed by spectral gradient methods, where the inverse Hessian is approximated by a scalar multiple of the identity, so that the direction remains the same to the negative gradient but still incorporates curvature information through a spectral coefficient.

2.3 Line Search Globalization

Once a search direction has been chosen, one needs a rule for determining the step size. In line search methods, one approach is to approximately solve the one-dimensional problem

$$\min_{\alpha > 0} \phi_k(\alpha), \quad \phi_k(\alpha) := f(x_k + \alpha p_k). \quad (2.19)$$

Exact minimization of ϕ_k is rarely necessary and often too expensive. Instead, practical methods rely on *inexact* step-size conditions that guarantee sufficient progress ([11, 58]).

Definition 2.3.1 (Armijo condition) *Let p_k be a descent direction and let $c_1 \in (0, 1)$. A step length $\alpha_k > 0$ satisfies the Armijo condition if*

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^\top p_k. \quad (2.20)$$

Since $\nabla f(x_k)^\top p_k < 0$, the right-hand side represents a decrease relative to the current objective value. The Armijo condition therefore requires a step that is not merely improving, but improving by a quantitatively meaningful amount.

A stronger criterion additionally controls the slope at the new point.

Definition 2.3.2 (Wolfe and strong Wolfe conditions) *Let $0 < c_1 < c_2 < 1$. The step size α_k is said to satisfy the Wolfe conditions if*

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^\top p_k, \quad (2.21)$$

$$\nabla f(x_k + \alpha_k p_k)^\top p_k \geq c_2 \nabla f(x_k)^\top p_k. \quad (2.22)$$

It satisfies the strong Wolfe conditions if (2.21) holds and

$$|\nabla f(x_k + \alpha_k p_k)^\top p_k| \leq c_2 |\nabla f(x_k)^\top p_k|. \quad (2.23)$$

The first inequality enforces sufficient decrease, while the second prevents the step from being excessively short. In particular, the strong Wolfe condition is well suited to quasi-Newton updates, since it helps maintain favorable curvature information.

A classical and inexpensive mechanism for producing Armijo steps is the *backtracking line search*. Starting from an initial trial value $\alpha_k^{(0)}$, one repeatedly reduces the step size by setting $\alpha_k = \beta^j \alpha_k^{(0)}$, where $\beta \in (0, 1)$ and $j \in \mathbb{N}_0$, until the Armijo condition is satisfied. This approach is particularly attractive when function evaluations are considerably cheaper than gradient evaluations.

When the step-size rule is combined with suitable assumptions on f and the search directions, one obtains global convergence results. A central statement in this direction is Zoutendijk's theorem, which shows that the total amount of decrease measured along the gradient is summable.

Theorem 2.3.1 (Zoutendijk condition) *Assume that f is continuously differentiable and bounded below on the level set*

$$\mathcal{L}(x_0) := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}, \quad (2.24)$$

and that, on an open neighborhood, ∇f is Lipschitz continuous of this set. Suppose that each p_k is a descent direction and each α_k satisfies the Wolfe conditions. If θ_k denotes the angle between $-\nabla f(x_k)$ and p_k , defined by

$$\cos \theta_k = \frac{-\nabla f(x_k)^\top p_k}{\|\nabla f(x_k)\|_2 \|p_k\|_2}, \quad (2.25)$$

then

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f(x_k)\|_2^2 < \infty. \quad (2.26)$$

This result implies that if the search directions remain uniformly separated from orthogonality with respect to the negative gradient,

then the gradient norm must tend to zero. In this sense, the geometry of the search directions is just as important as the step-size rule itself.

Although trust-region methods represent an important alternative globalization framework, the remainder of this chapter emphasizes line search methods, since these are more directly connected with the algorithms developed in later chapters.

2.4 Nonmonotone Line Search Techniques

Classical Armijo and Wolfe line searches require a monotone decrease in the objective or merit function. While this is safe and conceptually simple, it may be overly conservative in practice. In narrow curved valleys, in ill-scaled problems, or in methods using approximate curvature information, strict monotonicity often forces very short steps and may suppress the advantages of Newton, quasi-Newton, or spectral-gradient directions. This is precisely the setting in which nonmonotone line searches become useful, a perspective that is also emphasized later in this chapter.

The central idea of nonmonotone line search is to replace the current function value by a more permissive reference level based on recent history. This allows temporary increases in the objective, provided that the long-term behavior remains controlled.

A classical max-type nonmonotone criterion accepts a step ([29]) if

$$f(x_k + \alpha_k p_k) \leq R_k + c_1 \alpha_k \nabla f(x_k)^\top p_k, \quad (2.27)$$

where

$$R_k := \max_{0 \leq i \leq n_k} f(x_{k-i}), \quad 0 \leq n_k \leq M. \quad (2.28)$$

When $M = 0$, this reduces to the usual Armijo condition. For $M > 0$, the method compares the new point not only with the current value, but with the worst value among a window of recent iterates. This

relaxation can be especially helpful when full or nearly full steps are desirable, as in Newton or quasi-Newton methods.

A second important family uses an average-type reference rather than a window maximum ([75]). One recursively defines a reference sequence C_k and requires

$$f(x_k + \alpha_k p_k) \leq C_k + c_1 \alpha_k \nabla f(x_k)^\top p_k. \quad (2.29)$$

A common recursive construction is

$$Q_{k+1} = \eta_k Q_k + 1, \quad (2.30)$$

$$C_{k+1} = \frac{\eta_k Q_k}{Q_{k+1}} C_k + \frac{1}{Q_{k+1}} f(x_{k+1}), \quad (2.31)$$

with $Q_0 = 1$, $C_0 = f(x_0)$, and parameters $\eta_k \in [0, 1]$. In this construction, the nonmonotone behavior is controlled through a weighted aggregation of past values rather than a sharp maximum. This often leads to a smoother and more adaptive reference sequence.

The distinction between these two approaches is conceptually important. Max-type rules are easier to interpret and often more robust, whereas average-type rules are usually less sensitive to a single poor iterate and may better preserve fast local progress.

Another highly useful class of nonmonotone acceptance rules is designed for situations in which the step acceptance test should not rely on the assumption that the search direction is a descent direction. This may happen with indefinite quasi-Newton updates, derivative-free steps, inexact Newton methods, or line search procedures designed for residual minimization rather than direct function descent. In such cases one may introduce a summable relaxation sequence $\{\varepsilon_k\}$ and use an acceptance test of the form

$$f(x_k + \alpha_k p_k) \leq R_k + \varepsilon_k - \sigma \alpha_k^2 \beta_k, \quad (2.32)$$

where $\sigma > 0$, R_k is a nonmonotone reference value, and β_k is a positive control parameter reflecting the quality of the direction ([51, 48]). The summability condition

$$\sum_{k=0}^{\infty} \varepsilon_k < \infty \quad (2.33)$$

ensures that the relaxation does not destroy asymptotic convergence. This type of rule is useful precisely because it remains meaningful even when $\nabla f(x_k)^\top p_k$ is nonnegative or unavailable. That is one of the main reasons why nonmonotone techniques are naturally paired with SR1, spectral, inexact, or stochastic directions.

From a practical perspective, nonmonotone line searches should not be seen merely as a way of allowing objective increases. Their real value is that they create enough flexibility for the algorithm to exploit informative directions that may look unfavorable according to a purely monotone criterion. In many applications, the step that gives the best long-term behavior is not the one that yields the largest immediate decrease.

2.5 Constrained Nonlinear Optimization

Many optimization models arising in applications cannot be formulated without restrictions on the decision variables. Such restrictions may represent physical laws, balance relations, technological limitations, admissible ranges of parameters, or other structural requirements. This leads to the general nonlinear constrained optimization problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \quad i = 1, \dots, m, \\ & h_j(x) = 0, \quad j = 1, \dots, p, \end{aligned} \quad (2.34)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the objective function, while $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$ and $h_j : \mathbb{R}^n \rightarrow \mathbb{R}$ denote the inequality and equality constraint functions, respectively. Throughout, it is commonly assumed that these functions are at least continuously differentiable.

The feasible set is the set of all points that satisfy the constraints and is defined by

$$S := \{x \in \mathbb{R}^n : g_i(x) \leq 0, \ i = 1, \dots, m, \ h_j(x) = 0, \ j = 1, \dots, p\}.$$

A solution of (2.34) is sought among feasible points, that is, among points belonging to S .

Depending on the type of constraints involved, several important subclasses can be distinguished. If only equality constraints are present, one obtains an equality-constrained problem,

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad h_j(x) = 0, \quad j = 1, \dots, p. \quad (2.35)$$

These models arise when the variables are linked by exact relations that must be satisfied at every admissible point.

For a problem that involves only inequality constraints, then one has

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad g_i(x) \leq 0, \quad i = 1, \dots, m. \quad (2.36)$$

Such formulations are natural when the constraints describe upper or lower limits, safety margins, or resource capacities.

A particularly important case is the class of bound constrained (box constrained) problems, in which each variable is restricted independently by lower and upper bounds,

$$\ell \leq x \leq u. \quad (2.37)$$

Here, the bounds are understood componentwise, and they may be infinite: each ℓ_i may take values in $\mathbb{R} \cup \{-\infty\}$ and each u_i may take

values in $\mathbb{R} \cup \{+\infty\}$, with $\ell_i \leq u_i$ for all $i = 1, \dots, n$. The corresponding feasible region is a box in $\mathbb{R}^n$, and the special case $x \geq 0$ leads to nonnegativity constraints. Bound constraints are among the most common restrictions in optimization, both because of their modeling relevance and because of their relatively simple structure.

Another frequently studied subclass consists of linearly constrained problems, where the constraint functions are affine. In that case, the feasible set has a simpler geometric description, and one recovers important models such as linear programming and quadratic programming when the objective has a corresponding special form. By contrast, in general nonlinear programming both the objective function and the constraints can include nonlinear terms, which makes the analysis and numerical treatment significantly more challenging.

The presence of constraints changes the notion of optimality. In unconstrained optimization, stationary points are characterized by the vanishing of the gradient of the objective function. In the constrained setting, first-order optimality is typically expressed through the Karush–Kuhn–Tucker conditions, which combine gradient information from the objective and the active constraints through suitable multipliers. These conditions have a central place in both the theoretical analysis and the formulation of numerical algorithms for nonlinear constrained problems.

A broad class of numerical methods has been developed for solving problems of the form (2.34). Among the most classical and widely used approaches are:

- *penalty methods*, which incorporate the constraint violation into the objective function through an additional penalization term;
- *barrier and interior-point methods*, which enforce inequality constraints by preventing the iterates from approaching the boundary of the feasible set too aggressively;

- *augmented Lagrangian methods*, which combine ideas from penalty techniques and multiplier methods;
- *sequential quadratic programming methods*, which generate steps by solving quadratic programming subproblems that locally approximate the original nonlinear problem;
- *active-set methods*, which attempt to characterize the constraints that are active at a solution and solve reduced subproblems accordingly;
- *projected-gradient and feasible-direction methods*, especially in the presence of simple constraints such as boxes or balls in $\mathbb{R}^n$.

In addition, globalization mechanisms such as line search strategies, trust-region techniques, filter strategies, and merit functions are often incorporated in order to improve robustness and promote convergence from remote starting points.

Thus, constrained nonlinear optimization provides a general framework that encompasses a broad spectrum of models, ranging from simple bound-constrained problems to fully nonlinear systems with mixed equality and inequality restrictions. The specific structural features of the feasible set largely determines both the analytical features of the problem and the choice of an appropriate numerical method.

2.6 Optimality Conditions, Constraint Qualifications, and the KKT System

For constrained problems, derivative-based optimality conditions are expressed through the Lagrangian function.

Definition 2.6.1 (Lagrangian) *The Lagrangian associated with (2.34) is*

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^p \mu_j h_j(x),$$

where $\lambda \in \mathbb{R}^m$ is assigned to the inequality constraints, while $\mu \in \mathbb{R}^p$ corresponds to the equality constraints.

Let

$$A(x) := \{i \in \{1, \dots, m\} : g_i(x) = 0\} \quad (2.38)$$

denote the active set in inequality constraints at a feasible point x . The gradients of active constraints determine the local geometry of the feasible set.

Definition 2.6.2 (Karush–Kuhn–Tucker point) *A feasible point $x^* \in S$ is called a KKT point if there exist multipliers $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ such that*

$$\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0, \quad (2.39)$$

$$g_i(x^*) \leq 0, \quad i = 1, \dots, m, \quad (2.40)$$

$$h_j(x^*) = 0, \quad j = 1, \dots, p, \quad (2.41)$$

$$\lambda_i^* \geq 0, \quad i = 1, \dots, m, \quad (2.42)$$

$$\lambda_i^* g_i(x^*) = 0, \quad i = 1, \dots, m. \quad (2.43)$$

Stationarity equation (2.39) expresses the balance between the objective gradient and the gradients of the constraints that are relevant at x^* , namely the active inequality constraints and all equality constraints. Complementary slackness (2.43) indicates that only active inequalities may carry nonzero multipliers.

The validity of KKT conditions as necessary optimality conditions requires regularity assumptions, usually called constraint qualifications.

Definition 2.6.3 (LICQ) *The Linear Independence Constraint Qualification (LICQ) holds at a feasible point x^* if the family of gradients*

$$\{\nabla g_i(x^*) : i \in A(x^*)\} \cup \{\nabla h_j(x^*) : j = 1, \dots, p\} \quad (2.44)$$

is linearly independent.

Definition 2.6.4 (MFCQ) *The Mangasarian–Fromovitz Constraint Qualification (MFCQ) holds at a feasible point x^* if the gradients $\nabla h_1(x^*), \dots, \nabla h_p(x^*)$ are linearly independent and there exists a direction $d \in \mathbb{R}^n$ such that*

$$\nabla h_j(x^*)^\top d = 0, \quad j = 1, \dots, p, \quad \nabla g_i(x^*)^\top d < 0, \quad i \in A(x^*). \quad (2.45)$$

Definition 2.6.5 *For a convex constrained problem, the Slater condition requires the existence of a point $\bar{x}$ such that*

$$g_i(\bar{x}) < 0, \quad i = 1, \dots, m, \quad h_j(\bar{x}) = 0, \quad j = 1, \dots, p.$$

Theorem 2.6.1 (First-order optimality conditions) *Let x^* be a local minimizer of (2.34). If either LICQ or MFCQ is satisfied at x^* , then there exist multipliers λ^* and μ^* such that (x^*, λ^*, μ^*) satisfies the KKT system (2.39)–(2.43).*

To formulate second-order conditions, one introduces the *critical cone*. For a KKT point (x^*, λ^*, μ^*) , the critical directions are those that satisfy the linearized feasibility conditions and are orthogonal to the active linearized stationarity structure.

Definition 2.6.6 (Critical cone) *Let (x^*, λ^*, μ^*) be a KKT point. A critical cone at x^* is*

$$\tilde{\mathcal{S}}(x^*, \lambda^*) := \left\{ d \in \mathbb{R}^n : \begin{array}{ll} \nabla h_j(x^*)^\top d = 0, & j = 1, \dots, p, \\ \nabla g_i(x^*)^\top d = 0, & i \in A_+(x^*, \lambda^*), \\ \nabla g_i(x^*)^\top d \leq 0, & i \in A_0(x^*, \lambda^*) \end{array} \right\}, \quad (2.46)$$

where

$$A_+(x^*, \lambda^*) := \{i \in A(x^*) : \lambda_i^* > 0\}, \quad A_0(x^*, \lambda^*) := \{i \in A(x^*) : \lambda_i^* = 0\}.$$

Theorem 2.6.2 (Second-order sufficient condition) *Let (x^*, λ^*, μ^*) be a KKT point at which a suitable constraint qualification holds. If*

$$d^\top \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*, \mu^*) d > 0 \quad \text{for all } d \in \tilde{S}(x^*, \lambda^*) \setminus \{0\}, \quad (2.47)$$

then x^* is a strict local minimizer of (2.34).

2.7 Sequential Quadratic Programming

Sequential Quadratic Programming (SQP) methods are among the best known approaches for smooth constrained nonlinear optimization. At each iteration, they generate a search direction by solving a quadratic programming subproblem obtained from a local approximation of the original problem. Owing to their strong theoretical properties and good practical performance, SQP-type methods are often used as a standard benchmark in constrained optimization.

In this dissertation, SQP methods are not studied as the primary algorithmic framework, since the proposed methods are of a different type. They are included mainly because they provide a natural and well-established point of comparison in the numerical experiments. At the same time, classical SQP schemes may become computationally demanding due to the repeated solution of quadratic subproblems and their reliance on derivative and Hessian-related information, which can be less attractive in large-scale or sample-based settings.

At iteration k , one considers the subproblem

$$\begin{aligned}
 \min_{d \in \mathbb{R}^n} \quad & \nabla f(x_k)^\top d + \frac{1}{2} d^\top B_k d \\
 \text{s.t.} \quad & g_i(x_k) + \nabla g_i(x_k)^\top d \leq 0, \quad i = 1, \dots, m, \\
 & h_j(x_k) + \nabla h_j(x_k)^\top d = 0, \quad j = 1, \dots, p,
 \end{aligned} \tag{2.48}$$

where B_k is typically chosen as the Hessian of the Lagrangian, or as a quasi-Newton approximation thereof,

$$B_k \approx \nabla_{xx}^2 \mathcal{L}(x_k, \lambda_k, \mu_k). \tag{2.49}$$

The resulting direction d_k is then combined with a globalization strategy, such as a merit function, a filter method, or a line search rule. A generic SQP iteration therefore has the form

$$x_{k+1} = x_k + \alpha_k d_k, \tag{2.50}$$

where $\alpha_k \in (0, 1]$ is chosen to balance local model improvement and global stability.

Algorithm 1 Generic SQP framework

1. Choose an initial point x_0 , multiplier estimates (λ_0, μ_0) , and a symmetric matrix B_0 .
 2. For $k = 0, 1, 2, \dots$ do:
 - (a) Solve the quadratic subproblem (2.48) and obtain a step d_k ; if the corresponding KKT multipliers of the QP subproblem are available, use them to update the multiplier estimates $(\lambda_{k+1}, \mu_{k+1})$.
 - (b) If d_k is sufficiently small and the constraints are nearly satisfied, stop.
 - (c) Choose a step size α_k using a globalization mechanism.
 - (d) Set $x_{k+1} = x_k + \alpha_k d_k$.
 - (e) Update B_k if a quasi-Newton Hessian approximation is used.
-

SQP methods may be interpreted as Newton-type methods applied to the KKT system. Their appeal lies in the fact that they combine fast local convergence with a natural treatment of nonlinear constraints. At the same time, their efficiency depends strongly on how the QP subproblems are solved and how the step is globalized.

2.8 Projected Gradient Methods

We finally consider problems where the feasible set is simple enough for projection to be computed efficiently:

$$\min_{x \in S} f(x), \tag{2.51}$$

where $S \subseteq \mathbb{R}^n$ represents nonempty, closed, and convex set. In that setting, natural generalization of gradient descent is the projected gradient method.

Definition 2.8.1 (Projected gradient step) *Given a step size $\alpha_k > 0$, the projected gradient iteration is*

$$x_{k+1} = \pi_S(x_k - \alpha_k \nabla f(x_k)), \quad (2.52)$$

where π_S denotes the Euclidean projection onto S .

This iteration may be interpreted as taking an unconstrained gradient step and then restoring feasibility by projection. The method is particularly attractive when the projection has a closed form.

For box constraints

$$S = \{x \in \mathbb{R}^n : \ell \leq x \leq u\}, \quad (2.53)$$

the projection is componentwise clipping:

$$[\pi_S(z)]_i = \min\{u_i, \max\{\ell_i, z_i\}\}, \quad i = 1, \dots, n. \quad (2.54)$$

For the nonnegative orthant this reduces to

$$[\pi_{\mathbb{R}_+^n}(z)]_i = \max\{0, z_i\}. \quad (2.55)$$

For the simplex, the projection is also efficient, although it is somewhat more involved and usually requires sorting.

A useful quantity in projected methods is the projected gradient mapping introduced in Definition 1.1.10. In particular, the condition $G_\alpha(x) = 0$ provides an equivalent characterization of stationarity for the constrained problem $\min_{x \in S} f(x)$. This characterization makes projected methods especially convenient for the analysis.

When the objective splits as

$$\min_x f(x) + \psi(x),$$

where f is smooth and ψ is convex but possibly nonsmooth, the projected gradient framework extends to the proximal-gradient iteration

$$x_{k+1} = \text{prox}_{\alpha_k \psi}(x_k - \alpha_k \nabla f(x_k)),$$

where

$$\text{prox}_{\alpha \psi}(z) := \arg \min_{x \in \mathbb{R}^n} \left\{ \psi(x) + \frac{1}{2\alpha} \|x - z\|_2^2 \right\}. \quad (2.56)$$

This includes important instances such as ℓ_1 -regularized problems, sparse learning models, and constrained formulations that can be absorbed into ψ through indicator functions.

The appeal of projected and proximal-gradient methods lies in their balance between simplicity and structure preservation. Each iteration is relatively cheap, feasibility is maintained automatically, and the resulting schemes are well suited for large-scale problems whenever projection or proximal evaluation can be carried out efficiently. For these reasons, they provide a natural deterministic template for the stochastic and sample-based variants that appear in the later chapters.

Chapter 3

Stochastic Optimization

This chapter introduces the stochastic optimization framework that will be used throughout the remainder of the dissertation. The emphasis is placed on objective functions defined as large finite sums, which represent the core optimization models analyzed later in the thesis. While such formulations often originate from underlying mathematical expectations, working directly with finite sums allows for a more tractable computational framework. In contrast to deterministic optimization, exact evaluations of the objective function and its derivatives are often unavailable or computationally too expensive, so optimization algorithms must rely on sampled or approximate information. This naturally leads to questions of sample quality, estimator accuracy, and the balance between computational effort and reliability of the obtained search directions and acceptance decisions.

The chapter begins with the basic stochastic models and then explains the role of Sample Average Approximation as a bridge between expectation-based formulations and deterministic finite sum problems. After that, we discuss sampling mechanisms, mini-batch approximations, and the quality of stochastic estimators. A brief overview of stochastic first-order schemes is included only to the extent needed for

later developments, since the dissertation does not aim to survey the full literature on stochastic algorithms. The central part of the chapter is devoted to adaptive sample size strategies and additional sampling, which are particularly relevant for the methods developed later. The final sections focus on constrained stochastic optimization, with special attention to equality-constrained and simple bound-constrained finite sum models, because these are the settings in which the original contributions of the dissertation are formulated.

3.1 Stochastic Optimization Models

Many optimization problems are naturally modeled through randomness. In some settings, the random element describes uncertainty in the underlying data-generating mechanism; in others, it arises from simulation, sequential observations, or Monte Carlo approximation. A common feature in all these situations is that the true objective function is not given in a directly computable closed form, and optimization must therefore rely on stochastic surrogates or sample-based approximations [49, 60, 66, 68].

A stochastic optimization problem can be written in form

$$\min_{x \in X} F(x), \quad F(x) = \mathbb{E}[f(x; \xi)], \quad (3.1)$$

where $X \subseteq \mathbb{R}^n$ denotes the feasible set. Here, ξ is a random vector defined on an underlying probability space, and $f(x; \xi)$ denotes the corresponding random loss (or cost) evaluated at x . In statistical learning, $f(x; \xi)$ typically represents the loss induced by a single data point (i.e., one realization of ξ); in simulation-based optimization it may correspond to the outcome of one stochastic experiment; in data-driven applications it often quantifies the error associated with one observation or scenario [16, 49, 66].

A second model, especially important in machine learning, empirical risk minimization, and large-scale data analysis, is the finite sum problem

$$\min_{x \in S} F_N(x), \quad F_N(x) = \frac{1}{N} \sum_{i=1}^N f_i(x), \quad (3.2)$$

where the component functions f_i correspond to individual samples, observations, or data blocks. This model may be interpreted in two complementary ways. On the one hand, it is a deterministic optimization problem with a large objective sum. On the other hand, it may be viewed as a sample-based approximation of an expected-value problem. Both viewpoints are useful. The expectation formulation highlights the stochastic structure of the problem, while the finite sum representation is the natural starting point for mini-batching, subsampling, and variable sample size strategies [49, 66, 68].

The feasible set S may be unconstrained, but in this dissertation the constrained case is of primary interest. In particular, the later chapters focus on problems with nonlinear equality constraints and on problems with simple convex constraints such as box or nonnegativity conditions. For that reason, it is convenient to keep the feasible set explicit from the very beginning.

Definition 3.1.1 (Expectation and finite sum models) *A stochastic optimization problem is called an expected-value problem if its objective has the form*

$$F(x) = \mathbb{E}[f(x; \xi)]. \quad (3.3)$$

It is called a finite sum problem if its objective is expressed in the form

$$F_N(x) = \frac{1}{N} \sum_{i=1}^N f_i(x). \quad (3.4)$$

The regularity assumptions required in stochastic optimization depend on the algorithmic setting. One may assume continuity, convexity, differentiability, Lipschitz continuity of gradients, bounded second moments, or stronger smoothness properties. In this chapter, however, the assumptions are kept intentionally general. The exact hypotheses will be introduced later when needed. At the conceptual level, the main difficulty remains the same: exact function and derivative information is replaced by random or sample-based estimates, and this creates a trade-off between computational efficiency and statistical accuracy [49, 65, 66].

3.2 Sample Average Approximation and finite sum Reformulations

The most classical bridge between stochastic and deterministic optimization is provided by the Sample Average Approximation (SAA) principle. Instead of minimizing the expectation in (3.1) directly, one draws a sample $\xi_1, \dots, \xi_N$ and replaces the population objective by the empirical average. The resulting optimization problem is deterministic once the sample has been fixed, but its quality depends on the size and representativeness of the sample [63, 64, 65, 72].

Definition 3.2.1 (Sample Average Approximation) *Let $\xi_1, \dots, \xi_N$ be independent realizations of ξ . The Sample Average Approximation of the expected-value objective $F(x) = \mathbb{E}[f(x; \xi)]$ is defined by*

$$F_N(x) := \frac{1}{N} \sum_{i=1}^N f(x; \xi_i). \quad (3.5)$$

The associated SAA problem is

$$\min_{x \in S} F_N(x). \quad (3.6)$$

This construction is simple, but it has a fundamental position in stochastic programming and stochastic optimization. It turns an intractable expectation-based problem into a computable surrogate and thereby permits the employment of deterministic optimization techniques on a random approximation of the original model. At the same time, the approximation is useful only if optimal values and minimizers of the SAA problem reflect those of the underlying stochastic problem [63, 64, 72].

Theorem 3.2.1 (Uniform Law of Large Numbers) *Let $S \subset \mathbb{R}^n$ be nonempty and compact. Suppose that, for almost every ξ , the mapping $x \mapsto f(x; \xi)$ is continuous on S and that*

$$\mathbb{E} \left[\sup_{x \in S} |f(x; \xi)| \right] < \infty.$$

Then

$$\sup_{x \in S} |F_N(x) - F(x)| \rightarrow 0 \quad \text{almost surely as } N \rightarrow \infty,$$

where

$$F(x) = \mathbb{E}[f(x; \xi)], \quad F_N(x) = \frac{1}{N} \sum_{i=1}^N f(x; \xi_i),$$

Theorem 3.2.2 (Central Limit Theorem for SAA) *Under the assumptions of Theorem 3.2.1, suppose in addition that, for some $x^* \in S$,*

1. *the mapping $x \mapsto \nabla f(x; \xi)$ exists and is continuous in a neighborhood of x^* for almost every ξ ;*
2. $\mathbb{E}[\|\nabla f(x^*; \xi)\|^2] < \infty$;
3. *the Hessian $\nabla^2 F(x^*)$ exists and is positive definite.*

Then

$$\sqrt{N}(\nabla F_N(x^*) - \nabla F(x^*)) \xrightarrow{d} \mathcal{N}(0, \text{Var}(\nabla f(x^*; \xi))) \quad \text{as } N \rightarrow \infty.$$

Under standard continuity, compactness, and integrability assumptions, one expects $F_N(x)$ to converge to $F(x)$ as $N \rightarrow \infty$, and correspondingly the solutions of the SAA problems to approach the solutions for original problem. The following statement summarizes this principle.

Theorem 3.2.3 (Consistency of SAA) *Let $S \subset \mathbb{R}^n$ be compact and nonempty. Let $\{\xi^i\}_{i \geq 1}$ be i.i.d. samples with the same distribution as ξ , and define $F(x) := \mathbb{E}[f(x; \xi)]$, as in (3.5), with optimal values $v^* := \min_{x \in S} F(x)$ and $v_N^* := \min_{x \in S} F_N(x)$. Assume:*

1. *for almost every ξ , the mapping $x \mapsto f(x; \xi)$ is continuous on S ;*
2. $\mathbb{E}[\sup_{x \in S} |f(x; \xi)|] < \infty$;
3. *(uniform law of large numbers)*

$$\sup_{x \in S} |F_N(x) - F(x)| \rightarrow 0 \quad \text{almost surely as } N \rightarrow \infty.$$

Let $X^* := \arg \min_{x \in S} F(x)$ and $X_N^* := \arg \min_{x \in S} F_N(x)$, and let $x_N^* \in X_N^*$ be any selection. Then, almost surely:

1. $v_N^* \rightarrow v^*$;
2. *every accumulation point of $\{x_N^*\}$ belongs to X^* ;*
3. *equivalently, $\text{dist}(x_N^*, X^*) \rightarrow 0$, where $\text{dist}(x, X^*) := \inf_{y \in X^*} \|x - y\|$.*

If, in addition, $X^* = \{x^*\}$ is a singleton, then $x_N^* \rightarrow x^*$ almost surely.

Results of this type are classical in stochastic programming and form part of the theoretical justification for using SAA as an optimization model [64, 65, 72]. They explain why finite sum approximations are not merely convenient computational devices, but also meaningful surrogates of expectation-based problems.

For each fixed x , the central limit theorem further suggests that

$$\sqrt{N}(F_N(x) - F(x)) \longrightarrow \mathcal{N}(0, \sigma^2(x)), \quad (3.7)$$

where

$$\sigma^2(x) = \text{Var}(f(x; \xi)). \quad (3.8)$$

Thus, the approximation error typically decays at rate $N^{-1/2}$, which shows that statistical accuracy improves only gradually as the sample size grows [63, 64]. In large-scale problems, it is therefore often inefficient to insist on full sample evaluations at every iteration.

This observation is particularly relevant for the present dissertation. Even when the original model is already written in finite sum form, repeated evaluation of all N components may be unnecessarily expensive. This naturally motivates variable sample size strategies, in which smaller and cheaper approximations are used in early iterations and larger samples are employed only when the iterates enter regions where higher accuracy becomes algorithmically useful [32, 36, 38, 59].

Thesis-oriented remark. In the context of this dissertation, Sample Average Approximation should not be viewed only as a statistical approximation principle. It also provides the structural bridge between expectation-based stochastic models and the finite sum constrained optimization problems studied in the later chapters. In particular, it justifies why equality-constrained and bound-constrained finite sum models may be interpreted as computationally meaningful approximations of more general stochastic programs.

3.3 Sampling, Mini-Batches, and Estimator Quality

Once the objective is represented as a large finite sum, the next natural question is how much of that sum should be used at a single iteration. Evaluating all components yields exact objective and gradient information for the finite sum problem, but this may be prohibitively expensive when N is large. The standard alternative is to work with random subsamples or mini-batches [49, 66, 68].

Let

$$F_N(x) = \frac{1}{N} \sum_{i=1}^N f_i(x), \quad \nabla F_N(x) = \frac{1}{N} \sum_{i=1}^N \nabla f_i(x), \quad (3.9)$$

and let $\mathcal{B}_k \subseteq \{1, \dots, N\}$ be a random batch selected at iteration k . The corresponding mini-batch estimator of the gradient is

$$g_k = \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \nabla f_i(x_k), \quad (3.10)$$

possibly with an additional scaling factor depending on the sampling rule.

Definition 3.3.1 (Unbiased stochastic estimator) *For a random vector g_k to be called an unbiased estimator of $\nabla F_N(x_k)$ with respect to the information available at iteration k , we require that*

$$\mathbb{E}[g_k \mid \mathcal{F}_k] = \nabla F_N(x_k), \quad (3.11)$$

where $\mathcal{F}_k$ denotes the filtration containing all information revealed up to step k .

Unbiasedness is important, but it is not sufficient to characterize the quality of a stochastic estimator. The variance of the estimator is equally important. Smaller batches reduce computational cost, but they also introduce larger fluctuations. Larger batches improve reliability, but at a higher cost. This balance between estimator accuracy and computational efficiency is one of the fundamental themes in stochastic optimization [12, 49, 66].

Uniform mini-batching is the most common sampling rule, but it is not the only one. Typical sampling mechanisms for finite sum problems include:

- *uniform sampling*, where each component or subset has the same probability of being selected;
- *sampling without replacement*, where repeated selection within the same batch is excluded;
- *stratified sampling*, where data are partitioned into groups and each group is sampled separately;
- *importance sampling*, where index i is selected with probability $p_i > 0$ and the estimator is reweighted accordingly.

For example, under importance sampling one may use

$$g_k = \frac{1}{Np_{i_k}} \nabla f_{i_k}(x_k), \quad (3.12)$$

where i_k is drawn from $\{1, \dots, N\}$ according to probabilities p_i . With the correct scaling, the estimator remains unbiased, while a suitable choice of the sampling probabilities may reduce variance significantly [12, 49].

The quality of the estimator directly affects algorithmic decisions. In deterministic optimization, line search and globalization procedures

rely on exact comparisons of function values and directional information. In the stochastic setting, however, function values and derivatives are random or sample-based. Consequently, even the decision whether a step is good or not may become uncertain. Variance control is therefore not merely a technical issue: it influences the design of acceptance tests, descent conditions, and sample size adjustment mechanisms [23, 71].

Thesis-oriented remark. In the constrained methods studied later in this dissertation, the sample size is not just a computational parameter. It becomes part of the acceptance logic itself. For this reason, estimator quality and adaptive sample selection are more relevant here than a broad catalog of general-purpose sampling rules.

3.4 Stochastic First-Order Schemes and Variance Control

The most basic stochastic optimization method is stochastic gradient descent. In its simplest form, it updates the iterate by

$$x_{k+1} = x_k - \alpha_k g_k, \tag{3.13}$$

where g_k is a stochastic estimator of the gradient. In the unconstrained case, this is the direct stochastic analogue of gradient descent. In constrained settings, projection or proximal correction is usually added, which will be discussed later [49, 60, 66, 70].

The attractiveness of stochastic gradient methods lies in their very low per-iteration cost. Their main drawback is that estimator variance typically does not vanish automatically, so the iterates may continue to oscillate near a solution unless the step sizes are chosen carefully

or additional variance control is introduced. This observation motivated a wide range of methods based on larger batches, variable metric updates, variance reduction, or hybrid deterministic-stochastic phases [12, 13, 49, 68].

For the purposes of this dissertation, however, variance control is relevant primarily because it affects the reliability of local modeling and step acceptance in constrained finite sum optimization. The main algorithms studied later are not built around classical variance-reduction schemes as their central mechanism. Instead, the crucial issue is how sampled information enters penalty terms, projected steps, and acceptance tests. Accordingly, this chapter does not attempt to provide a broad survey of all first-order stochastic algorithms, but retains only the conceptual point that estimator variability must be controlled if one wants stochastic approximations to support reliable globalization mechanisms [23, 45, 49, 66].

3.5 Adaptive Sample Size, Additional Sampling, and Stochastic Step Acceptance

A central issue in stochastic optimization is that the quality of the information used at each iteration depends directly on the sample size. If one insists on using the full sample throughout the entire optimization process, then the resulting method may be accurate but unnecessarily expensive, especially in the early stages when the iterates are still far from stationarity. On the other hand, if one relies only on very small samples, then the objective and derivative approximations may be too noisy to support reliable step computation and acceptance decisions. This trade-off motivates the use of variable and adaptive sample size strategies [32, 36, 38, 43, 59].

The underlying principle is that optimization accuracy and sampling accuracy should be balanced rather than treated separately. When the current point is still far from a solution, a crude approximation of the objective is often sufficient, since the dominant source of error is the optimization error itself. As the method approaches a stationary point, however, the quality of the sample-based model becomes increasingly important, and larger samples are needed in order to prevent the algorithm from making decisions on the basis of unreliable information. In this sense, adaptive sampling can be viewed as a mechanism that coordinates computational effort with the level of accuracy actually required by the current stage of the optimization process [32, 36, 38, 47].

It is useful to distinguish between *variable sample size*, *adaptive sample size*, and *additional sampling*. A variable sample size method allows the number of sampled components to change with the iteration index. An adaptive sample size method goes one step further and updates the sample size according to quantities computed during the algorithm, such as progress indicators, stationarity measures, or estimates of sampling error. Additional sampling, in turn, refers to the use of a second sample, drawn after a trial point has been generated, whose role is not to define the step itself but to verify whether the proposed step behaves satisfactorily on information independent of the primary sample. This idea is central to the ASPEN and ASBOX methods developed in this dissertation [41, 42]; related sampling strategies can be found in [23, 39, 40, 45, 46].

Early variable sample size line search methods already showed that it is inefficient to keep approximation accuracy fixed throughout the entire optimization process [32, 36, 38, 59]. In such frameworks, the sample size is enlarged when the observed decrease becomes too small relative to the estimated inaccuracy of the current approximation. This idea is particularly natural in nonmonotone line search settings, where one seeks sufficient decrease only with respect to a local refer-

ence level instead of insisting on monotonic descent at every iteration [22, 30, 75]. In stochastic environments, however, such a logic becomes considerably more delicate, because both the search direction and the acceptance test are affected by sampling noise [23, 71].

A useful way to describe this balance is through two quantities of different nature. One is an indicator of optimization progress, which may take the form of a predicted decrease, a stationarity measure, or a merit reduction associated with the current sampled model. The other is an indicator of sampling uncertainty, which may be based on confidence intervals, variance proxies, or estimates of the error induced by the current sample size. The guiding principle is simple: if the observed progress is large compared with the sampling uncertainty, then the current approximation is considered informative enough and the sample size may be kept unchanged; if the progress becomes small relative to sampling error, then the sample should be enlarged, because the method is approaching the limit of what can be reliably inferred from the current model [32, 36, 38, 59].

This observation leads naturally to the notion of additional sampling. Suppose that a trial point is computed using a primary sample that defines the approximate objective, gradient, or local model. A decrease test based solely on that same sample may be overly optimistic, since the trial point was constructed precisely to perform well with respect to that approximate model. To reduce this bias, one may draw a second independent sample and use it as a control mechanism. The trial point is then accepted only if it also exhibits satisfactory behavior on this control sample. In this way, additional sampling strengthens the reliability of the acceptance criterion without requiring full sample evaluations at every iteration [23, 39, 40, 41, 42, 45, 46].

Additional sampling principle. Suppose that a trial point $\bar{x}_k$ is generated using a primary sample. The method is said to employ *additional sampling* if, after $\bar{x}_k$ has been computed, an auxiliary sample is drawn and used to verify whether the trial point satisfies a prescribed

acceptance condition on information independent of the primary sample.

From an algorithmic point of view, additional sampling plays two important roles. First, it improves robustness: a step that appears favorable only on the sample that generated it may fail on an independent control sample and can therefore be rejected. Second, it helps address a fundamental dependence issue in stochastic globalization procedures. In line search and trust-region frameworks, the sampled model, the trial step, and the function values at the trial point are otherwise strongly coupled. By introducing an auxiliary independent sample after the trial point has been formed, one can partially decouple these quantities and recover convergence analyses that are closer in spirit to deterministic arguments [23, 39]. A particularly attractive feature of this idea is that the control sample can remain very small, so the additional cost is modest compared with the gain in reliability [40, 41, 42, 45, 46].

The same philosophy has proved useful in several constrained settings that are directly relevant to this dissertation. In equality-constrained stochastic optimization, adaptive sampling has been combined with penalty techniques so that both the approximation quality and the strength of feasibility enforcement are adjusted as the iterates approach a feasible stationary point [37, 42, 47]. In projected methods for finite sum problems with simple constraints, additional sampling has been used together with nonmonotone descent tests in order to determine whether a projected trial step should be accepted on the basis of independent sampled information [40, 41, 44]. Related ideas have also been incorporated into second-order, trust-region, and proximal stochastic frameworks [23, 39, 45, 46].

A recurring conclusion in these works is that adaptive sampling should not merely reduce cost locally, but must eventually enforce sufficient accuracy. In bounded finite sum settings, this often means that the full sample is reached after finitely many iterations, so that the

method ultimately behaves like a deterministic algorithm applied to the target finite sum problem [36, 38, 40, 41]. In general expectation-based or general sample settings, the corresponding statement is that the effective sample size tends to infinity, which drives the approximation error to zero and supports convergence toward stationary or feasible stationary points [32, 47, 44, 65].

For the purposes of this dissertation, adaptive sample size and additional sampling are therefore not auxiliary implementation details, but structural ingredients of the methods studied later. They provide a principled way to combine cheap approximate models with reliable acceptance mechanisms, and they are particularly valuable when stochastic approximations must be reconciled with penalty terms, projected steps, or more complex constrained structures.

3.6 Constrained Stochastic Optimization

Since the main contributions of this dissertation concern constrained finite sum models, the constrained stochastic setting deserves particular emphasis. A general formulation is

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & F(x) \\ \text{s.t.} \quad & h(x) = 0, \\ & x \in S, \end{aligned} \tag{3.14}$$

where F is either an expected-value objective or a large finite sum, $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$ denotes a system of equality constraints, and S is a simple closed convex set, such as a box or the nonnegative orthant. This formulation already contains the two main classes studied later in the dissertation: finite sum problems with nonlinear equality constraints and finite sum problems with simple bound-type constraints.

Compared with unconstrained stochastic optimization, the constrained case is substantially more difficult. First, feasibility must

be controlled simultaneously with descent. Second, the relevant stationarity conditions involve both the objective function and the constraints, so stochastic errors propagate through the coupled system rather than through the single gradient expression. Third, globalization mechanisms such as line search or trust-region acceptance must assess not only predicted decrease, but also the effect of a step on feasibility or on a merit-type quantity. These features explain why constrained stochastic optimization generally requires more structure than basic unconstrained stochastic gradient methods [4, 24, 50, 54, 73].

A first important subclass is formed by *equality-constrained stochastic and finite sum problems*. In this setting, one may pursue feasibility either indirectly through penalty or augmented Lagrangian terms, or directly through local constrained models of SQP type. Penalty formulations replace the constrained problem by an unconstrained or simply constrained one expressed as

$$\Phi(x; \mu) = F(x) + \frac{\mu}{2} \|h(x)\|_2^2, \quad (3.15)$$

so that feasibility is enforced progressively through the penalty parameter μ [17, 33, 42, 47, 73]. This viewpoint is particularly attractive when combined with adaptive sampling, since the algorithm may simultaneously improve the fidelity of the sampled objective and the strength of the feasibility enforcement.

A more refined alternative is provided by augmented Lagrangian ideas, where one considers

$$\mathcal{L}_\mu(x, \lambda) = F(x) + \lambda^\top h(x) + \frac{\mu}{2} \|h(x)\|_2^2. \quad (3.16)$$

Stochastic variants of augmented Lagrangian methods update both primal variables and multipliers using sampled or approximate derivative information. Although such approaches are theoretically rich and

important in the broader literature [54, 55, 56, 61, 62], they are not the main methodological focus of this dissertation and will therefore serve primarily as related background.

A different but closely connected line of work is based on stochastic SQP-type methods, which combine sampled gradients and Hessian approximations with linearized constraints and local quadratic models. A generic stochastic SQP-type subproblem takes the form

$$\min_d g_k^\top d + \frac{1}{2} d^\top B_k d \quad \text{s.t.} \quad d \in \mathcal{S}_k, \quad (3.17)$$

where g_k and B_k are sampled or approximate first- and second-order quantities, and $\mathcal{S}_k$ is a polyhedral set representing the linearized constraints at the current iteration [2, 3, 4, 5, 20, 21, 24, 54]. These methods are highly relevant as comparative background for the present dissertation. Even though the algorithms developed later are not classical stochastic SQP methods in the strict sense, they share the same broad concern with local modeling, feasibility control, and globalization based on approximate information.

A second important subclass is formed by *stochastic problems with simple convex constraints*, especially box and nonnegativity constraints. In such settings, projected stochastic gradient methods and related proximal schemes provide natural baseline approaches, since projection onto the feasible set is computationally inexpensive. The basic projected update has the form

$$x_{k+1} = \pi_S(x_k - \alpha_k g_k), \quad (3.18)$$

where π_S denotes the Euclidean projection onto S [10, 25, 74]. In the present dissertation, however, the main interest is not in projection alone, but in projection combined with adaptive sample size rules, nonmonotone acceptance criteria, and additional sampling [40, 41, 44]. These ingredients are particularly useful in finite sum settings, where

one wishes to preserve the low cost of mini-batches while still making reliable acceptance decisions.

Other constrained stochastic approaches include interior-point and conditional-gradient techniques. Interior-point ideas are attractive for smooth constrained problems because they handle feasibility through barrier terms and preserve a rich local structure [17, 18]. Conditional-gradient or Frank–Wolfe-type methods avoid projections and instead rely on linear minimization over the feasible set, which can be advantageous for certain convex constrained problems [57]. Although these classes are relevant in the wider literature, they are not central to the present thesis and will not be developed in detail.

Another useful viewpoint is provided by local model methods. Even when the final algorithm is not formulated explicitly as an SQP method, it may still share the same conceptual components: a model of the objective, an approximate treatment of the constraints, and an acceptance test based on descent, feasibility, or merit reduction. This observation is important for understanding the methods developed later in the dissertation. The proposed algorithms are closer in spirit to penalty and projection frameworks with adaptive sampling, but they belong to the broader family of constrained stochastic optimization methods that seek to balance local model accuracy, computational efficiency, and feasibility control.

For this dissertation, two points are especially important. First, constrained finite sum optimization is the natural setting in which sample-based globalization mechanisms become most meaningful, since the algorithm must balance descent, feasibility, and estimator reliability at the same time. Second, the later chapters do not attempt to cover the full spectrum of constrained stochastic optimization methods, but focus on two specific and coherent lines: equality-constrained finite sum methods treated through penalty-based ideas, and box- or nonnegativity-constrained finite sum methods treated through projected steps enhanced by adaptive and additional sampling.

3.7 Remarks Specific to the Dissertation Setting

Although stochastic optimization is a broad and rapidly developing field, the present dissertation does not require an equally broad survey. The central objects in the later chapters are constrained finite sum problems, especially:

- equality-constrained finite sum models treated through penalty-type formulations, adaptive sample size rules, and additional sampling;
- simple convex constrained finite sum problems, particularly box and nonnegativity constraints, treated through projection-based ideas enhanced by nonmonotone acceptance criteria and additional sampling.

For this reason, the following aspects of the general literature on stochastic optimization problems are particularly important and relevant for the present dissertation:

1. the expected-value and finite sum formulations of stochastic objectives;
2. the Sample Average Approximation principle and the interpretation of large finite sums as empirical surrogates;
3. the use of mini-batches and variable sample sizes as a mechanism for balancing cost and accuracy;
4. the notion of stochastic estimators and the role of estimator quality in step computation and acceptance;
5. the difficulty of extending classical deterministic globalization mechanisms to noisy or sampled settings;

6. the use of additional sampling when the acceptance of a step cannot be decided reliably from the primary sample alone;
7. the constrained viewpoints based on projection, penalty terms, and local model subproblems.

By contrast, topics such as distributed implementations, asynchronous communication, or an extensive catalog of variance-reduced first-order methods are not essential for the aims of this dissertation and may therefore remain in the background. This selection reflects the methodological focus of the dissertation.

Chapter 4

Problem Formulation and Research Motivation

This chapter establishes the general optimization setting considered throughout the dissertation and explains the reasons for focusing on the two constrained problem classes studied later. The models of interest arise in large-scale environments in which the objective function is formed either as a mathematical expectation or as a finite sum of many component functions. At the same time, the decision variables are often subject to explicit restrictions expressing feasibility, physical admissibility, or structural relations among the unknowns. The combination of large-scale objective structure and constraints is therefore the central mathematical setting of this thesis.

The presentation in this chapter has three purposes. First, it introduces the constrained finite sum and stochastic models that provide the common basis for the subsequent developments. Second, it places these models in the context of representative applications in which large datasets, uncertainty, or repeated simulations make exact full sample computations impractical. Third, it fixes the notation and standing assumptions used later in the theoretical and numerical

analysis. Although the equality-constrained and bound-constrained settings lead to different algorithmic realizations, they are both rooted in the same sampled optimization framework, and it is therefore natural to introduce them within a unified chapter.

4.1 Research Motivation and Problem Classes

A large number of contemporary optimization problems are driven by data and characterized by scale. For problems of this type, the objective function is rarely ordered as a single closed expression of moderate cost. Much more often, it is assembled from many elementary contributions, each linked to one observation, one scenario, or one simulation output. This naturally leads to finite sum and expected-value formulations, which now play a fundamental role in machine learning, statistical estimation, inverse problems, and decision-making under uncertainty [16, 49, 66, 68].

The first class of problems we are interested in consists of equality-constrained optimization problems of the form

$$\min_{x \in S} f(x), \text{ where } S := \{x \in \mathbb{R}^n \mid h(x) = 0\}, \quad (4.1)$$

and $x \in \mathbb{R}^n$ represents the vector of decision variables. Set $S \subseteq \mathbb{R}^n$ is a simple feasible set, and $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$ represents nonlinear equality constraints. When randomness enters through data, noise, or uncertain scenarios, the objective often takes the form

$$f(x) = \mathbb{E}[\hat{f}(x; \xi)], \quad (4.2)$$

with ξ denoting a random vector and $\hat{f}(x; \xi)$ the loss associated with one realization. This formulation is particularly natural when performance is evaluated in expectation over a population or over future uncertain outcomes [49, 65, 66].

In many situations, however, one works with a finite dataset or with a finite family of scenarios. This finite-sum setting is the one adopted in this dissertation for both types of constrained optimization problems considered later: equality-constrained problems and box constrained problems. The objective function then becomes

$$f(x) = \frac{1}{N} \sum_{i=1}^N f_i(x), \quad (4.3)$$

or, more generally,

$$f(x) = \sum_{i=1}^N w_i f_i(x), \quad w_i \geq 0, \quad \sum_{i=1}^N w_i = 1. \quad (4.4)$$

The first expression (4.3) corresponds to the unweighted case, while the second (4.4) allows unequal influence of the component functions. Both formulations are useful in practice. The unweighted model is the standard empirical form in learning and fitting problems, whereas the weighted variant is appropriate when local contributions must be rebalanced according to importance, reliability, or frequency [32, 59].

The central mathematical objects of this dissertation belong to two subclasses of constrained finite sum optimization.

The first subclass is formed by nonlinear equality-constrained problems, which arise when the variables are linked by exact relations, balance laws, normalization requirements, or other structural equations that must be respected by admissible solutions. The presence of equality constraints fundamentally changes the geometry of the problem: progress cannot be judged only by reduction of the objective, but must also account for the extent to which the iterate satisfies the imposed relations. In the later chapters, this class is treated through penalty-based mechanisms coupled with sampled approximations and adaptive control of the sampling process [42, 47, 73].

The second subclass is formed by bound-constrained problems,

$$\min_{x \in S} f(x), \quad (4.5)$$

with

$$S = \{x \in \mathbb{R}^n \mid l_i \leq x_i \leq u_i, \ i = 1, \dots, n\}. \quad (4.6)$$

Here, the lower and upper bounds are allowed to be infinite, with $l_i \in \mathbb{R} \cup \{-\infty\}$ and $u_i \in \mathbb{R} \cup \{+\infty\}$ fulfilling $l_i \leq u_i$ for all $i = 1, \dots, n$. This formulation flexibly accommodates one-sided constraints or unconstrained variables within the same framework. A particularly important special case is obtained when only nonnegativity is imposed:

$$\min_{x \geq 0} f(x). \quad (4.7)$$

Problems of this kind are especially relevant when the variables represent physically interpretable quantities, proportions, intensities, or coefficients whose admissible range is known in advance. In this setting, projection becomes a natural algorithmic tool, since it provides a direct way to preserve feasibility while still allowing the use of sampled descent information [10, 25, 41].

The motivation for studying these two problem classes together is not only thematic, but also methodological. Although the treatment of equality constraints and bound constraints relies on different optimization devices, both settings raise the same core challenge: exact function and derivative information is often prohibitively expensive when the objective contains a very large number of terms. As a consequence, one is led to work with sampled approximations whose quality must be balanced against their computational cost [23, 36, 38].

To illustrate this point, consider the full gradient of the finite sum objective,

$$\nabla f(x) = \frac{1}{N} \sum_{i=1}^N \nabla f_i(x). \quad (4.8)$$

When N is large, the formation of $\nabla f(x)$ at every iteration may dominate the entire computational effort. A standard alternative is to replace it with a mini-batch approximation

$$g_k = \frac{1}{|B_k|} \sum_{i \in B_k} \nabla f_i(x_k), \quad (4.9)$$

where $B_k \subseteq \{1, \dots, N\}$ is a sampled subset of indices. Likewise, if the original objective is expressed in expectation form, then a finite sample may be used to define a Sample Average Approximation,

$$F_n(x) = \frac{1}{n} \sum_{j=1}^n f(x; \xi_j). \quad (4.10)$$

In both cases, the computational gain obtained by using only part of the available information is accompanied by a loss in accuracy. The resulting estimators are cheaper to compute, but they are also noisier, and this affects not only the search direction but also the reliability of step acceptance and descent tests [32, 36, 64, 72].

4.2 Application Settings and Modeling Motivation

The optimization models studied in this dissertation are not introduced for their formal structure alone. They arise naturally in applications where three features appear simultaneously: the objective function is built from many local contributions, the dimension of the problem is sufficiently large that full evaluations are expensive, and the decision variables are subject to explicit constraints expressing feasibility, admissibility, or structural relations. In such settings, constrained finite sum and stochastic formulations are not exceptional cases, but rather standard modeling tools.

A first important source of such problems is supervised learning and statistical estimation. If $\xi_1, \dots, \xi_N$ denote observations and $x \in \mathbb{R}^n$ is vector of model quantities, then the standard empirical formulation is

$$\min_{x \in S} \frac{1}{N} \sum_{i=1}^N \ell(x; \xi_i) + r(x), \quad (4.11)$$

where ℓ is the loss associated with one sample and r is a regularization term. This is the typical finite sum structure behind empirical risk minimization. In many models, the set S is not the whole space. Box restrictions may be introduced to control the magnitude of parameters, to enforce robustness or interpretability, or to reflect implementation limits, while nonnegativity appears whenever the parameters represent rates, weights, intensities, or proportions. In more structured formulations, one may also impose explicit side constraints,

$$\min_{x \in S} \frac{1}{N} \sum_{i=1}^N \ell(x; z_i) + r(x) \quad \text{subject to} \quad h(x) = 0, \quad (4.12)$$

which leads directly to constrained finite sum optimization. This point of view is especially relevant in large-scale machine learning, where optimization is inseparable from the statistical model itself and mini-batch approximations are essential because full-gradient computations are often too expensive to repeat throughout the entire run [19, 49, 68].

A closely related modeling perspective appears when one works not with a fixed dataset, but with an underlying data-generating mechanism. In that setting, the problem is naturally written as

$$\min_{x \in S} \mathbb{E}[f(x; \xi)], \quad (4.13)$$

possibly together with additional constraints. This formulation is appropriate in online learning, stochastic training, and data-driven decision problems in which observations are viewed as realizations of a random vector ξ ([65, 66, 72]).

Another broad class of examples is provided by inverse problems and reconstruction models. Here the variable x often represents a physical quantity that must satisfy basic admissibility conditions. A typical model has the form

$$\min_{x \in S} \frac{1}{2} \|Ax - y\|^2 + \rho(x), \quad (4.14)$$

where A is an observation or degradation operator, y denotes measured data, and ρ is a regularization term promoting a desired structure of the solution. In imaging, spectroscopy, deblurring, and related problems, nonnegativity is especially common because intensities, densities, or concentrations cannot take negative values. More generally, box constraints may encode physically meaningful ranges of admissible solutions. In that context, projected methods are natural because feasibility can be maintained explicitly, while sampled approximations allow the method to scale to large data regimes [10, 25, 41].

Scenario-based decision models form another setting in which the same mathematical structures appear in a particularly transparent way. If a decision must perform well across many realizations of an uncertain environment, then one typically evaluates it through a scenario aggregate,

$$\min_{x \in S} \frac{1}{N} \sum_{i=1}^N \phi(x; \xi_i), \quad (4.15)$$

or, at the population level, through an expected loss. Such models occur in risk-aware optimization, transport, allocation, and various forms of decision-making under uncertainty. Constraints are again indispensable: they encode budgets, capacities, or admissibility of the decisions.

Constrained optimization under uncertainty is also central in engineering and scientific computing. In these settings, equality constraints often represent the most important part of the model, because

they encode exact system relations rather than optional side conditions. A representative example is model predictive control, where at each decision epoch one solves a constrained optimization problem of the form

$$\begin{aligned} \min_{u_0, \dots, u_{V-1}} \quad & \sum_{v=0}^{V-1} \ell(x_v, u_v) + \varphi(x_V) \\ \text{subject to} \quad & x_{v+1} = F(x_v, u_v), \quad v = 0, \dots, V-1, \\ & x_v \in S, \quad u_v \in \mathcal{U}. \end{aligned} \tag{4.16}$$

The dynamic equations act as equality constraints, while the sets S and $\mathcal{U}$ impose operational limits. In large-scale or uncertainty-aware variants, the associated objective may again involve many scenarios, samples, or repeated evaluations. This explains why modern constrained stochastic optimization methods increasingly combine local modeling with sampled objective information, both in line search and in trust-region settings [18, 24, 54].

These application classes clarify the modeling logic behind the dissertation. The first class of methods developed later is tailored to problems where the decisive structural feature is the presence of nonlinear equality constraints. Such constraints are not merely technical complications; they encode exact relations that must be respected while the objective is minimized using sampled information. The second class is motivated by situations in which feasibility is naturally described by simple convex sets, especially boxes and the nonnegative orthant. In that case, the main challenge is to combine inexpensive feasibility preservation through projection with stochastic or finite sum information that is only partially observed at each iteration.

From this perspective, the emphasis on nonlinear equality constraints and simple bound constraints is not restrictive, but deliberate. These two classes are broad enough to include representative models from machine learning, reconstruction, scenario optimization,

and control, yet sufficiently structured to permit rigorous convergence analysis and computationally viable algorithms. They also expose a common methodological question that runs through the dissertation: once the objective function is large-scale, progress must be assessed not only in terms of descent, but also in terms of how much information is actually needed to make a reliable constrained optimization decision. That is precisely why sampled models, adaptive sample sizes, and additional sampling become essential ingredients of the methods studied in the subsequent chapters.

4.3 Sampled Models, and Standing Assumptions

In the sequel, we introduce some additional notation, finite-sum models defined on subsamples, and several key assumptions used in the subsequent parts of the dissertation.

The index set of the full finite sum problem is

$$\mathcal{N} := \{1, 2, \dots, N\}. \quad (4.17)$$

At iteration k , the subset $\mathcal{N}_k \subseteq \mathcal{N}$ denotes the primary sample used to construct approximate objective and gradient information. In addition, $\mathcal{D}_k \subseteq \mathcal{N}$ denotes an auxiliary sample used in adaptive or additional sampling procedures. Their sizes are denoted by

$$N_k := |\mathcal{N}_k|, \quad D_k := |\mathcal{D}_k|.$$

For the unweighted finite sum model, the sampled objective associated with $\mathcal{N}_k$ is defined by

$$f_{\mathcal{N}_k}(x) = \frac{1}{N_k} \sum_{i \in \mathcal{N}_k} f_i(x), \quad (4.18)$$

and the corresponding sampled gradient is

$$\nabla f_{\mathcal{N}_k}(x) = \frac{1}{N_k} \sum_{i \in \mathcal{N}_k} \nabla f_i(x). \quad (4.19)$$

Similarly, the auxiliary sample defines

$$f_{\mathcal{D}_k}(x) = \frac{1}{D_k} \sum_{i \in \mathcal{D}_k} f_i(x). \quad (4.20)$$

This auxiliary approximation is not intended to produce the search direction itself. Its purpose is to provide an external check of the candidate point and thereby support acceptance decisions based on information that is not fully tied to the sample from which the step was constructed.

In the equality-constrained context, the later methods rely on the quadratic penalty function

$$F(x, \mu) = f(x) + \frac{\mu}{2} \|h(x)\|^2, \quad (4.21)$$

where $\mu > 0$ is the penalty parameter. Replacing the exact objective with its sampled approximation yields the sampled penalty model

$$F_{\mathcal{N}_k}(x, \mu_k) = f_{\mathcal{N}_k}(x) + \frac{\mu_k}{2} \|h(x)\|^2. \quad (4.22)$$

This form will be the starting point for the penalty-based algorithms developed later for nonlinear equality-constrained finite sum problems.

In the bound-constrained framework, stationarity is characterized through the Euclidean projection onto the feasible set S . Using the projection formulas introduced in Section 2.8, we define the projected displacement

$$d(x) := \pi_S(x - \nabla f(x)) - x,$$

which serves as the natural stationarity measure in both the convergence analysis and the numerical experiments.

For equality constrained problems, first-order optimality is expressed through the Karush–Kuhn–Tucker conditions. A feasible value x^* is called a KKT point of (4.1) if there exists $\lambda^* \in \mathbb{R}^m$ such that

$$\nabla f(x^*) + \nabla h(x^*)^T \lambda^* = 0, \quad h(x^*) = 0. \quad (4.23)$$

The analysis carried out later rests on standard smoothness assumptions. We suppose that every function f_i , $i = 1, \dots, N$, is continuously differentiable and bounded from below. Moreover, there is a constant $L > 0$ obeying

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n, \quad \forall i = 1, \dots, N.$$

In the equality-constrained framework, we additionally assume sufficient regularity of the constraint mapping.

The mapping $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is continuously differentiable, and the function

$$q(x) = \frac{1}{2} \|h(x)\|^2 \quad (4.24)$$

has Lipschitz continuous gradient.

To characterize convergence to first-order constrained stationary points, we also use the Linear Independence Constraint Qualification.

At a feasible point x^* , the rows of $\nabla h(x^*)$ are linearly independent.

The later line search and acceptance conditions will involve a positive summable sequence $\{\varepsilon_k\}$ satisfying

$$\sum_{k=0}^{\infty} \varepsilon_k \leq \bar{\varepsilon} < \infty. \quad (4.25)$$

Moreover, the auxiliary sample used in additional sampling is assumed to be chosen independently of the primary sample from which the current trial point is generated. In the weighted setting, we also assume

that every local component has selection probability in the auxiliary sample.

Finally, in the numerical part of the dissertation, the computational effort of an algorithm is measured through the quantity FEV_k (Function Evaluation Value), which records the cumulative number of scalar products required to produce the iterate x_k starting from the initial point x_0 . This metric is particularly suitable in large-scale finite sum optimization, where inner products and gradient-related operations dominate the computational cost. It also allows a more informative comparison of methods with different sample sizes and therefore substantially different per-iteration workloads.

The framework introduced in this chapter provides the common mathematical basis for the remainder of the dissertation. It identifies the optimization models that motivate the later developments, clarifies their application background, and fixes the notation and assumptions under which the proposed methods are studied.

Chapter 5

Additional Sampling Method for Equality Constraints

In this section, we consider a class of optimization problems with nonlinear equality constraints and develop the method called ASPEN (*Additional Sampling Penalty method for Equality Nonlinear constraints*), which represents one of the central methods of this dissertation. The results presented in this section represent an original scientific contribution and have been published in the paper:

- N. KREJIĆ, N. K. JERINKIĆ, T. OSTOJIĆ, AND N. VUČIĆEVIĆ, ASPEN: An Additional Sampling Penalty Method for Finite Sum Optimization Problems with Nonlinear Equality Constraints, *arXiv preprint arXiv:2508.02299*, 2025.

Building on the general framework introduced in the previous chapter, the focus here is on problems where the objective function has a finite sum structure, while the constraints are given in the form of nonlinear equalities. The main idea of the method is to alleviate the high computational cost of large-scale problems by means of adaptive

sampling, while at the same time preserving the relevant theoretical and practical properties of the algorithm.

5.1 Problem formulation

Using the definition stated earlier, the problem considered here is of the form

$$\min_x f(x) := \frac{1}{N} \sum_{i=1}^N f_i(x), \quad \text{subject to} \quad h(x) = 0, \quad (5.1)$$

where the local objective functions $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \dots, N$, as well as the constraint mapping $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$, are continuously differentiable. Thus, we consider a problem in which the overall objective function is obtained by aggregating a large number of local cost functions, while feasibility is determined by nonlinear equality conditions.

Although the general theoretical framework has already been introduced in the previous chapter, in the case of equality constraints it is natural to rely on a penalty-based approach. Instead of working directly with the constrained problem, one introduces an appropriate penalty function in which constraint violation is incorporated into the objective function by means of a penalty parameter. Such an approach is particularly suitable in the finite sum setting, since it enables the construction of an algorithm relying on sample-based approximations of the objective function and its gradient, while avoiding explicit projection steps that may be computationally expensive.

In this sense, ASPEN is conceived as a first-order method that combines adaptive sampling, penalization, and additional sampling, where the acceptance of a candidate point and the possible increase of the sample size are determined on the basis of information obtained from an independent additional sample.

From the perspective of the dissertation's contributions, this approach provides a transition from the general theoretical framework to a concrete algorithm designed for nonlinear equality-constrained problems. Accordingly, in the remainder of this section, we first introduce the penalty formulation of the problem, then construct the ASPEN algorithm, and finally analyze its convergence properties and numerical behavior on appropriate test problems.

5.2 ASPEN algorithm

The method introduced in this section – ASPEN – belongs to the class of first-order methods, meaning that only the objective function and its first-order derivatives are assumed to be available. In what follows, we present the necessary framework and the description of the algorithm itself.

As a basis for approximating the objective function, we use the standard *Sample Average Approximation* (SAA) construction. More precisely, at iteration k we consider a subset of indices $\mathcal{N}_k \subseteq \mathcal{N} := \{1, 2, \dots, N\}$, where $N_k := |\mathcal{N}_k|$, and define the approximation of the objective function and its gradient by

$$f_{\mathcal{N}_k}(x) = \frac{1}{N_k} \sum_{i \in \mathcal{N}_k} f_i(x), \quad \nabla f_{\mathcal{N}_k}(x) = \frac{1}{N_k} \sum_{i \in \mathcal{N}_k} \nabla f_i(x). \quad (5.2)$$

Thus, the same sample $\mathcal{N}_k$ is used both for approximating the objective function and for approximating its gradient. This choice simplifies the construction of the algorithm and ensures that at each iteration one has a consistent description of the local model of the problem. At the same time, we emphasize that the choice of the sample $\mathcal{N}_k$ remains arbitrary, that is, no special restrictions are imposed on the sampling strategy within the construction of the method. This leaves room

for different implementation variants, depending on the nature of the problem and the available computational resources.

ASPEN is based on a penalty approach. Instead of solving the constrained problem directly, an unconstrained problem of the form $\min_{x \in \mathbb{R}^n} F(x, \mu)$ is considered, where a quadratic penalty function associated with the original problem (5.1) is introduced and defined by

$$F(x, \mu) = f(x) + \frac{\mu}{2} \|h(x)\|^2 =: f(x) + \mu q(x).$$

Here, $\mu > 0$ is the penalty parameter, while q denotes the constraint violation function (4.24), defined by the squared norm of the constraint mapping. In other words, the term $\|h(x)\|^2$ measures how far away the current point is from the feasible set, whereas the penalty parameter determines how strongly the algorithm enforces the reduction of this violation.

Since the algorithm uses a subsampled objective approximation, the corresponding approximate penalty function is introduced as

$$F_{\mathcal{N}_k}(x, \mu_k) = f_{\mathcal{N}_k}(x) + \frac{\mu_k}{2} \|h(x)\|^2,$$

where μ_k denotes the value of the penalty parameter at iteration k . The gradient of this function with respect to the variable x is

$$g_k := \nabla F_{\mathcal{N}_k}(x_k, \mu_k) = \nabla f_{\mathcal{N}_k}(x_k) + \mu_k \nabla h(x_k)^T h(x_k). \quad (5.3)$$

It is important to emphasize that only the part originating from the objective function is subsampled, whereas the constraint $h(x)$ and its derivative are computed exactly. The reason for this is twofold. On the one hand, the objective function has a finite sum structure, and this is precisely where the dominant computational cost appears, so subsampling is both the most natural and the most beneficial in that part. On the other hand, the constraints typically do not have the structure of a large sum and are usually of much smaller dimensional

or structural complexity, so their exact computation does not represent the main computational bottleneck. In this way, the algorithm retains reliable information about feasibility, while computational savings are achieved where they are most needed – in the evaluation of the objective function and its gradient.

Since we work with approximate functions, a monotone line search may be too restrictive and, in practice, often inefficient. For this reason, we introduce a nonmonotone Armijo-type line search with respect to the approximate penalty function. The search direction is taken as the negative gradient of the penalized approximation,

$$p_k = -g_k,$$

and then the step size α_k is determined by a backtracking procedure so that

$$F_{\mathcal{N}_k}(x_k + \alpha_k p_k, \mu_k) \leq F_{\mathcal{N}_k}(x_k, \mu_k) + \eta \alpha_k g_k^T p_k + \epsilon_k, \quad (5.4)$$

where $\eta \in (0, 1)$ is a line search constant, and the sequence $\{\epsilon_k\}_{k \in \mathbb{N}}$ is a sequence of positive summable numbers, i.e.,

$$\sum_{k=0}^{\infty} \epsilon_k < \infty.$$

The role of the term ϵ_k is to allow controlled nonmonotonicity: the algorithm does not require strict decrease at every iteration, but rather permits small oscillations, which are natural and often useful in a stochastic environment. After the line search procedure, the candidate point is defined by

$$\bar{x}_k = x_k + \alpha_k p_k,$$

and is then subjected to an additional test.

Depending on the behavior of the sample size throughout the iterations, ASPEN may operate in two regimes. The first regime is

the *mini-batch* (MB) regime, in which $N_k < N$ throughout the whole optimization process, meaning that the algorithm always uses only a subset of the data. The second regime is the *full sample* (FS) regime, in which for sufficiently large k one has $N_k = N$, so the algorithm switches to the full sample and then behaves as a deterministic first-order penalty method. Intuitively, the MB phase is responsible for computational efficiency and adaptivity, since it enables inexpensive iterations in the early stages of optimization, while the FS phase, if reached, provides more accurate behavior near stationary points of the penalized problem. For this reason, the transition between these two phases is one of the key structural features of the method.

If at iteration k the method is in the FS regime, that is, if $N_k = N$, then one sets $\mathcal{N}_k = \mathcal{N}$ and the candidate point is accepted without any additional test, i.e., $x_{k+1} = \bar{x}_k$. Since the sample size sequence is nondecreasing, in this phase the algorithm essentially behaves like a deterministic penalty method: one approximately solves the minimization problem for the penalty function with fixed μ_k , and once the norm of the gradient becomes sufficiently small, the penalty parameter is increased and the next penalized subproblem is considered. Nevertheless, it should be kept in mind that the entire sequence of iterates remains stochastic, since the moment at which the method switches to the full sample depends on the previous random course of the algorithm.

The essential novelty of ASPEN lies in the MB phase. When $N_k < N$, after the candidate point has been computed, an additional test of its qualitative acceptability is performed by means of an independent additional sample $\mathcal{D}_k \subseteq \mathcal{N}$. This additional sample is, in general, small, selected randomly and without replacement, and serves to verify whether the candidate point is sufficiently good from the perspective of new, independent information about the objective function. In this way, the risk is reduced that a step is accepted merely because it was favorable for the particular main sample $\mathcal{N}_k$, but not for the problem

as a whole.

On the additional sample, we define the function

$$F_{\mathcal{D}_k}(x, \mu_k) := \frac{1}{D_k} \sum_{i \in \mathcal{D}_k} f_i(x) + \frac{\mu_k}{2} \|h(x)\|^2,$$

where $D_k := |\mathcal{D}_k| \leq N_k$. One then checks whether the following condition holds:

$$F_{\mathcal{D}_k}(\bar{x}_k, \mu_k) \leq F_{\mathcal{D}_k}(x_k, \mu_k) - c \|\nabla F_{\mathcal{D}_k}(x_k, \mu_k)\|^2 + C\epsilon_k, \quad (5.5)$$

where $c > 0$ and $C > 0$ are constants, with C possibly large. If this condition is satisfied, the candidate point is accepted. Otherwise, the step is rejected and the method remains at the point x_k .

This additional test plays a dual role. First, it serves as the acceptance criterion for the candidate point, since it evaluates whether sufficient progress has been achieved with respect to an independent additional sample. Second, it simultaneously governs the update of the main sample size. If condition (5.5) is satisfied, the current approximation level is regarded as sufficiently accurate, and the same sample size is retained in the next iteration. If the condition is not satisfied, it is concluded that the current sample is not sufficiently representative, and the sample size is increased to some value

$$N_{k+1} \in \{N_k + 1, \dots, N\}.$$

It is important to emphasize that this increase is arbitrary, which gives the algorithm additional flexibility and allows it to adapt to different types of problems. Thus, additional sampling in ASPEN simultaneously controls both the acceptance of the point and the growth of the sample size, which is one of the key ideas of the method.

In the MB phase, the penalty parameter is updated according to the constraint violation measure $\|h(x_k)\|$. If the current iterate is

still relatively far from the feasible set, i.e., if $\|h(x_k)\| > \epsilon_k$, then the penalty parameter is increased so that the algorithm more strongly enforces progress toward feasibility. Otherwise, the penalty parameter remains unchanged. In this way, a balance is achieved between decreasing the objective function and improving feasibility.

The complete procedure is formalized in Algorithm 2.

The logic of the algorithm can be explained as follows. First, based on the main sample, the approximate gradient of the penalized function is formed and the search direction is determined. Then, a candidate point is obtained through a nonmonotone line search. After that, an independent additional sample checks whether the achieved progress is sufficiently representative. Depending on the outcome of this test, the candidate point is either accepted or rejected, and the sample size is either kept unchanged or increased. Simultaneously, the penalty parameter is adjusted according to the level of constraint violation. In this way, ASPEN combines the efficiency of subsampling with a controlled approach to stationarity and feasibility.

5.3 Convergence analysis

Within this subsection, we establish almost sure convergence results for ASPEN. The analysis is naturally split according to the two regimes in which the method may operate: the mini-batch (MB) regime, where subsampling is used indefinitely, and the full sample (FS) regime, where the method eventually switches to the full sample. We first analyze these two scenarios separately and then combine the obtained conclusions into the final convergence result.

We start by stating the basic regularity assumption under which the subsequent analysis is carried out.

Algorithm 2 ASPEN: Additional Sampling Penalty Method for Equality Nonlinear Constraints

```

1: Input:  $x_0 \in \mathbb{R}^n$ ,  $N_0 \in \mathbb{N}$ ,  $c, \beta \in (0, 1)$ ,  $C$ ,  $\gamma$ ,  $\mu_0 > 0$ ,  $\{\epsilon_k\}_{k=0}^\infty$ 
2: for  $k = 0, 1, 2, \dots$  do
3:   if  $N_k < N$  then
4:     Choose  $\mathcal{N}_k \subseteq \mathcal{N}$ 
5:   else
6:     Set  $\mathcal{N}_k = \mathcal{N}$ 
7:   end if
8:   Calculate  $p_k = -g_k$  via (5.3)
9:   Find the smallest  $j \in \mathbb{N}_0$  such that  $\alpha_k = \beta^j$  satisfies (5.4)
10:  Set  $\bar{x}_k = x_k + \alpha_k p_k$ 
11:  if  $N_k = N$  then
12:    Set  $x_{k+1} = \bar{x}_k$ 
13:    if  $\|\nabla F_{\mathcal{N}_k}(x_k, \mu_k)\| < \frac{1}{\mu_k}$  then
14:      Set  $\mu_{k+1} = \gamma \mu_k$ 
15:    else
16:      Set  $\mu_{k+1} = \mu_k$ 
17:    end if
18:  else
19:    Choose  $\mathcal{D}_k \subseteq \mathcal{N}$  randomly and uniformly, without replacement
20:    if (5.5) holds then
21:      Set  $x_{k+1} = \bar{x}_k$  and  $N_{k+1} = N_k$ 
22:    else
23:      Set  $x_{k+1} = x_k$  and choose  $N_{k+1} \in \{N_k + 1, \dots, N\}$ 
24:    end if
25:    if  $\|h(x_k)\| > \epsilon_k$  then
26:      Set  $\mu_{k+1} = \gamma \mu_k$ 
27:    else
28:      Set  $\mu_{k+1} = \mu_k$ 
29:    end if
30:  end if
31: end for

```

Assumption A 1 *The functions f_i , $i = 1, \dots, N$, are bounded from below and continuously differentiable with L -Lipschitz continuous gradients. Moreover, the constraint violation function q is continuously differentiable with L -Lipschitz continuous gradient.*

Assumption 1 plays a key role in the analysis. First, the lower boundedness of the local cost functions excludes the possibility of an unbounded decrease of the objective values and therefore makes a descent-based argument meaningful. Second, the Lipschitz continuity of the gradients is the standard required smoothness property that allows one to control the local behavior of the sampled penalty functions by means of the descent lemma. In particular, it guarantees that the backtracking line search doesn't reduce the step size indefinitely, but instead yields a strictly positive lower bound on the accepted step length. This property is crucial in proving that every accepted step produces a quantitatively nontrivial amount of progress.

Notice that Assumption 1 immediately implies that every sampled function f_{N_k} is also bounded from below and continuously differentiable with L -Lipschitz continuous gradient. Consequently, the sampled penalty functions $F_{N_k}(x, \mu_k)$ are bounded from below and continuously differentiable as well. Moreover, their gradients are Lipschitz continuous with constant $(1 + \mu_k)L$. Since the search direction is chosen as the negative gradient of the sampled penalty function, the standard descent-lemma argument leads to the following result.

Lemma 5.3.1 *Suppose that Assumption 1 holds. Then the backtracking line search defined by (5.4) is well defined. Moreover, the accepted step size α_k satisfies*

$$\alpha_k \geq \min \left\{ 1, \frac{2\beta(1 - \eta)}{(1 + \mu_k)L} \right\} =: \bar{\alpha}_{\mu_k}. \quad (5.6)$$

Proof. Fix an iteration k . By Assumption 1, each sampled function $f_{\mathcal{N}_k}$ is continuously differentiable and has L -Lipschitz continuous gradient. Since the constraint violation function q also has L -Lipschitz continuous gradient, it follows that the sampled penalty function

$$F_{\mathcal{N}_k}(x, \mu_k) = f_{\mathcal{N}_k}(x) + \mu_k q(x)$$

has $(1 + \mu_k)L$ -Lipschitz continuous gradient.

Now let

$$g_k = \nabla F_{\mathcal{N}_k}(x_k, \mu_k), \quad p_k = -g_k.$$

Applying the descent lemma to $F_{\mathcal{N}_k}(\cdot, \mu_k)$ at the point x_k in the direction p_k , we obtain that for every $\alpha \in [0, 1]$,

$$F_{\mathcal{N}_k}(x_k + \alpha p_k, \mu_k) \leq F_{\mathcal{N}_k}(x_k, \mu_k) + \alpha g_k^T p_k + \frac{(1 + \mu_k)L}{2} \alpha^2 \|p_k\|^2.$$

Since $p_k = -g_k$, we have

$$g_k^T p_k = -\|g_k\|^2 = -\|p_k\|^2,$$

and therefore

$$F_{\mathcal{N}_k}(x_k + \alpha p_k, \mu_k) \leq F_{\mathcal{N}_k}(x_k, \mu_k) + \alpha \left(1 - \frac{(1 + \mu_k)L}{2} \alpha \right) g_k^T p_k.$$

Hence, if

$$\alpha \leq \frac{2(1 - \eta)}{(1 + \mu_k)L},$$

then

$$1 - \frac{(1 + \mu_k)L}{2} \alpha \geq \eta,$$

which yields

$$F_{\mathcal{N}_k}(x_k + \alpha p_k, \mu_k) \leq F_{\mathcal{N}_k}(x_k, \mu_k) + \eta \alpha g_k^T p_k.$$

Since $\epsilon_k > 0$, this implies the Armijo-type condition (5.4). Thus, every step size satisfying

$$\alpha \leq \frac{2(1-\eta)}{(1+\mu_k)L}$$

is acceptable, which proves that the backtracking procedure is well defined.

It remains to derive the lower bound on the accepted step. If

$$\frac{2(1-\eta)}{(1+\mu_k)L} \geq 1,$$

then $\alpha = 1$ already satisfies (5.4), so $\alpha_k = 1$. Otherwise, the backtracking procedure tests the geometric sequence $1, \beta, \beta^2, \dots$, and since every sufficiently small step is acceptable, the first accepted step satisfies

$$\alpha_k \geq \beta \frac{2(1-\eta)}{(1+\mu_k)L}.$$

Combining the two cases gives

$$\alpha_k \geq \min \left\{ 1, \frac{2\beta(1-\eta)}{(1+\mu_k)L} \right\},$$

which proves the claim. $\square$

We now turn to the MB phase of ASPEN. As is standard in additional-sampling-based methods, we introduce the sets of “good” and “bad” realizations of the additional sample. More precisely, let $\mathcal{D}_k^+$ denote the set of all possible realizations of $\mathcal{D}_k$ at iteration k for which the acceptance test (5.5) is satisfied, namely

$$\mathcal{D}_k^+ = \left\{ \mathcal{D}_k \subseteq \mathcal{N} \mid F_{\mathcal{D}_k}(\bar{x}_k, \mu_k) \leq F_{\mathcal{D}_k}(x_k, \mu_k) - c\|\nabla F_{\mathcal{D}_k}(x_k, \mu_k)\|^2 + C\epsilon_k \right\},$$

and let $\mathcal{D}_k^-$ denote its complement, i.e.,

$$\mathcal{D}_k^- = \left\{ \mathcal{D}_k \subseteq \mathcal{N} \mid F_{\mathcal{D}_k}(\bar{x}_k, \mu_k) > F_{\mathcal{D}_k}(x_k, \mu_k) - c\|\nabla F_{\mathcal{D}_k}(x_k, \mu_k)\|^2 + C\epsilon_k \right\}.$$

The next lemma shows that if the method stays forever in the MB regime, then after finitely many iterations the additional-sampling test must accept *all* admissible additional samples.

Lemma 5.3.2 *Suppose that Assumption 1 holds. If $N_k < N$ for all $k \in \mathbb{N}$, then a.s. there exists $k_1 \in \mathbb{N}$ such that $\mathcal{D}_k^- = \emptyset$ for all $k \geq k_1$.*

Proof. Since we are in the MB scenario, the sequence $\{N_k\}$ is non-decreasing and bounded above by $N - 1$. Hence, there exists an index $k_0 \in \mathbb{N}$ and a number $\bar{N} \in \{1, \dots, N - 1\}$ such that

$$N_k = \bar{N} \quad \text{for all } k \geq k_0.$$

In other words, if the full sample is never attained, then the monotonicity of the sample size sequence forces it to stabilize after finitely many iterations. This simple observation is the starting point of the contradiction argument below.

Assume now, aiming for contradiction, that the statement of the lemma is false. Then there exists an infinite set of indices $K \subseteq \mathbb{N}$ such that

$$\mathcal{D}_k^- \neq \emptyset \quad \text{for all } k \in K.$$

Without loss of generality, we may also assume that $K \subseteq \{k \in \mathbb{N} : k \geq k_0\}$, so that $N_k = \bar{N}$ for every $k \in K$. The condition $\mathcal{D}_k^- \neq \emptyset$ means that, at each iteration $k \in K$, there exists at least one admissible realization of the additional sample that violates the acceptance condition (5.5).

Now observe that, for each $k \in K$, the size of the additional sample satisfies $1 \leq D_k \leq N_k = \bar{N}$. Since the full index set $\mathcal{N}$ is finite, there are only finitely many subsets of $\mathcal{N}$ of cardinality at most $\bar{N}$. Therefore, the family of all admissible realizations of $\mathcal{D}_k$ that may occur for indices $k \in K$ is itself finite. Since the additional sample is chosen randomly and uniformly from the admissible family at each

iteration, every admissible realization has strictly positive probability. Taking the minimum over a finite collection of positive numbers, we conclude that there exists a constant $q > 0$ such that

$$\mathbb{P}(\mathcal{D}_k \in \mathcal{D}_k^-) \geq q \quad \text{for all } k \in K.$$

By independence of the additional samples across iterations, the probability that *all* these infinitely many tests are successful is bounded by

$$\mathbb{P}(\mathcal{D}_k \in \mathcal{D}_k^+, k \in K) \leq \prod_{k \in K} (1 - q) = 0.$$

Hence, a.s. there exists an iteration $\tilde{k} \in K$ such that

$$\mathcal{D}_{\tilde{k}} \in \mathcal{D}_{\tilde{k}}^-.$$

According to ASPEN, this implies that the sample size must be increased at iteration $\tilde{k}$, that is,

$$N_{\tilde{k}+1} > N_{\tilde{k}} = \bar{N}.$$

But this contradicts the stabilization property $N_k = \bar{N}$ for all $k \geq k_0$. Therefore, the assumption was false, and the conclusion follows. $\square$

Lemma 5.3.2 has a clear intuitive meaning: if the algorithm remains forever in the MB regime, then the current sample size must eventually become sufficiently informative so that the additional-sampling test accepts every admissible auxiliary sample. In other words, once the sample size stops changing, the quality test can no longer keep failing.

The next lemma shows that the additional-sampling acceptance condition eventually implies a descent inequality for the *full* penalty function, even though the algorithm keeps operating with subsamples.

Lemma 5.3.3 *Suppose that the assumptions of Lemma 5.3.2 hold. Then, a.s., the following inequality holds for all $k \geq k_1$:*

$$F(x_{k+1}, \mu_k) \leq F(x_k, \mu_k) - c \|\nabla F(x_k, \mu_k)\|^2 + C\epsilon_k.$$

Proof. By Lemma 5.3.2, a.s. we have $\mathcal{D}_k^- = \emptyset$ for all $k \geq k_1$. Hence, for all sufficiently large k , the acceptance condition (5.5) is satisfied for every admissible realization of the additional sample. In particular, the candidate point is accepted, so that

$$x_{k+1} = \bar{x}_k \quad \text{for all } k \geq k_1.$$

We now pass from the additional-sampling inequality to inequalities involving the individual local penalty functions. The key observation is that, once every admissible additional sample is accepted, no local component can violate the corresponding decrease condition. Otherwise, one could construct an admissible additional sample built around such a violating component, contradicting the fact that $\mathcal{D}_k^- = \emptyset$.

Therefore, for each $i \in \mathcal{N}$ and each $k \geq k_1$, we obtain

$$F_i(x_{k+1}, \mu_k) \leq F_i(x_k, \mu_k) - c \|\nabla F_i(x_k, \mu_k)\|^2 + C\epsilon_k, \quad (5.7)$$

where

$$F_i(x, \mu_k) := f_i(x) + \frac{\mu_k}{2} \|h(x)\|^2.$$

Summing (5.7) over $i = 1, \dots, N$ and dividing by N , we get

$$F(x_{k+1}, \mu_k) \leq F(x_k, \mu_k) - c \frac{1}{N} \sum_{i=1}^N \|\nabla F_i(x_k, \mu_k)\|^2 + C\epsilon_k. \quad (5.8)$$

Finally, by convexity of the squared norm,

$$\|\nabla F(x_k, \mu_k)\|^2 = \left\| \frac{1}{N} \sum_{i=1}^N \nabla F_i(x_k, \mu_k) \right\|^2 \leq \frac{1}{N} \sum_{i=1}^N \|\nabla F_i(x_k, \mu_k)\|^2.$$

Combining this estimate with (5.8) yields

$$F(x_{k+1}, \mu_k) \leq F(x_k, \mu_k) - c \|\nabla F(x_k, \mu_k)\|^2 + C\epsilon_k,$$

which proves the claim. $\square$

Before stating the next theorem, let us denote by $\mathbb{E}_{MB}(\cdot)$ and $\mathbb{E}_{FS}(\cdot)$ the conditional expectations with respect to all sample paths of ASPEN that fall into the MB and FS scenarios, respectively. Similarly, we denote by $\mathbb{P}_{MB}(\cdot)$ and $\mathbb{P}_{FS}(\cdot)$ the corresponding conditional probabilities.

We now prove that, in the MB scenario, the norm of the gradient of the full penalty function converges to zero almost surely.

Theorem 5.3.1 *Suppose that Assumption 1 holds and that the sequence $\{x_k\}_{k \in \mathbb{N}}$ generated by ASPEN is bounded. Then*

$$\mathbb{P}_{MB}\left(\lim_{k \rightarrow \infty} \nabla F(x_k, \mu_k) = 0\right) = 1.$$

Proof. Since we restrict attention to sample paths belonging to the MB scenario, Lemma 5.3.3 implies that, a.s., for every $l \in \mathbb{N}$,

$$F(x_{k_1+l}, \mu_{k_1+l}) \leq F(x_{k_1}, \mu_{k_1}) - c \sum_{j=0}^{l-1} \|\nabla F(x_{k_1+j}, \mu_{k_1+j})\|^2 + C \sum_{j=0}^{l-1} \epsilon_{k_1+j}.$$

The boundedness of the iterates ensures that the values of the penalty function along the generated sequence remain controlled from below, while the summability of $\{\epsilon_k\}$ implies that the cumulative error term is finite. Therefore, after taking the conditional expectation $\mathbb{E}_{MB}(\cdot)$ and letting $l \rightarrow \infty$, we obtain

$$\sum_{j=0}^{\infty} \mathbb{E}_{MB}(\|\nabla F(x_{k_1+j}, \mu_{k_1+j})\|^2) < \infty.$$

In particular, the sequence of nonnegative random variables

$$\|\nabla F(x_{k_1+j}, \mu_{k_1+j})\|^2$$

is summable in expectation. By the extended Markov inequality, for every $\delta > 0$,

$$\sum_{j=0}^{\infty} \mathbb{P}_{MB}(\|\nabla F(x_{k_1+j}, \mu_{k_1+j})\| > \delta) < \infty.$$

An application of the Borel–Cantelli lemma then shows that, for every fixed $\delta > 0$, the event

$$\|\nabla F(x_{k_1+j}, \mu_{k_1+j})\| > \delta$$

can occur only finitely many times almost surely. Since $\delta > 0$ is arbitrary, this yields

$$\lim_{k \rightarrow \infty} \nabla F(x_k, \mu_k) = 0 \quad \text{a.s. on the MB scenario.}$$

This proves the theorem. $\square$

We next show that, under LICQ, every accumulation point generated in the MB scenario is a KKT point of the original equality-constrained problem. The argument splits naturally according to the asymptotic behavior of the penalty parameter sequence.

Theorem 5.3.2 *Let the assumptions of Theorem 5.3.1 hold and assume that $N_k < N$ for every $k \in \mathbb{N}$. Then, a.s., every accumulation point of the sequence $\{x_k\}_{k \in \mathbb{N}}$ at which LICQ holds is a KKT point of problem (5.1).*

Proof. By Theorem 5.3.1, we know that

$$\lim_{k \rightarrow \infty} \nabla F(x_k, \mu_k) = 0 \quad \text{a.s.}$$

Let x^* be an arbitrary accumulation point of the sequence $\{x_k\}$, that is,

$$\lim_{k \in K} x_k = x^*$$

for some infinite index set $K \subseteq \mathbb{N}$. We distinguish two possible cases according to the behavior of the penalty parameter sequence.

Case 1: The sequence $\{\mu_k\}$ is bounded.

Since $\{\mu_k\}$ is nondecreasing, boundedness implies that there exists $\bar{\mu} > 0$ such that

$$\mu_k = \bar{\mu} \quad \text{for all sufficiently large } k.$$

From the update rule in the MB phase, this means that for all sufficiently large k we have

$$\|h(x_k)\| \leq \epsilon_k.$$

Because the sequence $\{\epsilon_k\}$ is summable, it follows in particular that $\epsilon_k \rightarrow 0$, and hence

$$\lim_{k \rightarrow \infty} h(x_k) = 0.$$

Therefore, every accumulation point is feasible, so

$$h(x^*) = 0.$$

Moreover, since $\mu_k = \bar{\mu}$ for large k and $\nabla F(x_k, \mu_k) \rightarrow 0$, a.s., we obtain along the subsequence K ,

$$0 = \lim_{k \in K} \nabla F(x_k, \mu_k) = \lim_{k \in K} (\nabla f(x_k) + \mu_k \nabla h(x_k)^T h(x_k)) = \nabla f(x^*),$$

because $h(x_k) \rightarrow 0$ and all relevant mappings are continuous. Thus,

$$\nabla f(x^*) = 0, \quad h(x^*) = 0,$$

so x^* is almost surely a KKT point with multiplier $\lambda^* = 0$.

Case 2: The sequence $\{\mu_k\}$ is unbounded.

Assume now that

$$\lim_{k \rightarrow \infty} \mu_k = \infty.$$

From the identity

$$\nabla F(x_k, \mu_k) = \nabla f(x_k) + \mu_k \nabla h(x_k)^T h(x_k),$$

we obtain

$$\|\nabla h(x_k)^T h(x_k)\| \leq \frac{1}{\mu_k} \left(\|\nabla F(x_k, \mu_k)\| + \|\nabla f(x_k)\| \right).$$

Passing to the limit along K and using $\mu_k \rightarrow \infty$ together with the boundedness of the iterates, we conclude that

$$\|\nabla h(x^*)^T h(x^*)\| = 0.$$

Since LICQ holds at x^* , the rows of $\nabla h(x^*)$ are linearly independent. Equivalently, the matrix

$$\nabla h(x^*) \nabla h(x^*)^T$$

is nonsingular. Therefore, the equality

$$\nabla h(x^*)^T h(x^*) = 0$$

can only hold if

$$h(x^*) = 0.$$

Hence, x^* is feasible.

To identify the Lagrange multiplier, define

$$\lambda_k := \mu_k h(x_k).$$

Since LICQ holds at x^* and $\nabla h(\cdot)$ is continuous, the matrix

$$\nabla h(x_k) \nabla h(x_k)^T$$

is nonsingular for all sufficiently large $k \in K$. Multiplying the identity

$$\nabla F(x_k, \mu_k) - \nabla f(x_k) = \nabla h(x_k)^T \lambda_k$$

from the left by

$$(\nabla h(x_k) \nabla h(x_k)^T)^{-1} \nabla h(x_k),$$

we obtain

$$\lambda_k = (\nabla h(x_k) \nabla h(x_k)^T)^{-1} \nabla h(x_k) (\nabla F(x_k, \mu_k) - \nabla f(x_k)). \quad (5.9)$$

Passing to the limit along K , we find

$$\lim_{k \in K} \lambda_k = -(\nabla h(x^*) \nabla h(x^*)^T)^{-1} \nabla h(x^*) \nabla f(x^*) =: \lambda^*. \quad (5.10)$$

Finally, taking limits in

$$\nabla F(x_k, \mu_k) = \nabla f(x_k) + \nabla h(x_k)^T \lambda_k$$

gives

$$0 = \lim_{k \in K} \nabla F(x_k, \mu_k) = \nabla f(x^*) + \nabla h(x^*)^T \lambda^*, \text{ a.s.} \quad (5.11)$$

Together with $h(x^*) = 0$, this is precisely the KKT system for problem (5.1). Hence, x^* is almost surely a KKT point. $\square$

The previous proof clearly separates the two possible asymptotic mechanisms. If the penalty parameter eventually stabilizes, then feasibility is recovered directly through the update rule, and stationarity follows from the vanishing gradient of the penalty function. If, on the other hand, the penalty parameter tends to infinity, then the scaled constraint residuals $\lambda_k = \mu_k h(x_k)$ play the role of approximate multipliers, and the KKT system is obtained in the limit.

We now consider the FS scenario.

Theorem 5.3.3 *Suppose that Assumption 1 holds and that the sequence $\{x_k\}_{k \in \mathbb{N}}$ generated by ASPEN is bounded. Moreover, suppose that there exists $\tilde{k}$ such that $N_k = N$ for all $k \geq \tilde{k}$. Then, a.s., there exists an accumulation point of the sequence $\{x_k\}_{k \in \mathbb{N}}$ which is a KKT point of problem (5.1), provided that LICQ holds at that point.*

Proof. Let us consider iterations $k \geq \tilde{k}$. Then $N_k = N$, and the line search condition reduces to

$$F(x_{k+1}, \mu_k) \leq F(x_k, \mu_k) - \eta \alpha_k \|\nabla F(x_k, \mu_k)\|^2 + \epsilon_k.$$

In the FS phase, the penalty parameter is increased only if

$$\|\nabla F(x_k, \mu_k)\| < \frac{1}{\mu_k}.$$

Assume, for contradiction, that μ_k remains fixed at some value $\bar{\mu}$ for infinitely many consecutive iterations. Then, using the lower bound (5.6) on the accepted step sizes, we obtain

$$F(x_{k+1}, \bar{\mu}) \leq F(x_k, \bar{\mu}) - \eta \bar{\alpha}_{\bar{\mu}} \|\nabla F(x_k, \bar{\mu})\|^2 + \epsilon_k.$$

Since the sequence of iterates is bounded and $\{\epsilon_k\}$ is summable, the usual descent argument implies that

$$\sum_{k=\tilde{k}}^{\infty} \|\nabla F(x_k, \bar{\mu})\|^2 < \infty,$$

and therefore

$$\|\nabla F(x_k, \bar{\mu})\| \rightarrow 0.$$

But then, after finitely many iterations, we must have

$$\|\nabla F(x_k, \bar{\mu})\| < \frac{1}{\bar{\mu}},$$

which triggers an increase of the penalty parameter. This contradicts the assumption that μ_k stays equal to $\bar{\mu}$ indefinitely. Hence, the penalty parameter cannot remain fixed forever.

We conclude that μ_k must be increased infinitely many times, and therefore

$$\lim_{k \rightarrow \infty} \mu_k = \infty.$$

Moreover, there exists an infinite subset of indices $\tilde{K} \subseteq \mathbb{N}$ such that

$$\|\nabla F(x_k, \mu_k)\| < \frac{1}{\mu_k} \quad \text{for all } k \in \tilde{K}.$$

Since $\mu_k \rightarrow \infty$, it follows that

$$\lim_{k \in \tilde{K}} \nabla F(x_k, \mu_k) = 0.$$

Because the sequence $\{x_k\}$ is bounded, there exist an accumulation point x^* and an infinite subset $\tilde{K}_1 \subseteq \tilde{K}$ such that

$$\lim_{k \in \tilde{K}_1} x_k = x^*.$$

At this point the argument is identical to the second part of the proof of Theorem 5.3.2: the fact that $\mu_k \rightarrow \infty$ and $\nabla F(x_k, \mu_k) \rightarrow 0$ along $\tilde{K}_1$ yields feasibility of x^* and, via the multiplier sequence $\lambda_k = \mu_k h(x_k)$, the KKT system at the limit point. Therefore, x^* is a KKT point of problem (5.1). $\square$

We are now ready to combine the two possible scenarios into the final convergence statement.

Theorem 5.3.4 *Suppose that Assumption 1 holds and that the sequence $\{x_k\}_{k \in \mathbb{N}}$ generated by ASPEN is bounded. Then, a.s., there exists an accumulation point of the sequence $\{x_k\}_{k \in \mathbb{N}}$ which is a KKT point of problem (5.1), provided that LICQ holds.*

Proof. The conclusion follows directly by combining Theorems 5.3.2 and 5.3.3. Indeed, every sample path of ASPEN belongs either to the MB scenario or to the FS scenario. In the first case, every accumulation point satisfying LICQ is a KKT point by Theorem 5.3.2. In the second case, at least one accumulation point satisfying LICQ is a KKT point by Theorem 5.3.3. Therefore, in all admissible sample paths, ASPEN admits a KKT accumulation point almost surely. $\square$

The convergence analysis above covers both structural regimes of ASPEN. In the MB regime, additional sampling guarantees that the method eventually behaves like a descent method for the full penalty function, which leads to vanishing gradient norms and, under LICQ, to KKT points of the original problem. In the FS regime, the method reduces to a deterministic penalty scheme with adaptive updates of the penalty parameter, and the same KKT conclusion is recovered. Hence, regardless of whether the method remains in the mini-batch phase or eventually reaches the full sample, ASPEN converges almost surely to a KKT point in the sense stated above.

5.4 Extension to the weighted finite sum setting

The ASPEN method can also be applied to the case in which the objective function is given in the weighted finite sum form. In that setting, instead of the problem as in (5.1), we consider

$$\min_x f(x) := \sum_{i=1}^N w_i f_i(x), \quad \text{subject to} \quad h(x) = 0, \quad (5.12)$$

where the weights satisfy

$$w_i \geq 0, \quad \sum_{i=1}^N w_i = 1,$$

and the functions $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \dots, N$, as well as the constraint mapping $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$, are continuously differentiable.

This formulation is relevant in situations where different local objective functions do not have the same contribution to the overall model. In such cases, the weights reflect the relative importance of the local terms and must therefore be incorporated directly into the sampling procedure. In the present extension, this is done in exactly the same spirit as in the bound-constrained weighted setting: the weights are used as sampling probabilities.

More precisely, at iteration k , let

$$\mathcal{N}_k = \{i_1^k, \dots, i_{N_k}^k\}$$

denote the main sample, where each sampled index i_j^k takes the value $s \in \mathcal{N} := \{1, \dots, N\}$ with probability

$$\mathbb{P}(i_j^k = s) = w_s, \quad s = 1, \dots, N, \quad j = 1, \dots, N_k. \quad (5.13)$$

The corresponding sample-based approximation of the objective function is then defined by

$$f_{\mathcal{N}_k}(x) := \frac{1}{N_k} \sum_{j=1}^{N_k} f_{i_j^k}(x), \quad (5.14)$$

and its gradient approximation is given by

$$\nabla f_{\mathcal{N}_k}(x) := \frac{1}{N_k} \sum_{j=1}^{N_k} \nabla f_{i_j^k}(x). \quad (5.15)$$

With this construction, the sampled objective and gradient are unbiased estimators of the weighted objective and its gradient. Indeed,

$$\mathbb{E}(f_{\mathcal{N}_k}(x) \mid x) = \frac{1}{N_k} \sum_{j=1}^{N_k} \mathbb{E}(f_{i_j^k}(x) \mid x) = \frac{1}{N_k} \sum_{j=1}^{N_k} \sum_{s=1}^N w_s f_s(x) = f(x).$$

Analogously,

$$\mathbb{E}(\nabla f_{\mathcal{N}_k}(x) \mid x) = \nabla f(x).$$

As in the unweighted case, the equality constraints are handled through a quadratic penalty reformulation. Thus, we define

$$F(x, \mu) = f(x) + \frac{\mu}{2} \|h(x)\|^2,$$

and, at iteration k , the corresponding sampled penalty function becomes

$$F_{\mathcal{N}_k}(x, \mu_k) = f_{\mathcal{N}_k}(x) + \frac{\mu_k}{2} \|h(x)\|^2. \quad (5.16)$$

Its gradient is therefore

$$\nabla F_{\mathcal{N}_k}(x, \mu_k) = \nabla f_{\mathcal{N}_k}(x) + \mu_k \nabla h(x)^T h(x). \quad (5.17)$$

Consequently, the search direction in ASPEN is still chosen as

$$p_k = -\nabla F_{\mathcal{N}_k}(x_k, \mu_k),$$

and the nonmonotone line search condition keeps exactly the same form as in the standard finite sum case.

The additional sample is constructed analogously. Namely, let

$$\mathcal{D}_k = \{\ell_1^k, \dots, \ell_{D_k}^k\},$$

where each ℓ_j^k takes the value $s \in \mathcal{N}$ with probability

$$\mathbb{P}(\ell_j^k = s) = w_s, \quad s = 1, \dots, N, \quad j = 1, \dots, D_k. \quad (5.18)$$

Then the auxiliary sampled penalty function used in the acceptance test is defined by

$$F_{\mathcal{D}_k}(x, \mu_k) := \frac{1}{D_k} \sum_{j=1}^{D_k} f_{\ell_j^k}(x) + \frac{\mu_k}{2} \|h(x)\|^2. \quad (5.19)$$

Accordingly, the additional-sampling acceptance test retains the same structure as before, but with the weighted-sampling-based estimator:

$$F_{\mathcal{D}_k}(\bar{x}_k, \mu_k) \leq F_{\mathcal{D}_k}(x_k, \mu_k) - c \|\nabla F_{\mathcal{D}_k}(x_k, \mu_k)\|^2 + C\epsilon_k.$$

Therefore, the algorithmic form of ASPEN does not change in any essential way. The only modification lies in the fact that the main and additional samples are now drawn according to the probability distribution induced by the weights. Once the sample reaches the full size, that is, when $N_k = N$, the method again switches to the deterministic regime by using the full objective function.

From the theoretical point of view, the convergence analysis remains almost identical. The arguments based on smoothness of the penalty function, lower bounds on the line search step size, and the distinction between the MB and FS regimes are unaffected by the presence of weights. The only place where the proof needs a slight adaptation is in the passage from local inequalities to the full objective function. In the unweighted case this is done by averaging over all local terms, whereas in the weighted case one must form the corresponding convex combination with coefficients w_i .

Indeed, if

$$F_i(x, \mu_k) := f_i(x) + \frac{\mu_k}{2} \|h(x)\|^2,$$

then the full penalty function can be written as

$$F(x, \mu_k) = \sum_{i=1}^N w_i F_i(x, \mu_k),$$

and the full gradient satisfies

$$\nabla F(x_k, \mu_k) = \sum_{i=1}^N w_i \nabla F_i(x_k, \mu_k).$$

Hence, in the weighted analogue of Lemma 5.3.3, instead of dividing by N , one obtains

$$F(x_{k+1}, \mu_k) \leq F(x_k, \mu_k) - c \sum_{i=1}^N w_i \|\nabla F_i(x_k, \mu_k)\|^2 + C\epsilon_k.$$

Using convexity of the squared norm, we then have

$$\|\nabla F(x_k, \mu_k)\|^2 = \left\| \sum_{i=1}^N w_i \nabla F_i(x_k, \mu_k) \right\|^2 \leq \sum_{i=1}^N w_i \|\nabla F_i(x_k, \mu_k)\|^2,$$

which yields the same type of descent estimate as in the unweighted case.

Thus, the weighted finite sum setting does not require a new ASPEN scheme, but only a weighted interpretation of the sampling process. The structure of the method, the penalty mechanism, the line search, the additional-sampling test, and the distinction between the MB and FS regimes all remain the same. The only substantive modification is that uniform averaging is replaced by the convex combination induced by the weights.

This leads to the following conclusion.

Theorem 5.4.1 *Suppose that all assumptions required in Theorem 5.3.4 are satisfied for problem (5.12), and assume additionally that the main and additional samples are generated according to the weighted distributions (5.13) and (5.18). Then the convergence results established for ASPEN in Theorem 5.3.4 remain valid in the weighted*

setting as well. In particular, if the generated sequence $\{x_k\}$ is bounded and LICQ holds at the corresponding accumulation point, then the sequence generated by ASPEN has, almost surely, an accumulation point that is a KKT point of problem (5.12).

Proof. The proof follows the same line as in the standard finite sum case. The algorithmic structure is unchanged, and all arguments based on smoothness, line search well-definedness, and the MB/FS split remain valid without modification. The only adjustment appears in those steps where one passes from local descent inequalities to the full penalty function. In the weighted case, the arithmetic mean is replaced by the weighted sum with coefficients w_i , and the necessary estimate follows from convexity of the squared norm applied to the convex combination

$$\sum_{i=1}^N w_i \nabla F_i(x_k, \mu_k).$$

Therefore, the convergence analysis carries over directly. $\square$

In summary, the weighted equality-constrained finite sum problem is treated within the ASPEN framework in the same way as in the weighted bound-constrained setting: the weights are incorporated through the sampling distribution, while the overall structure of the method and its convergence analysis remain essentially unchanged.

Chapter 6

Additional Sampling Method for Bound-Constrained Problems

In the present section, we focus on the bound-constrained case and develop methods based on additional sampling and projected first-order directions. The results presented in this section represent an original scientific contribution and have been published in the paper

- N. KREJIĆ, N. KRKLEC JERINKIĆ, T. OSTOJIĆ, & N. VUČIĆEVIĆ, AS-BOX: Additional sampling method for weighted sum problems with box constraints, *Numerical Algorithms*, 1-26., 2025.

The analysis is carried out progressively. We first examine the simpler case of nonnegativity constraints, where the projection and active-set structure are particularly easy to describe. This allows the main ideas of the method to be introduced in a clearer form. After that, the same framework is extended to general box constraints. Thus, this section serves both as a methodological bridge from the general

problem setting introduced earlier and as the basis for the algorithms and convergence results developed for bound-constrained finite sum optimization problems.

6.1 Methods for Nonnegativity-Constrained Problems

6.1.1 Problem formulation

Within this subsection, we consider an important special case of the general bound-constrained problem (4.5), namely

$$\min_{x \geq 0} f(x), \tag{6.1}$$

where the objective function f is the same as in (4.4), and the inequalities $x \geq 0$ are understood componentwise. In other words, each component of the decision vector is required to remain nonnegative. Although this class of problems is a particular case of the more general box-constrained framework, it is sufficiently rich to capture many relevant applications and, at the same time, simple enough to make the structure of the proposed method more transparent.

A central feature of the method is that the objective function is not necessarily processed through full information at every iteration. Instead, the algorithm works with sampled approximations built from a main sample and an additional sample. This is particularly important in large-scale settings, where evaluating the whole finite sum and its gradient at each step may be unnecessarily expensive.

At iteration k , the main sample is denoted by

$$\mathcal{N}_k = \{i_1^k, \dots, i_{N_k}^k\},$$

where N_k is the sample size, and the sampled gradient is formed from the components selected through this sample. In the weighted finite

sum setting, the random indices are chosen according to the weights of the objective function. More precisely, for each draw we assume

$$P(i_j^k = s) = w_s, \quad s = 1, \dots, N, \quad j = 1, \dots, N_k, \quad (6.2)$$

where the weights satisfy $w_s \geq 0$ and $\sum_{s=1}^N w_s = 1$. In the special unweighted case, this reduces to the uniform rule $w_s = \frac{1}{N}$ for all $s = 1, \dots, N$. Thus, the sample is generated in a way that is consistent with the structure of the finite sum objective itself. The corresponding sampled gradient is written as

$$\nabla f_{N_k}(x) = \frac{1}{N_k} \sum_{j=1}^{N_k} \nabla f_{i_j^k}(x).$$

Besides the main sample, the method also uses an additional sample

$$\mathcal{D}_k = \{l_1^k, \dots, l_{D_k}^k\},$$

with sample size D_k . This second sample is not introduced to compute the search direction itself, but rather to provide an external check of the structural information extracted from the main sample. The indices in the additional sample are generated according to the same probability law,

$$P(l_j^k = s) = w_s, \quad s = 1, \dots, N, \quad j = 1, \dots, D_k. \quad (6.3)$$

The point of using the same distribution in (6.2) and (6.3) is that both samples represent the same underlying finite sum model, while still remaining distinct realizations. In this way, the additional sample can be used to verify whether the projection pattern suggested by the main sample is reliable. In practical implementations, D_k is typically chosen much smaller than N_k , since its purpose is validation rather

than direction construction. The associated sampled gradient is

$$\nabla f_{\mathcal{D}_k}(x) = \frac{1}{D_k} \sum_{j=1}^{D_k} \nabla f_{l_j^k}(x).$$

Compared to the general box-constrained setting, the feasible set in the nonnegativity-constrained case has a particularly simple geometry. As a consequence, the orthogonal projection onto the feasible set takes the componentwise form

$$[\pi_S(y)]_i = \max \{y_i, 0\}, \quad i = 1, \dots, n.$$

This explicit formula makes it possible to analyze the projected search direction in a very direct way. Since the direction used by the method is of the form

$$p_k = \pi_S(x_k - \nabla f_{\mathcal{N}_k}(x_k)) - x_k, \tag{6.4}$$

each component of p_k can be described according to whether the projected trial point hits the boundary of the feasible set or remains in its interior. More precisely, for every component $i \in \{1, \dots, n\}$, we distinguish the following two cases:

$$[p_k]_i = -[x_k]_i \quad \text{if} \quad [x_k]_i < [\nabla f_{\mathcal{N}_k}(x_k)]_i,$$

and

$$[p_k]_i = -[\nabla f_{\mathcal{N}_k}(x_k)]_i \quad \text{if} \quad [x_k]_i \geq [\nabla f_{\mathcal{N}_k}(x_k)]_i.$$

Hence, the projected direction either moves the iterate exactly to the boundary in a given component, or coincides with the negative gradient component whenever the projected point remains feasible without truncation. This simple dichotomy is one of the main reasons why the nonnegativity-constrained case is particularly suitable as the first step in the development of the method.

What matters here is not only the magnitude of the sampled gradient, but also the qualitative information it conveys about the active structure of the constraint. Indeed, in the present setting, each component of the projected point either stays strictly positive or is pushed to zero by the projection. Since this distinction determines the local geometry of the step, it is natural to encode it explicitly.

In order to describe this componentwise structure more systematically, we introduce the indicator vector $I_{\mathcal{N}_k}$ associated with the event $x_k < \nabla f_{\mathcal{N}_k}(x_k)$, where the inequality is understood componentwise. More precisely, for $i = 1, \dots, n$, we define

$$[I_{\mathcal{N}_k}]_i = \begin{cases} 1, & [x_k]_i < [\nabla f_{\mathcal{N}_k}(x_k)]_i, \\ 0, & [x_k]_i \geq [\nabla f_{\mathcal{N}_k}(x_k)]_i. \end{cases} \quad (6.5)$$

Thus, $[I_{\mathcal{N}_k}]_i = 1$ marks those components for which the sampled gradient would drive the projected trial point to the boundary, while $[I_{\mathcal{N}_k}]_i = 0$ corresponds to components for which the projection does not activate the nonnegativity constraint.

Analogously, using the additional sample $\mathcal{D}_k$, we define the indicator vector $I_{\mathcal{D}_k}$ by replacing $\nabla f_{\mathcal{N}_k}(x_k)$ with $\nabla f_{\mathcal{D}_k}(x_k)$. We then introduce the quantity

$$r_{\mathcal{D}_k} := \|I_{\mathcal{N}_k} - I_{\mathcal{D}_k}\|. \quad (6.6)$$

Since $\mathcal{N}_k$, $\mathcal{D}_k$, and x_k are random, the values $r_{\mathcal{D}_k}$ are random as well. Their role is to measure whether the projected directions generated by the two samples $\mathcal{N}_k$ and $\mathcal{D}_k$ have the same structural pattern, that is, whether they identify the same components as active at the boundary. In particular, if $r_{\mathcal{D}_k} = 0$, then the pattern of zero entries in $\pi_S(x_k - \nabla f_{\mathcal{N}_k}(x_k))$ coincides with the corresponding pattern in $\pi_S(x_k - \nabla f_{\mathcal{D}_k}(x_k))$. This means that the two samples induce the same active structure of the projection, even if the actual gradient values may differ.

This comparison of indicator vectors is a natural way to test whether two samples lead to the same projection behavior. In the nonnegativity-constrained setting, the decisive geometric information is not only how large the sampled gradient components are, but also whether they are large enough to force a coordinate to the boundary. The main sample provides one estimate of that behavior, while the additional sample offers an independent check drawn from the same probabilistic model. If the two samples produce the same activation pattern, then the current structural information can be regarded as stable. On the other hand, a discrepancy between I_{N_k} and I_{D_k} indicates that the present sample size may still be too small to identify the active set reliably.

For this reason, the quantities introduced through (6.2), (6.3), and (6.6) play a more important role than a simple auxiliary test. They provide the mechanism through which the method evaluates the consistency of the sampled projection structure. This observation will be essential later, when additional sampling is used both in the acceptance of candidate iterates and in the adaptive decision of whether the current sample size should be enlarged.

6.1.2 AS-NC algorithm

We now present the method designed for solving problem (6.1). Since the feasible set is the nonnegative orthant, the projection step admits a particularly simple componentwise form, which makes it possible to combine projected gradient directions with adaptive subsampling in a transparent way. The resulting procedure, denoted by AS-NC, is a stochastic projected-gradient-type method equipped with a nonmonotone line search and an additional sampling test used to decide both whether a candidate point should be accepted and whether the current sample size should be increased.

The algorithm starts from a feasible initial point and therefore

preserves feasibility throughout the whole iterative process. At each iteration, a main sample is first selected and used to compute the projected search direction. Then a backtracking nonmonotone line search is performed with respect to the sampled objective function. The resulting candidate point is afterwards checked on an independent additional sample. If this additional test confirms that the candidate provides sufficient progress and that the induced projection structure is compatible with the one obtained from the main sample, then the current sample size is retained. Otherwise, the sample size is enlarged. In this way, the method adapts its computational effort to the local difficulty of the problem.

The behavior of AS-NC can be naturally separated into two regimes. In the *mini-batch* (MB) regime, the full sample is never reached and the method continues to operate with subsampled information. In the *full sample* (FS) regime, there exists a finite iteration from which onward the algorithm always uses the complete objective function. In that case, the stochastic component of the method disappears from the objective and gradient evaluations, and the method reduces to a deterministic projected gradient scheme, although the entire trajectory still depends on the random choices made in the earlier MB stage.

In the MB phase, sampling appears twice: first in the construction of the main sample $\mathcal{N}_k$, and then in the construction of the additional sample $\mathcal{D}_k$. Although (6.2) gives one natural unbiased sampling rule for $\mathcal{N}_k$, the convergence analysis does not require this particular choice, and in fact allows arbitrary strategies for the main sample. On the other hand, the additional sample must satisfy two structural properties: it has to be chosen independently of the current main sample, and every local objective function must have a positive probability of being selected. These conditions ensure that the additional-sampling test remains informative and can detect lack of representativeness of the current main sample.

Algorithm 3 AS-NC: Additional Sampling for Nonnegativity Constraints

S0 *Initialization.* Input: $x_0 \geq 0$, $N_0 \in \mathbb{N}$, $\beta, c, c_1 \in (0, 1)$, $C > 0$, $\{\varepsilon_k\}$ satisfying (4.25). Set $k := 0$.

S1 *Subsampling.* If $N_k < N$, choose $\mathcal{N}_k$ such that (6.2) holds. Otherwise, set $f_{\mathcal{N}_k} = f$.

S2 *Search direction.* Compute

$$p_k = \pi_S(x_k - \nabla f_{\mathcal{N}_k}(x_k)) - x_k.$$

S3 *Step size.* Find the smallest $j \in \mathbb{N}_0$ such that $t_k = \beta^j$ satisfies

$$f_{\mathcal{N}_k}(x_k + t_k p_k) \leq f_{\mathcal{N}_k}(x_k) + c_1 t_k (\nabla f_{\mathcal{N}_k}(x_k))^T p_k + \varepsilon_k. \quad (6.7)$$

Set $\bar{x}_k = x_k + t_k p_k$.

S4 *Additional sampling.* If $N_k = N$, set $x_{k+1} = \bar{x}_k$, $k := k + 1$, and go to Step **S1**. Otherwise, choose $\mathcal{D}_k$ via (6.3) and compute

$$s_k = \pi_S(x_k - \nabla f_{\mathcal{D}_k}(x_k)) - x_k \quad (6.8)$$

and

$$r_{\mathcal{D}_k} = \|I_{\mathcal{N}_k} - I_{\mathcal{D}_k}\|.$$

S5 *Sample size update.* If

$$r_{\mathcal{D}_k} = 0 \quad \text{and} \quad f_{\mathcal{D}_k}(\bar{x}_k) \leq f_{\mathcal{D}_k}(x_k) - c\|s_k\|^2 + C\varepsilon_k, \quad (6.9)$$

set $N_{k+1} = N_k$.

Otherwise, choose $N_{k+1} \in \{N_k + 1, \dots, N\}$.

S6 *Iterate update.* If

$$f_{\mathcal{D}_k}(\bar{x}_k) \leq f_{\mathcal{D}_k}(x_k) - c\|s_k\|^2 + C\varepsilon_k,$$

set $x_{k+1} = \bar{x}_k$. Otherwise, set $x_{k+1} = x_k$.

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

The feasibility of all iterates follows directly from the construction of the algorithm. Since $x_0 \in S$ and the projected direction p_k is built from the projection operator, the segment joining x_k and $x_k + p_k$ remains in the feasible set. Consequently, the line search point $\bar{x}_k = x_k + t_k p_k$ is feasible for every accepted step length $t_k \in [0, 1]$. In the additional-sampling step, the quantity s_k plays the projected-gradient-type direction associated with the independently sampled function $f_{\mathcal{D}_k}$.

The criterion (6.9) should be interpreted as a combined test of quality and structural consistency. The condition

$$f_{\mathcal{D}_k}(\bar{x}_k) \leq f_{\mathcal{D}_k}(x_k) - c\|s_k\|^2 + C\varepsilon_k$$

checks whether the candidate point yields a sufficient amount of progress on independent data, while the condition $r_{\mathcal{D}_k} = 0$ verifies that the projection structure induced by $\mathcal{N}_k$ and $\mathcal{D}_k$ is consistent. If both requirements are satisfied, then the current sample size is regarded as adequate. Otherwise, the algorithm increases the size of the main sample in order to obtain a more reliable approximation of the full objective. Thus, the additional-sampling mechanism simultaneously controls point acceptance and sample size growth.

For clarity, the role of the additional-sampling step in AS-NC can be summarized as follows. After the projected direction p_k and the trial point $\bar{x}_k$ have been computed using the primary sample $\mathcal{N}_k$, an independent auxiliary sample $\mathcal{D}_k$ is used to compute the projected direction s_k and the discrepancy measure $r_{\mathcal{D}_k}$. The decrease condition evaluated on $f_{\mathcal{D}_k}$ determines whether the trial point is accepted, whereas the stronger test combining this decrease condition with $r_{\mathcal{D}_k} = 0$ determines whether the current sample size is kept or increased. Thus, the additional sample provides an independent check for the step acceptance decision and, at the same time, controls the adaptive update of the sample size.

6.1.3 Convergence analysis

In this subsection, we establish almost sure convergence properties of the AS-NC method and derive a corresponding complexity estimate. The analysis follows the general philosophy of additional-sampling methods: one first proves a descent inequality for the full objective function, then derives almost sure convergence of the projected stationarity measure, and finally deduces complexity bounds by quantifying the expected time needed to reach the full sample and the expected number of subsequent deterministic iterations.

We begin with the standard smoothness assumption imposed on the local objective functions.

Assumption A 2 *Each function f_i , $i = 1, \dots, N$, is continuously differentiable, bounded from below by a constant f_{low} , and has L -Lipschitz continuous gradient.*

Under this assumption, every sampled function $f_{\mathcal{N}_k}$ is also continuously differentiable and bounded from below, with the same Lipschitz constant for its gradient. This regularity is essential for the line search analysis, since it ensures that local first-order information can be used to control the variation of the sampled objective function along projected directions.

As is customary in the analysis of additional-sampling methods, we split the set of all admissible realizations of the additional sample into two complementary classes. The first class contains those realizations for which the additional-sampling acceptance condition is satisfied, while the second contains those for which it fails. More precisely, we define

$$\mathcal{D}_k^+ = \{ \mathcal{D}_k \subset \mathcal{N} \mid r_{\mathcal{D}_k} = 0, \quad f_{\mathcal{D}_k}(\bar{x}_k) \leq f_{\mathcal{D}_k}(x_k) - c\|s_k\|^2 + C\varepsilon_k \},$$

and

$$\mathcal{D}_k^- = \{ \mathcal{D}_k \subset \mathcal{N} \mid r_{\mathcal{D}_k} > 0 \quad \text{or} \quad f_{\mathcal{D}_k}(\bar{x}_k) > f_{\mathcal{D}_k}(x_k) - c\|s_k\|^2 + C\varepsilon_k \}.$$

The next lemma shows that, if the full sample is never reached, then after finitely many iterations every admissible realization of the additional sample must satisfy the acceptance criterion.

Lemma 6.1.1 *Suppose that Assumption 2 holds. If $N_k < N$ for all $k \in \mathbb{N}$, then a.s. there exists $k_1 \in \mathbb{N}$ such that $\mathcal{D}_k^- = \emptyset$ for all $k \geq k_1$.*

Proof. Given that the sequence $\{N_k\}$ is nondecreasing and bounded above by $N - 1$ in the present scenario, there exists an index $k_0 \in \mathbb{N}$ and an integer $\bar{N} \in \{1, \dots, N - 1\}$ such that

$$N_k = \bar{N} \quad \text{for all } k \geq k_0.$$

Indeed, a monotone sequence of integers that never reaches N must eventually become constant.

Assume now, aiming at a contradiction, that the statement of the lemma does not hold. Then there exists an infinite set $K \subseteq \mathbb{N}$ such that

$$\mathcal{D}_k^- \neq \emptyset \quad \text{for all } k \in K.$$

Without loss of generality, we may suppose that $K \subseteq \{k \in \mathbb{N} : k \geq k_0\}$, so that $N_k = \bar{N}$ for every $k \in K$.

The condition $\mathcal{D}_k^- \neq \emptyset$ means that for every $k \in K$ there exists at least one admissible realization of the additional sample for which the acceptance criterion fails. Since $D_k \leq N_k = \bar{N}$, only finitely many values of D_k are possible, and for each such value only finitely many realizations of $\mathcal{D}_k$ can occur. Hence the family of all admissible realizations that may appear for indices in K is finite.

Because the additional sample is generated in such a way that each local objective has positive probability of being selected, every admissible realization of $\mathcal{D}_k$ has strictly positive probability. Taking the minimum over a finite family of positive numbers, we conclude that there exists $q > 0$ such that

$$\mathbb{P}(\mathcal{D}_k \in \mathcal{D}_k^-) \geq q \quad \text{for all } k \in K.$$

Therefore,

$$\mathbb{P}(\mathcal{D}_k \in \mathcal{D}_k^+, k \in K) \leq \prod_{k \in K} (1 - q) = 0.$$

This implies that, a.s., there exists an iteration $\tilde{k} \in K$ for which

$$\mathcal{D}_{\tilde{k}} \in \mathcal{D}_{\tilde{k}}^-.$$

According to Step S5 of AS-NC, this necessarily yields

$$N_{\tilde{k}+1} > N_{\tilde{k}} = \bar{N},$$

which contradicts the fact that the sequence $\{N_k\}$ has already stabilized at the value $\bar{N}$. The contradiction proves the result. $\square$

The following lemma gives the standard lower bound for the step size produced by the backtracking line search.

Lemma 6.1.2 *Suppose that Assumption 2 holds. Then the step size t_k obtained in Step S3 satisfies*

$$t_k \geq t_{\min} := \min \left\{ 1, \frac{2\beta(1 - c_1)}{L} \right\}.$$

Proof. Fix an iteration k . Since $\nabla f_{\mathcal{N}_k}$ is L -Lipschitz continuous, the descent lemma implies that for every $t \in [0, 1]$,

$$f_{\mathcal{N}_k}(x_k + tp_k) \leq f_{\mathcal{N}_k}(x_k) + t \nabla f_{\mathcal{N}_k}(x_k)^T p_k + \frac{L}{2} t^2 \|p_k\|^2.$$

From the properties of the projected gradient, we have that the projected direction satisfies

$$\nabla f_{\mathcal{N}_k}(x_k)^T p_k \leq -\|p_k\|^2.$$

Hence,

$$f_{\mathcal{N}_k}(x_k + tp_k) \leq f_{\mathcal{N}_k}(x_k) + t\nabla f_{\mathcal{N}_k}(x_k)^T p_k - \frac{L}{2}t^2\nabla f_{\mathcal{N}_k}(x_k)^T p_k.$$

Therefore, if

$$t \leq \frac{2(1 - c_1)}{L},$$

then

$$f_{\mathcal{N}_k}(x_k + tp_k) \leq f_{\mathcal{N}_k}(x_k) + c_1 t \nabla f_{\mathcal{N}_k}(x_k)^T p_k.$$

Since $\varepsilon_k > 0$, this certainly implies the line search condition (6.7). Thus every step size not exceeding $2(1 - c_1)/L$ is acceptable. Because the backtracking procedure tests the sequence $1, \beta, \beta^2, \dots$, the accepted step is bounded below by

$$t_k \geq \min \left\{ 1, \beta \frac{2(1 - c_1)}{L} \right\} = \min \left\{ 1, \frac{2\beta(1 - c_1)}{L} \right\}.$$

This proves the claim. $\square$

We now prove the central descent property of AS-NC. The statement is expressed in terms of the full objective function and the exact projected gradient mapping

$$d(x) := \pi_S(x - \nabla f(x)) - x,$$

regardless of whether the method is in the FS or MB regime.

Theorem 6.1.1 *Suppose that Assumption 2 holds. Then a.s. there exists a finite random iteration $\tilde{k}$ such that for all $k \geq \tilde{k}$,*

$$f(x_{k+1}) \leq f(x_k) - \bar{c}\|d(x_k)\|^2 + \bar{C}\varepsilon_k, \quad (6.10)$$

where

$$\bar{c} = \min \left\{ c, c_1, \frac{2c_1(1 - c_1)\beta}{L} \right\}, \quad \bar{C} = \max\{1, C\}.$$

Proof. We begin with the FS scenario. In this case there exists a finite random index $\tilde{k}_1$ such that for all $k \geq \tilde{k}_1$ the full sample is used, hence $f_{\mathcal{N}_k} = f$. The line search condition (6.7) then gives

$$f(x_{k+1}) \leq f(x_k) + c_1 t_k \nabla f(x_k)^T d(x_k) + \varepsilon_k.$$

Using Lemma 1.1.6 (1.34), we obtain

$$f(x_{k+1}) \leq f(x_k) - c_1 t_k \|d(x_k)\|^2 + \varepsilon_k.$$

By Lemma 6.1.2, $t_k \geq t_{min}$, and therefore

$$f(x_{k+1}) \leq f(x_k) - c_1 t_{min} \|d(x_k)\|^2 + \varepsilon_k. \quad (6.11)$$

We now turn to the MB scenario. By Lemma 6.1.1, a.s. there exists a finite random index k_1 such that $\mathcal{D}_k^- = \emptyset$ for all $k \geq k_1$. This means that for all sufficiently large iterations the acceptance condition (6.9) is satisfied for every admissible realization of the additional sample.

In particular, the second part of (6.9) must hold for all singleton realizations corresponding to individual local functions. Therefore, for each $j \in \mathcal{N}$ and every $k \geq k_1$,

$$f_j(\bar{x}_k) \leq f_j(x_k) - c \|s_k^j\|^2 + C\varepsilon_k,$$

where

$$s_k^j := \pi_S(x_k - \nabla f_j(x_k)) - x_k.$$

Since in this regime the candidate point is accepted, i.e., $x_{k+1} = \bar{x}_k$, multiplying the above inequality by w_j and summing over $j = 1, \dots, N$, we obtain

$$f(x_{k+1}) \leq f(x_k) - c \sum_{j=1}^N w_j \|s_k^j\|^2 + C\varepsilon_k. \quad (6.12)$$

It remains to relate the full projected gradient mapping $d(x_k)$ to the family of local projected directions $\{s_k^j\}_{j=1}^N$. For this purpose, consider the first condition in (6.9). Define the index set

$$\mathcal{A}_{N_k} := \{i \in \{1, \dots, n\} \mid [x_k]_i < [\nabla f_{N_k}(x_k)]_i\}. \quad (6.13)$$

Thus, $\mathcal{A}_{N_k}$ contains exactly those components for which the projected step with respect to the main sample hits the boundary.

Now, because $\mathcal{D}_k^- = \emptyset$ for all large k , the condition $r_{\mathcal{D}_k} = 0$ must also hold for singleton additional samples. Consequently, for all $k \geq k_1$ one has

$$\mathcal{A}_{N_k} = \mathcal{A}_k^1 = \dots = \mathcal{A}_k^N, \quad (6.14)$$

where

$$\mathcal{A}_k^j := \{i \in \{1, \dots, n\} \mid [x_k]_i < [\nabla f_j(x_k)]_i\}, \quad j = 1, \dots, N.$$

In other words, for large iterations all local projected directions have the same active set as the one generated by the main sample.

Therefore, for every $i \in \mathcal{A}_{N_k}$ and every $j \in \mathcal{N}$,

$$[x_k]_i < [\nabla f_j(x_k)]_i.$$

Multiplying this inequality by w_j and summing over j , we obtain

$$[x_k]_i < [\nabla f(x_k)]_i \quad \text{for all } i \in \mathcal{A}_{N_k}.$$

Similarly,

$$[x_k]_i \geq [\nabla f(x_k)]_i \quad \text{for all } i \notin \mathcal{A}_{N_k}.$$

Hence, by the definition of the full projected gradient mapping,

$$[d(x_k)]_i = \begin{cases} -[x_k]_i, & i \in \mathcal{A}_{N_k}, \\ -[\nabla f(x_k)]_i, & i \notin \mathcal{A}_{N_k}. \end{cases} \quad (6.15)$$

We may now estimate the norm of $d(x_k)$:

$$\begin{aligned}\|d(x_k)\|^2 &= \sum_{i=1}^n ([d(x_k)]_i)^2 \\ &= \sum_{i \in \mathcal{A}_{N_k}} ([x_k]_i)^2 + \sum_{i \notin \mathcal{A}_{N_k}} ([\nabla f(x_k)]_i)^2.\end{aligned}\quad (6.16)$$

By (6.14), for every $k \geq k_1$, every $j \in \mathcal{N}$, and every $i \in \mathcal{A}_{N_k}$, one has

$$[s_k^j]_i = -[x_k]_i,$$

so that

$$\sum_{j=1}^N w_j ([s_k^j]_i)^2 = ([x_k]_i)^2, \quad i \in \mathcal{A}_{N_k}.\quad (6.17)$$

Summing over all such indices yields

$$\sum_{i \in \mathcal{A}_{N_k}} \sum_{j=1}^N w_j ([s_k^j]_i)^2 = \sum_{i \in \mathcal{A}_{N_k}} ([x_k]_i)^2.\quad (6.18)$$

On the other hand, for every $i \notin \mathcal{A}_{N_k}$ and every $j \in \mathcal{N}$, one has

$$[s_k^j]_i = -[\nabla f_j(x_k)]_i.$$

Therefore, by convexity of the square function,

$$([\nabla f(x_k)]_i)^2 = \left(\sum_{j=1}^N w_j [\nabla f_j(x_k)]_i \right)^2 \leq \sum_{j=1}^N w_j ([\nabla f_j(x_k)]_i)^2 = \sum_{j=1}^N w_j ([s_k^j]_i)^2.\quad (6.19)$$

Summing this estimate over all $i \notin \mathcal{A}_{N_k}$ gives

$$\sum_{i \notin \mathcal{A}_{N_k}} ([\nabla f(x_k)]_i)^2 \leq \sum_{i \notin \mathcal{A}_{N_k}} \sum_{j=1}^N w_j ([s_k^j]_i)^2.\quad (6.20)$$

Combining (6.16), (6.18), and (6.20), we obtain

$$\begin{aligned} \|d(x_k)\|^2 &\leq \sum_{i \in \mathcal{A}_{\mathcal{N}_k}} \sum_{j=1}^N w_j ([s_k^j]_i)^2 + \sum_{i \notin \mathcal{A}_{\mathcal{N}_k}} \sum_{j=1}^N w_j ([s_k^j]_i)^2 \\ &= \sum_{j=1}^N w_j \|s_k^j\|^2. \end{aligned} \quad (6.21)$$

Substituting this estimate into (6.12), we conclude that for all $k \geq k_1$,

$$f(x_{k+1}) \leq f(x_k) - c\|d(x_k)\|^2 + C\varepsilon_k. \quad (6.22)$$

Finally, both scenarios lead to the same form of sufficient decrease for the full objective function. In the FS case this is given by (6.11), whereas in the MB case it is given by (6.22). Hence, choosing

$$\tilde{k} := \max\{\tilde{k}_1, k_1\},$$

we obtain (6.10) for all $k \geq \tilde{k}$, which completes the proof. $\square$

The significance of Theorem 6.1.1 is that, independently of whether the method remains in the mini-batch regime or eventually reaches the full sample, one ultimately obtains the same type of descent inequality for the *full* objective function. This estimate is the key tool from which both almost sure convergence and complexity bounds follow.

To derive almost sure convergence, we impose the following integrability condition, standard in the additional-sampling literature.

Assumption A 3 *There exists a constant C_b such that*

$$\mathbb{E}(|f(x_{\tilde{k}})|) \leq C_b,$$

where $\tilde{k}$ denotes the random iteration index used in Theorem 6.1.1.

This assumption is automatically satisfied whenever the sequence $\{f(x_k)\}$ is bounded. In particular, this happens in many practically relevant settings, for example when the feasible set is compact. However, Assumption 3 is more general: it does not require pathwise boundedness of the sampled objective values, but only boundedness of their expectation at the random time $\tilde{k}$.

The notation $\mathbb{E}_{FS}(\cdot) := \mathbb{E}(\cdot \mid FS)$ is used for the conditional expectation with respect to sample paths belonging to the FS scenario, and analogously by $\mathbb{E}_{MB}(\cdot) := \mathbb{E}(\cdot \mid MB)$ the conditional expectation with respect to MB sample paths. Then, as shown in the additional-sampling literature [45, 46], Assumption 3 implies the existence of constants C_b^{FS} and C_b^{MB} such that

$$\mathbb{E}_{FS}(|f(x_{\tilde{k}_1})|) \leq C_b^{FS}, \quad \mathbb{E}_{MB}(|f(x_{k_1})|) \leq C_b^{MB}, \quad (6.23)$$

where $\tilde{k}_1$ and k_1 are the random indices introduced in the proof of Theorem 6.1.1.

We can now state the main almost sure convergence result for AS-NC.

Theorem 6.1.2 *Suppose that Assumptions 2 and 3 hold. Then a.s. every accumulation point of the sequence $\{x_k\}_{k \in \mathbb{N}}$ generated by AS-NC is a stationary point of problem (6.1).*

Proof. By Theorem 6.1.1, a.s. for every $l \in \mathbb{N}$,

$$f(x_{\tilde{k}+l}) \leq f(x_{\tilde{k}}) - \bar{c} \sum_{j=0}^{l-1} \|d(x_{\tilde{k}+j})\|^2 + \bar{C} \sum_{j=0}^{l-1} \varepsilon_{\tilde{k}+j}.$$

Taking expectations, using Assumption 3, and letting $l \rightarrow \infty$, we obtain

$$\sum_{j=0}^{\infty} \mathbb{E}(\|d(x_{\tilde{k}+j})\|^2) < \infty.$$

Hence,

$$\sum_{j=0}^{\infty} \mathbb{P}(\|d(x_{\tilde{k}+j})\| > \delta) < \infty \quad \text{for every } \delta > 0,$$

by the extended Markov inequality. The Borel–Cantelli lemma ([63]) then yields

$$\mathbb{P}\left(\lim_{k \rightarrow \infty} d(x_k) = 0\right) = 1. \quad (6.24)$$

Suppose that x^* is an arbitrary accumulation point of $\{x_k\}$, and let $K_0 \subset \mathbb{N}$ be an infinite subsequence such that

$$\lim_{k \in K_0} x_k = x^*.$$

Passing to the limit along K_0 in (6.24), and using continuity of both the gradient and the projection operator, we obtain a.s.

$$\begin{aligned} 0 &= \lim_{k \in K_0} d(x_k) = \lim_{k \in K_0} (\pi_S(x_k - \nabla f(x_k)) - x_k) \\ &= \pi_S(x^* - \nabla f(x^*)) - x^* = d(x^*). \end{aligned}$$

Since every iterate is feasible and feasibility is preserved by the algorithm, we have $x^* \in S$. Therefore, Definition 1.1.10 implies that x^* is a stationary point of problem (6.1). This proves the claim. $\square$

We next address complexity. The aim is to estimate the expected number of iterations required to reach an iterate whose projected stationarity measure is below a prescribed tolerance.

Assumption A 4 *For each k there exists at least one function f_i such that the condition (6.9),*

$$r_{\mathcal{D}_k} = 0 \quad \text{and} \quad f_{\mathcal{D}_k}(\bar{x}_k) \leq f_{\mathcal{D}_k}(x_k) - c\|s_k\|^2 + C\varepsilon_k,$$

is violated.

This assumption reflects the idea that the local objective functions are not completely homogeneous. In heterogeneous data-fitting problems such behavior is natural: the direction that decreases one local function may be much less favorable, or even unfavorable, for another. Such variability is precisely what makes adaptive sample size control meaningful. In the following theorem, we use the constant $\bar{\varepsilon}$ arising from the summability of the sequence $\{\varepsilon_k\}$, (4.25).

Theorem 6.1.3 *Suppose that Assumptions 2, 3, and 4 hold. Then the expected number of iterations needed to reach $\|d(x_k)\| < \nu$ is bounded above by*

$$\hat{k}_E = \left\lceil \frac{N-1}{q} \right\rceil + \left\lceil \frac{C_b^{FS} - f_{low} + \bar{\varepsilon}}{\bar{c}\nu^2} \right\rceil,$$

where $\bar{c}$ is as in Theorem 6.1.1, C_b^{FS} is as in (6.23), and

$$q = \min\{w_1, \dots, w_N\}^{N-1}.$$

Proof. By Assumption 4, at every iteration there exists at least one local function whose singleton additional sample would violate the acceptance condition. Therefore, by the sampling rule (6.3),

$$\mathbb{P}(\mathcal{D}_k \in \mathcal{D}_k^-) \geq \min\{w_1, \dots, w_N\}^{D_k} \geq \min\{w_1, \dots, w_N\}^{N-1} = q.$$

Let S_k denote the number of sample size increases that occur up to iteration k . Then

$$S_k = I_1 + \dots + I_k,$$

where I_k is the indicator of the event $\{N_k > N_{k-1}\}$. By Step S5 of the algorithm, this event occurs exactly when $\mathcal{D}_k \in \mathcal{D}_k^-$. Hence,

$$\mathbb{E}(I_k) = \mathbb{P}(\mathcal{D}_k \in \mathcal{D}_k^-) \geq q,$$

and therefore

$$\mathbb{E}(S_k) \geq kq. \tag{6.25}$$

Let $\tilde{N}$ denote the total number of sample size increases needed to reach the full sample. Then, the expected iteration count required to enter the FS regime is bounded above by $\lceil \tilde{N}/q \rceil$, and consequently by

$$\left\lceil \frac{N-1}{q} \right\rceil. \quad (6.26)$$

Now let $\tilde{k}_1$ be the first iteration of the FS phase. On FS sample paths, the descent estimate (6.11) holds. Combining it with Assumptions 2, 3, and the summability of $\{\varepsilon_k\}$ gives, for every $j \in \mathbb{N}$,

$$\sum_{k=\tilde{k}_1}^{\tilde{k}_1+j} \mathbb{E}_{FS}(\|d(x_k)\|^2) \leq \frac{C_b^{FS} - f_{low} + \bar{\varepsilon}}{\bar{c}}. \quad (6.27)$$

Let T be the number of FS iterations needed to reach $\|d(x_k)\| < \nu$. Then, before this event occurs, one has $\|d(x_k)\| \geq \nu$, so

$$\sum_{k=\tilde{k}_1}^{\tilde{k}_1+T-1} \mathbb{E}_{FS}(\|d(x_k)\|^2) \geq T\nu^2.$$

Together with (6.27), this yields

$$T \leq \frac{C_b^{FS} - f_{low} + \bar{\varepsilon}}{\bar{c}\nu^2}.$$

Finally, combining this estimate with (6.26) proves the stated bound. $\square$

Remark 6.1.1 *The upper bound $\hat{k}_E$ is generally conservative. First, in the FS phase one may replace $\bar{c}$ by the more precise constant $c_1 t_{min}$. Second, instead of the crude bound $N-1$, one may use the exact number $\tilde{N}$ of increases required to reach the full sample, which reflects*

the actual growth rule used in Step S5. Finally, the parameter q also depends on the size of the additional sample. In particular, if $D_k = 1$ for all k , then

$$q = \min\{w_1, \dots, w_N\},$$

which may be much larger than $\min\{w_1, \dots, w_N\}^{N-1}$. In terms of order, the obtained complexity estimate is $\mathcal{O}(\nu^{-2})$, which matches the deterministic projected gradient rate. If one sets $N_k = N$ for all k , the method becomes deterministic from the start, and the bound reduces to

$$\hat{k} = \left\lceil \frac{f(x_0) - f_{\text{low}} + \bar{\varepsilon}}{\bar{c} \nu^2} \right\rceil.$$

We conclude this part of the analysis with the strongly convex case. Under Assumption A2, strong convexity guarantees that problem (6.1) admits a unique solution, denoted by x^* .

Theorem 6.1.4 *Suppose that Assumption 2 holds and that the sequence $\{x_k\}_{k \in \mathbb{N}}$ generated by AS-NC is bounded. If the function f is strongly convex, then*

$$\lim_{k \rightarrow \infty} x_k = x^* \quad \text{a.s.},$$

where x^* is the unique solution of problem (6.1).

Proof. Boundedness of the iterates implies Assumption 3, so Theorem 6.1.2 applies and shows that every accumulation point of $\{x_k\}$ is almost surely a stationary point of (6.1). On the other hand, strong convexity implies that problem (6.1) has a unique stationary point, namely the unique minimizer x^* . Therefore every accumulation point of the sequence must coincide with x^* almost surely. Since a bounded sequence all of whose accumulation points are equal must converge to that point, we conclude that $x_k \rightarrow x^*$ a.s. This completes the proof. $\square$

6.2 Box Constraints

We now extend the additional-sampling framework developed for non-negativity constraints to the general box-constrained case. The transition from AS-NC to AS-BOX is conceptually natural: the same overall architecture of the method is retained, namely projected sampled directions, nonmonotone line search, and an additional-sampling acceptance test, while only those components that depend explicitly on the geometry of the feasible set need to be modified. In particular, the projection operator is no longer binary in nature, as in the nonnegativity case, but instead distinguishes between lower-bound activity, interior components, and upper-bound activity. As a consequence, the indicator used to compare the projection structures of the main and additional samples must also be adapted. Apart from these changes, both the algorithmic design and the convergence arguments remain strongly aligned with those of AS-NC.

6.2.1 Problem formulation

In this subsection, we consider the general box-constrained optimization problem (4.4). Compared with the nonnegativity-constrained case, the passage to box constraints requires only the corresponding modification of the projection operator and of the comparison between the sampled projected-gradient directions. After these changes are introduced, the convergence analysis follows the same general arguments as in the previous case.

Consider the feasible set

$$S = \{x \in \mathbb{R}^n \mid l_i \leq x_i \leq u_i, \ i = 1, \dots, n\}.$$

In this setting, the projected direction defined in (6.4) takes the com-

ponentwise form

$$[p_k]_i = \begin{cases} l_i - [x_k]_i, & [x_k]_i - [\nabla f_{\mathcal{N}_k}(x_k)]_i < l_i, \\ -[\nabla f_{\mathcal{N}_k}(x_k)]_i, & l_i \leq [x_k]_i - [\nabla f_{\mathcal{N}_k}(x_k)]_i \leq u_i, \\ u_i - [x_k]_i, & [x_k]_i - [\nabla f_{\mathcal{N}_k}(x_k)]_i > u_i, \end{cases} \quad i = 1, \dots, n. \quad (6.28)$$

Thus, depending on the relative position of the trial point $x_k - \nabla f_{\mathcal{N}_k}(x_k)$ with respect to the interval $[l_i, u_i]$, the projected step either moves the i -th component to the lower bound, remains in the interior, or moves it to the upper bound.

Compared to the nonnegativity-constrained case, where each component had only two possible structural states, the box-constrained setting leads to a three-way classification. This distinction is reflected in the indicator vector associated with the main sample. Namely, analogously to (6.5), we define

$$[\tilde{I}_{\mathcal{N}_k}]_i = \begin{cases} 1, & \text{if } [x_k]_i - [\nabla f_{\mathcal{N}_k}(x_k)]_i < l_i, \\ 2, & \text{if } l_i \leq [x_k]_i - [\nabla f_{\mathcal{N}_k}(x_k)]_i \leq u_i, \\ 3, & \text{if } [x_k]_i - [\nabla f_{\mathcal{N}_k}(x_k)]_i > u_i, \end{cases} \quad i = 1, \dots, n.$$

The indicator vector $\tilde{I}_{\mathcal{D}_k}$ associated with the additional sample is introduced in the same way. The discrepancy between the two projection structures is then measured by

$$\tilde{r}_{\mathcal{D}_k} = \|\tilde{I}_{\mathcal{N}_k} - \tilde{I}_{\mathcal{D}_k}\|. \quad (6.29)$$

The role of $\tilde{r}_{\mathcal{D}_k}$ is entirely analogous to that of $r_{\mathcal{D}_k}$ in the nonnegativity-constrained case, but now the comparison is finer. The value $\tilde{r}_{\mathcal{D}_k} = 0$ means that the main and additional samples induce exactly the same partition of the coordinates into three groups: those that hit the lower bound, those that remain in the interior of the box,

and those that hit the upper bound. Hence, in the box-constrained setting, the structural compatibility test must track not only whether a component becomes active, but also which bound becomes active. This is precisely why the binary indicator from AS-NC is replaced here by a ternary one.

6.2.2 AS-BOX algorithm

The algorithm proposed for the general box-constrained case differs from AS-NC only in those parts that explicitly involve the projection structure. More precisely, the only modifications concern Step S2, where the projected search direction is computed, and Step S4, where the structural discrepancy between the main and additional samples is evaluated. The remaining components of the method – the initialization, the nonmonotone line search, the additional-sampling acceptance test, the sample size update, and the iterate update – remain unchanged. For completeness, the full algorithm is stated below.

The interpretation of the algorithm is essentially the same as in the nonnegativity-constrained case. The main sample $\mathcal{N}_k$ is used to construct a projected descent direction and to generate a candidate point through a nonmonotone line search. The additional sample $\mathcal{D}_k$ is then used to assess whether this candidate is satisfactory from the viewpoint of fresh independent information and whether it induces the same boundary/interior structure as the main sample. If the additional test succeeds, the method keeps the current sample size. If not, the candidate point may be rejected and the sample size is increased.

The passage from AS-NC to AS-BOX therefore affects only the form of the projected direction and the corresponding discrepancy measure $\tilde{r}_{\mathcal{D}_k}$. The remaining algorithmic structure and the convergence argument remain the same as in the nonnegativity-constrained case.

Algorithm 4 AS-BOX: Additional Sampling for BOX Constraints

S0 *Initialization.* Input: $x_0 \in S$, $N_0 \in \mathbb{N}$, $\beta, c, c_1 \in (0, 1)$, $C > 0$, $\{\varepsilon_k\}$ satisfying (4.25). Set $k := 0$.

S1 *Subsampling.* If $N_k < N$, choose N_k via (6.2). Otherwise, set $f_{N_k} = f$.

S2 *Search direction.* Compute p_k via (6.28).

S3 *Step size.* Find the smallest $j \in \mathbb{N}_0$ such that $t_k = \beta^j$ satisfies

$$f_{N_k}(x_k + t_k p_k) \leq f_{N_k}(x_k) + c_1 t_k (\nabla f_{N_k}(x_k))^T p_k + \varepsilon_k. \quad (6.30)$$

Set $\bar{x}_k = x_k + t_k p_k$.

S4 *Additional sampling.* If $N_k = N$, set $x_{k+1} = \bar{x}_k$, $k := k + 1$, and go to Step **S1**.

Otherwise, choose $\mathcal{D}_k$ via (6.3) and compute

$$s_k = \pi_S(x_k - \nabla f_{\mathcal{D}_k}(x_k)) - x_k \quad (6.31)$$

and $\tilde{r}_{\mathcal{D}_k}$ defined by (6.29).

S5 *Sample size update.* If

$$\tilde{r}_{\mathcal{D}_k} = 0 \quad \text{and} \quad f_{\mathcal{D}_k}(\bar{x}_k) \leq f_{\mathcal{D}_k}(x_k) - c \|s_k\|^2 + C\varepsilon_k, \quad (6.32)$$

set $N_{k+1} = N_k$.

Otherwise, choose $N_{k+1} \in \{N_k + 1, \dots, N\}$.

S6 *Iterate update.* If

$$f_{\mathcal{D}_k}(\bar{x}_k) \leq f_{\mathcal{D}_k}(x_k) - c \|s_k\|^2 + C\varepsilon_k,$$

set $x_{k+1} = \bar{x}_k$.

Otherwise, set $x_{k+1} = x_k$.

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

6.2.3 Convergence analysis

The convergence analysis is performed under the same assumptions as in the nonnegativity-constrained case. Moreover, the analogues of Lemmas 6.1.1 and 6.1.2 remain valid in the present setting. More precisely, if one defines the sets $\mathcal{D}_k^+$ and $\mathcal{D}_k^-$ by replacing the condition (6.9) with (6.32), and $r_{\mathcal{D}_k}$ with $\tilde{r}_{\mathcal{D}_k}$, then the proof of eventual emptiness of $\mathcal{D}_k^-$ in the MB regime is unchanged. Likewise, the lower bound on the accepted step size follows from the same backtracking argument, since the projected direction still satisfies the descent property from Theorem 6.1.1.

The only result whose proof needs to be genuinely adapted is the analogue of Theorem 6.1.1, because in the box-constrained setting the projected direction has three different coordinate types instead of two. The next theorem provides the corresponding full-objective descent estimate.

Theorem 6.2.1 *Suppose that Assumption 2 holds. Then a.s. there exists a finite random iteration $\tilde{k}$ such that for all $k \geq \tilde{k}$,*

$$f(x_{k+1}) \leq f(x_k) - \bar{c}\|d(x_k)\|^2 + \bar{C}\varepsilon_k,$$

where

$$\bar{c} = \min \left\{ c, \ c_1, \ \frac{2c_1(1-c_1)\beta}{L} \right\}, \quad \bar{C} = \max\{1, C\}.$$

Proof. The argument begins exactly as in the proof of Theorem 6.1.1. In the FS regime, there exists a finite random index $\tilde{k}_1$ such that for all $k \geq \tilde{k}_1$ the full sample is used, and therefore the line search inequality yields

$$f(x_{k+1}) \leq f(x_k) - c_1 t_{\min} \|d(x_k)\|^2 + \varepsilon_k. \quad (6.33)$$

We now consider the MB regime. Since the analogue of Lemma 6.1.1 also holds for AS-BOX, there exists a finite random index k_1

such that $\mathcal{D}_k^- = \emptyset$ for all $k \geq k_1$. Consequently, for all sufficiently large k the acceptance test is satisfied for all admissible realizations of the additional sample. As in the nonnegativity-constrained case, this implies that the candidate point is accepted and that one has, for all $k \geq k_1$,

$$f(x_{k+1}) \leq f(x_k) - c \sum_{j=1}^N w_j \|s_k^j\|^2 + C\varepsilon_k, \quad (6.34)$$

where

$$s_k^j := \pi_S(x_k - \nabla f_j(x_k)) - x_k.$$

The remaining task is to relate the exact projected gradient mapping

$$d(x_k) = \pi_S(x_k - \nabla f(x_k)) - x_k$$

to the family of local projected directions $\{s_k^j\}_{j=1}^N$. To this end, we partition the coordinate set according to the position of the projected trial point relative to the interval bounds. Define

$$\begin{aligned} \mathcal{L}_{N_k} &:= \{i \in \{1, \dots, n\} \mid [x_k]_i - [\nabla f_{N_k}(x_k)]_i < l_i\}, \\ \mathcal{J}_{N_k} &:= \{i \in \{1, \dots, n\} \mid l_i \leq [x_k]_i - [\nabla f_{N_k}(x_k)]_i \leq u_i\}, \\ \mathcal{U}_{N_k} &:= \{i \in \{1, \dots, n\} \mid [x_k]_i - [\nabla f_{N_k}(x_k)]_i > u_i\}. \end{aligned} \quad (6.35)$$

Thus, the set $\mathcal{L}_{N_k}$ contains the indices whose projected trial point hits the lower bound, $\mathcal{J}_{N_k}$ contains those that remain in the interior, and $\mathcal{U}_{N_k}$ contains those that hit the upper bound.

Exactly as in the nonnegativity-constrained case, the fact that $\mathcal{D}_k^- = \emptyset$ implies that the structural discrepancy indicator vanishes for all singleton additional samples. Hence, for all $k \geq k_1$,

$$\mathcal{L}_{N_k} = \mathcal{L}_k^1 = \dots = \mathcal{L}_k^N, \mathcal{J}_{N_k} = \mathcal{J}_k^1 = \dots = \mathcal{J}_k^N, \mathcal{U}_{N_k} = \mathcal{U}_k^1 = \dots = \mathcal{U}_k^N, \quad (6.36)$$

where, for each $j = 1, \dots, N$,

$$\begin{aligned}\mathcal{L}_k^j &:= \{i \in \{1, \dots, n\} \mid [x_k]_i - [\nabla f_j(x_k)]_i < l_i\}, \\ \mathcal{J}_k^j &:= \{i \in \{1, \dots, n\} \mid l_i \leq [x_k]_i - [\nabla f_j(x_k)]_i \leq u_i\}, \\ \mathcal{U}_k^j &:= \{i \in \{1, \dots, n\} \mid [x_k]_i - [\nabla f_j(x_k)]_i > u_i\}.\end{aligned}$$

This equality of sets means that, once the additional-sampling test stops failing, all local projected directions exhibit the same lower-bound/interior/upper-bound structure. We can therefore infer the position of the full projected direction coordinatewise. Indeed, for any $i \in \mathcal{L}_{N_k}$ and every $j \in \mathcal{N}$, one has

$$[x_k]_i - [\nabla f_j(x_k)]_i < l_i,$$

which implies

$$[x_k]_i < [\nabla f_j(x_k)]_i + l_i.$$

Multiplying by w_j and summing over j yields

$$[x_k]_i < [\nabla f(x_k)]_i + l_i, \quad i \in \mathcal{L}_{N_k}.$$

Similarly, for $i \in \mathcal{U}_{N_k}$ we obtain

$$[x_k]_i - [\nabla f(x_k)]_i > u_i,$$

while for $i \in \mathcal{J}_{N_k}$ we have

$$l_i \leq [x_k]_i - [\nabla f(x_k)]_i \leq u_i.$$

Consequently, for all $k \geq k_1$,

$$[d(x_k)]_i = \begin{cases} l_i - [x_k]_i, & i \in \mathcal{L}_{N_k}, \\ -[\nabla f(x_k)]_i, & i \in \mathcal{J}_{N_k}, \\ u_i - [x_k]_i, & i \in \mathcal{U}_{N_k}. \end{cases} \quad (6.37)$$

We now estimate the squared norm of $d(x_k)$:

$$\begin{aligned} \|d(x_k)\|^2 &= \sum_{i=1}^n ([d(x_k)]_i)^2 \\ &= \sum_{i \in \mathcal{L}_{N_k}} (l_i - [x_k]_i)^2 + \sum_{i \in \mathcal{J}_{N_k}} ([\nabla f(x_k)]_i)^2 \end{aligned} \quad (6.38)$$

$$+ \sum_{i \in \mathcal{U}_{N_k}} (u_i - [x_k]_i)^2. \quad (6.39)$$

On the other hand, by the definition of the local projected directions and (6.36), one has for all $k \geq k_1$ and all $j \in \mathcal{N}$,

$$[s_k^j]_i = \begin{cases} l_i - [x_k]_i, & i \in \mathcal{L}_{N_k}, \\ -[\nabla f_j(x_k)]_i, & i \in \mathcal{J}_{N_k}, \\ u_i - [x_k]_i, & i \in \mathcal{U}_{N_k}. \end{cases}$$

Hence, for every $i \in \mathcal{L}_{N_k}$,

$$\sum_{j=1}^N w_j ([s_k^j]_i)^2 = (l_i - [x_k]_i)^2,$$

and therefore

$$\sum_{i \in \mathcal{L}_{N_k}} \sum_{j=1}^N w_j ([s_k^j]_i)^2 = \sum_{i \in \mathcal{L}_{N_k}} (l_i - [x_k]_i)^2. \quad (6.40)$$

Likewise, for every $i \in \mathcal{U}_{N_k}$,

$$\sum_{j=1}^N w_j ([s_k^j]_i)^2 = (u_i - [x_k]_i)^2,$$

and hence

$$\sum_{i \in \mathcal{U}_{N_k}} \sum_{j=1}^N w_j ([s_k^j]_i)^2 = \sum_{i \in \mathcal{U}_{N_k}} (u_i - [x_k]_i)^2. \quad (6.41)$$

For indices in the interior set, convexity of the square function gives

$$([\nabla f(x_k)]_i)^2 = \left(\sum_{j=1}^N w_j [\nabla f_j(x_k)]_i \right)^2 \leq \sum_{j=1}^N w_j ([\nabla f_j(x_k)]_i)^2 = \sum_{j=1}^N w_j ([s_k^j]_i)^2,$$

and summing over $i \in \mathcal{J}_{N_k}$ yields

$$\sum_{i \in \mathcal{J}_{N_k}} ([\nabla f(x_k)]_i)^2 \leq \sum_{i \in \mathcal{J}_{N_k}} \sum_{j=1}^N w_j ([s_k^j]_i)^2. \quad (6.42)$$

Combining (6.38), (6.40), (6.41), and (6.42), we obtain

$$\begin{aligned} \|d(x_k)\|^2 &\leq \sum_{i \in \mathcal{L}_{N_k}} \sum_{j=1}^N w_j ([s_k^j]_i)^2 + \sum_{i \in \mathcal{J}_{N_k}} \sum_{j=1}^N w_j ([s_k^j]_i)^2 + \sum_{i \in \mathcal{U}_{N_k}} \sum_{j=1}^N w_j ([s_k^j]_i)^2 \\ &= \sum_{j=1}^N w_j \|s_k^j\|^2. \end{aligned} \quad (6.43)$$

Substituting this bound into (6.34), we conclude that for all $k \geq k_1$,

$$f(x_{k+1}) \leq f(x_k) - c \|d(x_k)\|^2 + C \varepsilon_k. \quad (6.44)$$

Thus, in the FS regime we have (6.33), and in the MB regime we have (6.44). Choosing k sufficiently large so that the corresponding estimate holds in the realized scenario, we obtain the desired conclusion.

□

We next record the almost sure convergence, complexity, and strong-convexity results for AS-BOX. Although these statements are fully analogous to those obtained for AS-NC, we provide explicit proofs here for completeness.

Theorem 6.2.2 *Suppose that Assumptions 2 and 3 hold. Then a.s. every accumulation point of the sequence $\{x_k\}_{k \in \mathbb{N}}$ generated by AS-BOX is a stationary point of problem (4.5).*

Proof. The argument is identical in structure to the proof of Theorem 6.1.2, with the only difference being that the feasible set is now the general box S . By Theorem 6.2.1, there exists a finite random index k such that for all $l \in \mathbb{N}$,

$$f(x_{\tilde{k}+l}) \leq f(x_{\tilde{k}}) - \bar{c} \sum_{j=0}^{l-1} \|d(x_{\tilde{k}+j})\|^2 + \bar{C} \sum_{j=0}^{l-1} \varepsilon_{\tilde{k}+j}.$$

Taking expectations, using Assumption 3, and letting $l \rightarrow \infty$, we conclude that

$$\sum_{j=0}^{\infty} \mathbb{E}(\|d(x_{\tilde{k}+j})\|^2) < \infty.$$

Hence, by the extended Markov inequality and the Borel–Cantelli lemma,

$$\mathbb{P}\left(\lim_{k \rightarrow \infty} d(x_k) = 0\right) = 1.$$

Let x^* be an arbitrary accumulation point of the sequence $\{x_k\}$, and let $K_0 \subset \mathbb{N}$ be an infinite subsequence such that $x_k \rightarrow x^*$ as $k \in K_0$. Since all iterates are feasible, one has $x^* \in S$. Passing to the limit in the identity

$$d(x_k) = \pi_S(x_k - \nabla f(x_k)) - x_k,$$

and using continuity of both ∇f and the projection operator, we obtain

$$0 = \lim_{k \in K_0} d(x_k) = \pi_S(x^* - \nabla f(x^*)) - x^* = d(x^*).$$

By Proposition 1.1.1 b), the equality $d(x^*) = 0$ is equivalent to stationarity of x^* for problem (4.5). Therefore every accumulation point is almost surely stationary. $\square$

To derive the complexity estimate, we use the same heterogeneity assumption as in the nonnegativity-constrained case, now interpreted through the box-constrained acceptance condition (6.32). More precisely, Assumption 4 is understood as requiring that at every iteration there exists at least one local function for which the box-constrained additional-sampling test fails.

Theorem 6.2.3 *Suppose that Assumptions 2, 3, and 4 hold. Then the expected number of iterations required to reach $\|d(x_k)\| < \nu$ is upper bounded by*

$$\hat{k}_E = \left\lceil \frac{N-1}{q} \right\rceil + \left\lceil \frac{C_b^{FS} - f_{low} + \bar{\varepsilon}}{\bar{c} \nu^2} \right\rceil,$$

where $\bar{c}$ is the constant from Theorem 6.2.1, C_b^{FS} is defined through (6.23), and

$$q = \min\{w_1, \dots, w_N\}^{N-1}.$$

Proof. The reasoning is the same as in the proof of Theorem 6.1.3. Assumption 4 guarantees that for every iteration k there exists at least one local function for which the acceptance condition (6.32) fails. Therefore, by the sampling rule for the additional sample,

$$\mathbb{P}(\mathcal{D}_k \in \mathcal{D}_k^-) \geq \min\{w_1, \dots, w_N\}^{D_k} \geq \min\{w_1, \dots, w_N\}^{N-1} = q.$$

Let S_k denote the number of sample size increases up to iteration k , and let I_k be the indicator of the event $\{N_k > N_{k-1}\}$. Then

$$S_k = I_1 + \dots + I_k, \quad \mathbb{E}(I_k) = \mathbb{P}(\mathcal{D}_k \in \mathcal{D}_k^-) \geq q.$$

Consequently,

$$\mathbb{E}(S_k) \geq kq.$$

If $\tilde{N}$ denotes the number of increases needed to reach the full sample, then the expected number of iterations necessary to enter the FS regime is bounded above by $\lceil \tilde{N}/q \rceil$, and hence by

$$\left\lceil \frac{N-1}{q} \right\rceil.$$

Once the algorithm enters the FS regime, Theorem 6.2.1 yields the deterministic-type descent estimate for the full objective. Arguing exactly as in the proof of Theorem 6.1.3, one obtains

$$\sum_{k=\tilde{k}_1}^{\tilde{k}_1+j} \mathbb{E}_{FS}(\|d(x_k)\|^2) \leq \frac{C_b^{FS} - f_{low} + \bar{\varepsilon}}{\bar{c}} \quad \text{for all } j \in \mathbb{N}.$$

If T denotes the number of FS iterations needed to reach $\|d(x_k)\| < \nu$, then prior to this event we must have $\|d(x_k)\| \geq \nu$, so

$$T\nu^2 \leq \sum_{k=\tilde{k}_1}^{\tilde{k}_1+T-1} \mathbb{E}_{FS}(\|d(x_k)\|^2) \leq \frac{C_b^{FS} - f_{low} + \bar{\varepsilon}}{\bar{c}}.$$

Thus,

$$T \leq \frac{C_b^{FS} - f_{low} + \bar{\varepsilon}}{\bar{c}\nu^2}.$$

Adding the expected time needed to enter the FS phase proves the stated bound. $\square$

Theorem 6.2.4 *Suppose that Assumption 2 holds and that the sequence $\{x_k\}_{k \in \mathbb{N}}$ generated by AS-BOX is bounded. If the function f is strongly convex, then*

$$\lim_{k \rightarrow \infty} x_k = x^* \quad a.s.,$$

where x^* is the unique solution of problem (4.5).

Proof. Boundedness of the iterates implies Assumption 3, so Theorem 6.2.2 shows that every accumulation point of $\{x_k\}$ is almost surely a stationary point of the box-constrained problem. On the other hand, strong convexity of f implies that problem (4.5) admits a unique stationary point, which necessarily coincides with the unique minimizer x^* . Therefore all accumulation points of the generated sequence must be equal to x^* almost surely. Since a bounded sequence whose accumulation points all coincide with the same vector must converge to that vector, we conclude that

$$x_k \rightarrow x^* \quad \text{a.s.}$$

This proves the theorem. □

The AS-BOX framework may therefore be regarded as a genuine extension of AS-NC rather than as a fundamentally different method. The algorithmic deviations are minimal and stem exclusively from the more general form of the projection operator and the corresponding three-way active-set structure. The same is true at the theoretical level: once the projection geometry is carefully incorporated into the estimates, the almost sure convergence result, the expected complexity bound, and the strong-convexity convergence statement all follow in the same spirit as in the nonnegativity-constrained case.

Chapter 7

Numerical Experiments

This chapter is devoted to the numerical assessment of the methods developed in the previous chapters 5 and 6. The exposition is organized according to the two principal classes of constrained finite sum problems considered in the dissertation. We begin with equality-constrained problems and the ASPEN framework, where the experiments focus on computational efficiency, comparison with benchmark methods, and the influence of data heterogeneity on sample size adaptation. After that, we turn to box-constrained problems and the ASBOX methodology, where the same general philosophy is tested in the projection-based setting. In this way, the chapter provides both a problem-dependent and a method-dependent view of the practical behavior of the proposed algorithms.

7.1 Equality-Constrained Problems: ASPEN

In this section, we present the numerical results for the equality-constrained framework. The experiments are organized progressively:

we first describe the computational environment and benchmark problems, then compare ASPEN with natural baseline strategies, after that with stochastic SQP-type competitors, and finally examine how the adaptive sample size mechanism reacts to controlled changes in the level of heterogeneity among local objective functions.

7.1.1 Computational setting and benchmark suite

In this subsection, we examine the practical performance of ASPEN on a collection of equality-constrained machine learning problems. Our goal is twofold. First, we investigate whether the proposed adaptive sampling strategy can provide a computational advantage over natural deterministic and heuristic alternatives. Second, we study how the algorithm behaves across datasets of different sizes and structures, as well as how its sample size mechanism reacts to different levels of heterogeneity in the local objective functions.

The experimental study is conducted on several binary classification datasets from the LIBSVM collection [14], listed in Table 7.1. In order to complement these real-data experiments with a more controlled synthetic setting, we also consider the academic problem HS24 from the CUTEst collection [27], enriched by artificial noise. This additional experiment is used to illustrate the sensitivity of ASPEN to different levels of dissimilarity among local cost functions. Some of these datasets originate from standard machine-learning repositories, such as the UCI Machine Learning Repository ([52, 53]).

We begin by comparing ASPEN with two natural reference strategies. The first one is a deterministic penalty method, denoted by *Full*, in which the complete sample is used at every iteration, that is, $N_k = N$ for all $k \in \mathbb{N}$. In this case, each penalized subproblem is solved approximately until the condition

$$\|\nabla F_{N_k}(x_k, \mu_k)\| < \frac{1}{\mu_k}$$

is met, after which the penalty parameter is updated according to

$$\mu_{k+1} = \gamma\mu_k.$$

The second baseline is a heuristic adaptive strategy, denoted by *Heur*. This method starts from the same initial sample size as ASPEN and increases it according to the rule

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

whenever

$$\|\nabla F_{N_k}(x_k, \mu_k)\| < \frac{1}{\mu_k}.$$

The penalty parameter is updated in the same way as in the Full method, and the same nonmonotone line search is used for all three methods.

The parameter values employed in ASPEN are chosen in accordance with the algorithmic framework introduced earlier. The initial penalty parameter is set to $\mu_0 = 1$, while the initial sample size is chosen as $N_0 = \lceil 0.01N \rceil$. The additional sampling step is performed with $D_k = 1$, $c = 10^{-4}$, and $C = 1$. In the line search procedure we use $\beta = 0.1$, $\eta = 10^{-4}$, and the error sequence $\epsilon_k = k^{-1.1}$. Whenever the algorithm decides to enlarge the main sample, the increase is taken to be minimal, that is,

$$N_{k+1} = N_k + 1,$$

as prescribed in Line 23 of ASPEN. The same starting values of μ_0 and N_0 are used for the heuristic method, and throughout all experiments we fix $\gamma = 1.1$.

Initial points are identical for all compared methods and are obtained by normalizing Gaussian random vectors.

Table 7.1: Binary classification data set details [14].

Dataset	Dimension (n)	Datapoints (N)
a9a	123	32,561
Australian	14	690
Heart	13	270
Mushrooms	112	8,124
Splice	60	1,000
MNIST	784	60000

As the underlying test model, we consider the equality-constrained logistic regression problem

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{N} \sum_{i=1}^N \log \left(1 + e^{-b_i a_i^\top x} \right) \quad \text{s.t.} \quad \|x\|_2^2 = 1, \quad (7.1)$$

where $a_i \in \mathbb{R}^n$ denotes the feature vector associated with the i -th data point, and $b_i \in \{-1, 1\}$ is the corresponding binary label. This formulation is a standard benchmark in constrained machine learning and is particularly suitable for the present context, since it combines a finite sum structure with a nonlinear equality constraint.

The computational effort is measured by the *Function Evaluation Value* (FEV_k), which counts the total number of scalar products required by a given method in order to compute the iterate x_k , starting from the initial point x_0 .

This performance metric is more informative than a pure iteration count, because the compared methods operate with different sample sizes and therefore incur different per-iteration costs. Measuring progress against FEV_k makes it possible to compare the methods on a common computational scale.

In the first group of experiments, for each dataset from Table 7.1, we report the following quantities:

1. the distance to the reference solution, $\|x_k - x^*\|$, plotted against the computational budget FEV_k (part a);
2. the evolution of the sample size across iterations (part b);
3. the evolution of the penalty parameter across iterations (part c).

The corresponding results are displayed in Figures 7.1–7.6.

This first group of experiments establishes the common computational framework used in the remainder of the ASPEN study. Having fixed the model problem, the datasets, and all algorithmic parameters, we can now examine in a direct way how ASPEN compares with simpler penalty-based reference strategies before moving on to more specialized stochastic constrained optimization methods.

7.1.2 Comparative Performance of ASPEN

The first set of experiments highlights the advantage of ASPEN over the deterministic Full strategy and over the heuristic adaptive rule *Heur*. Across most datasets, ASPEN reaches a smaller distance to the reference solution while requiring substantially fewer scalar products. This behavior indicates that the additional-sampling mechanism is able to balance progress and cost more effectively than either always using the full sample or enlarging the sample according to a fixed heuristic rule.

A particularly important aspect of these experiments is the evolution of the sample size. The plots in part b) show that ASPEN does not increase the sample in a uniform or preassigned way. Instead, the growth of N_k depends strongly on the dataset under consideration, which confirms the adaptive character of the method. For example, the MNIST experiment in Figure 7.1 exhibits a noticeably faster increase of the sample size, which is consistent with the greater diversity and complexity of that dataset. In contrast, the Mushrooms

dataset in Figure 7.2 allows the method to remain effective with much smaller sample sizes, suggesting that the local objective functions are significantly more aligned and that a modest mini-batch is already sufficiently informative.

Another striking feature of these experiments is that none of the considered runs reaches the full sample. This indicates that, on the tested problems, ASPEN is able to produce competitive and often superior solutions while remaining entirely in the subsampled regime. The plots in part c) further show that the penalty parameter in ASPEN tends to increase more aggressively than in Full and Heur. Rather than being detrimental, this appears to be beneficial: the method enforces feasibility more actively while still maintaining a favorable balance between progress and cost.

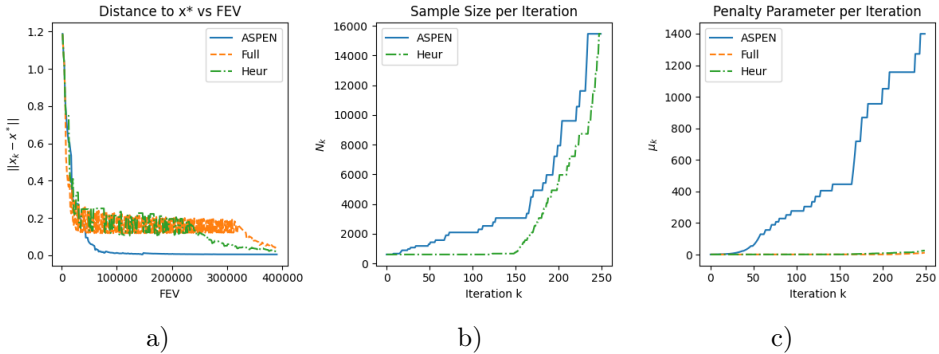


Figure 7.1: *MNIST* dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).

Figure 7.7 reports the relative sample size N_k/N rather than the absolute sample size N_k . This normalization is particularly useful in the present setting because the data sets listed in Table 7.1 differ substantially in the number of data points, ranging from 270 to 60 000.

The plots show that the adaptive methods exhibit noticeably dif-

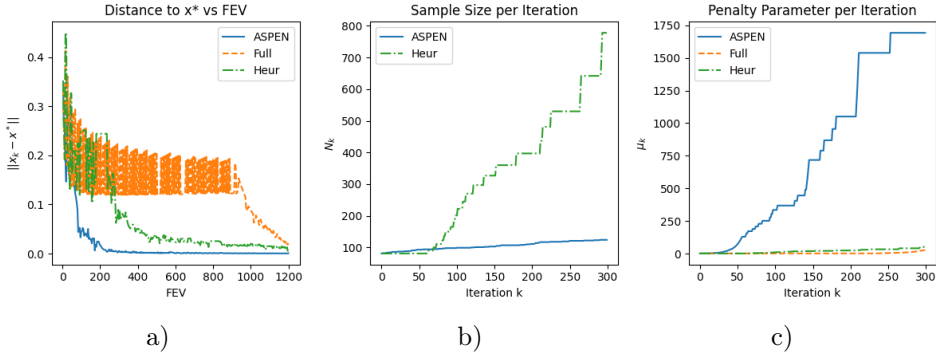


Figure 7.2: *Mushrooms* dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).

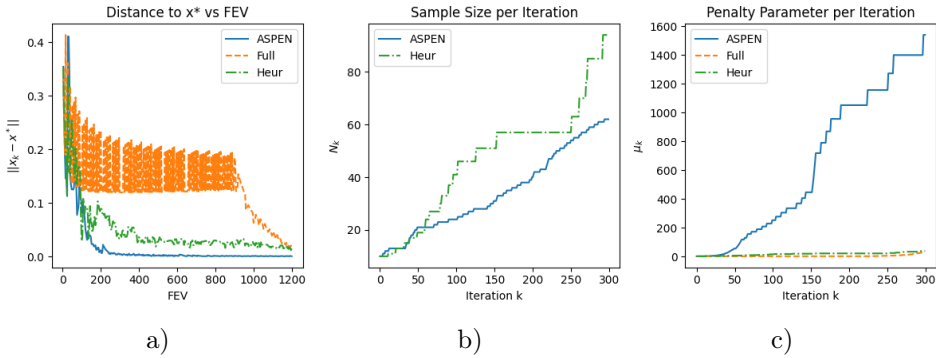


Figure 7.3: *Splice* dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).

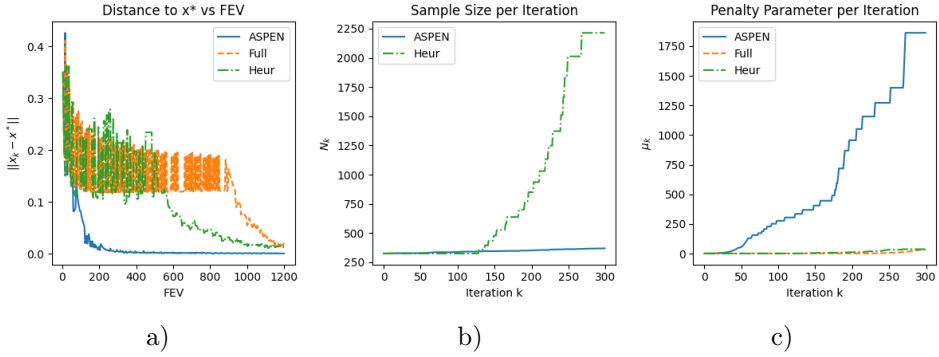


Figure 7.4: *a9a* dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).

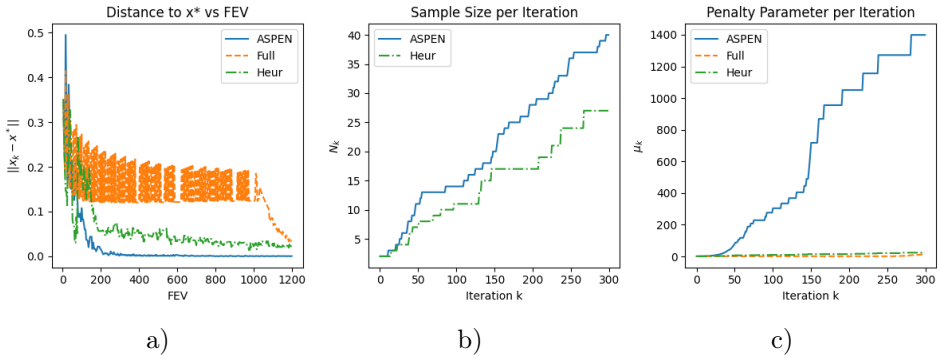


Figure 7.5: *Heart* dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).

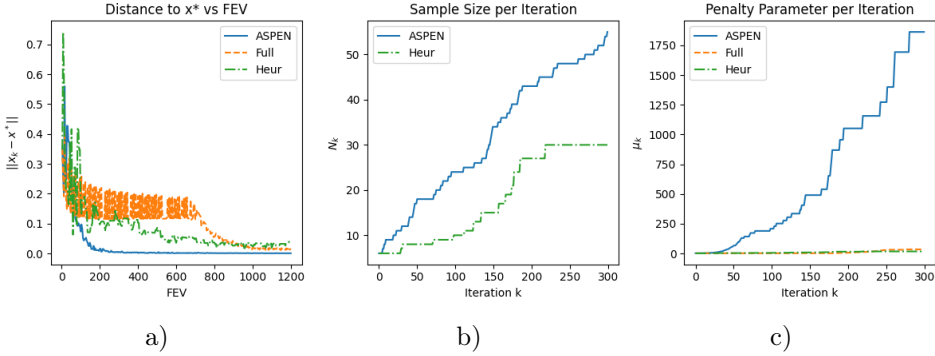


Figure 7.6: *Australian* dataset. ASPEN vs. Full and Heur: optimality gap vs. FEV (a); sample size vs. iteration (b); penalty parameter vs. iteration (c).

ferent sample-growth behaviors across the considered problems. In particular, on the *Australian* and *Splice* data sets, the Heur strategy increases the sample size more aggressively and reaches a considerably larger fraction of the data than ASPEN. On *MNIST*, both adaptive approaches remain well below the full sample regime throughout the considered iteration budget, which indicates substantial computational savings on the largest data set. On *Heart* and *a9a*, ASPEN displays a pronounced late increase and eventually reaches the full sample, suggesting that more accurate gradient information becomes desirable in the final stage of the optimization process, whereas Heur remains below 1. On *Mushrooms*, Heur uses a larger fraction of the data during the middle phase, while ASPEN catches up later and eventually overtakes it near the end. Overall, these results indicate that ASPEN is generally more conservative in the early and middle iterations and tends to enlarge the sample only when the progress of the method requires more accurate information, whereas Heur often follows a smoother and earlier growth pattern.

From a broader perspective, these experiments support the cen-

tral idea behind ASPEN: additional sampling is not used merely as a safeguard, but as a mechanism that shapes the whole optimization process. It determines when the current sample is sufficiently informative and when more precision is required, thereby producing a genuinely data-dependent sample size schedule.

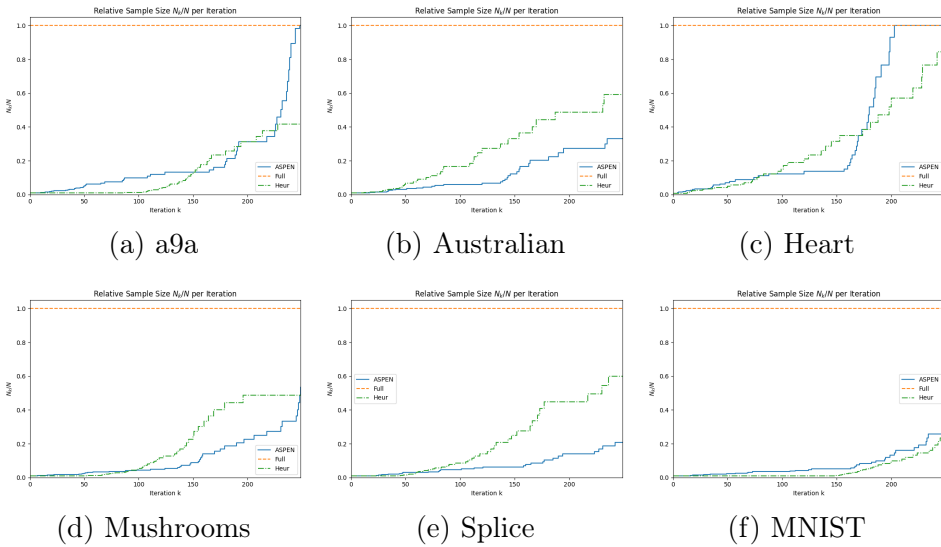


Figure 7.7: Relative sample size N_k/N per iteration for ASPEN, Heur, and Full on the considered binary classification data sets.

The comparison with Full and Heur shows how ASPEN behaves relative to two conceptually simple reference rules that share the same penalty framework. This provides a useful baseline, but it does not yet position ASPEN with respect to stochastic methods specifically designed for constrained optimization. For that reason, the next subsection broadens the comparison and considers SQP-type competitors that represent a more demanding benchmark.

7.1.3 Comparison with Stochastic SQP-Type Methods

We next compare ASPEN with two established stochastic optimization methods from the equality-constrained literature, namely Sto-SQP [4] and SVR-STO [5]. In these experiments, ASPEN is used with exactly the same parameter values as above, while the competing methods are run with parameter choices obtained through empirical tuning and consistent with the recommendations given in the corresponding references. After preliminary validation, the selected values were fixed and used uniformly across all datasets, following the experimental protocol adopted in [5]. More specifically, Sto-SQP is run with

$$\theta = 10^4, \tilde{\tau}_{-1} = 0.1, \epsilon_\tau = 10^{-6}, \tilde{\xi}_{-1} = 0.1, \epsilon_\xi = 10^{-2}, \sigma_{\text{Sto-SQP}} = 0.5,$$

while SVR-STO uses

$$\begin{aligned} \sigma_{\text{SVR-STO}} &= 0.5, \theta = 10^4, \tilde{\tau}_{-1,0} = 0.1, \epsilon_\tau = 10^{-6}, \alpha_u = 10^6, \\ \beta_{\text{SVR-STO}} &= 1. \end{aligned}$$

The comparison presented in Figure 7.8 indicates that ASPEN is competitive with these state-of-the-art schemes and, on most datasets, provides the most favorable trade-off between solution quality and computational effort. Since the behavior is not identical across all problems, we comment on each dataset separately.

Figure 7.8 a) corresponds to the *Mushrooms* dataset. In this experiment, ASPEN exhibits a very rapid initial decrease of the optimality gap and continues to improve throughout the available computational window. Sto-SQP also decreases in a stable and monotone manner, but its rate is visibly slower, meaning that it requires more scalar products to achieve a comparable level of accuracy. The SVR-SQP variant with batch size $b = 16$ improves more slowly than Sto-SQP, while the

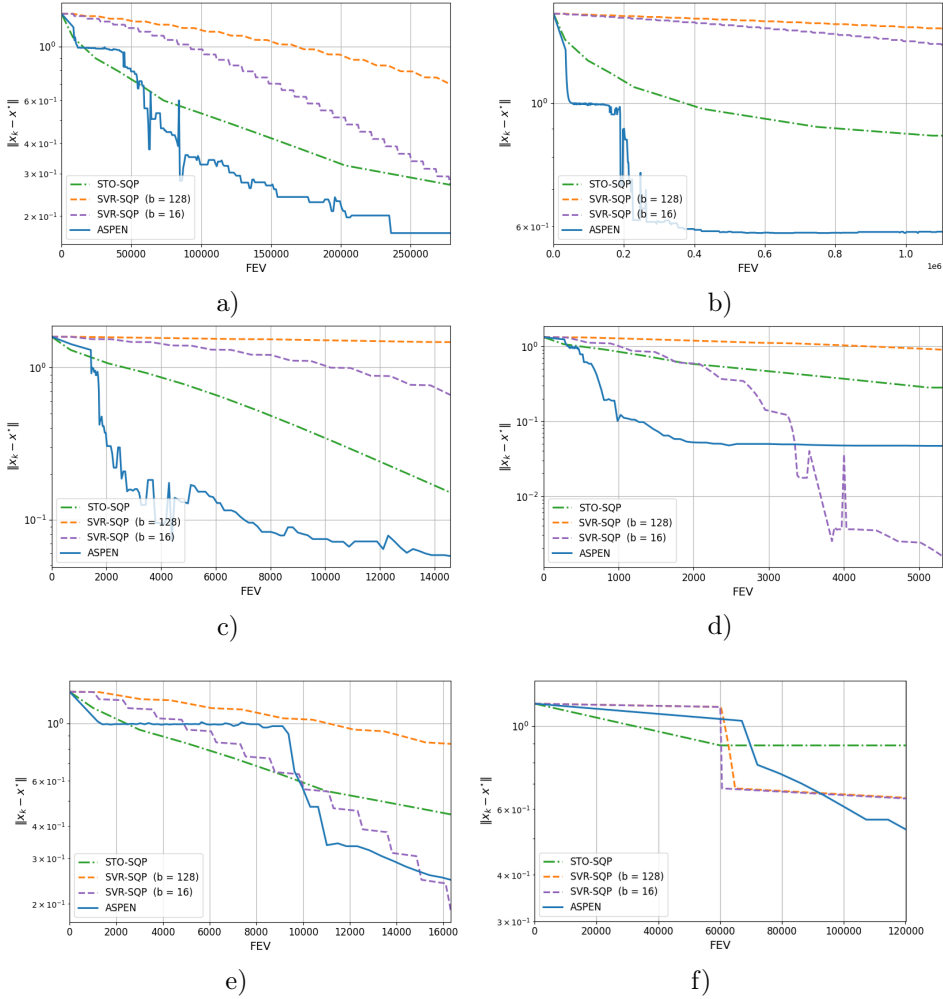


Figure 7.8: ASPEN vs. STO-SQP and two variants of SVR-SQP. Distance to the optimal solution versus the number of scalar products (FEV): a) *Mushrooms* dataset; b) *a9a* dataset; c) *Australian* dataset; d) *Heart* dataset; e) *Splice* dataset; f) *MNIST* dataset.

variant with $b = 128$ shows only a marginal reduction of the distance within the considered budget.

Figure 7.8b) shows the results on the *a9a* dataset. Here ASPEN is again the only method that achieves a distinctly pronounced reduction in the error. After a short plateau, its progress resumes, which suggests that the adaptive sample size strategy is able to react to the local difficulty of the problem. Sto-SQP yields a steady but only moderate decrease, while both SVR-SQP variants remain relatively flat, indicating that on this dataset they would likely require either a larger computational budget or a different internal batch-size regime.

Figure 7.8c) refers to the *Australian* dataset. On this problem, ASPEN again attains the largest reduction of the optimality gap and keeps the best performance level over the full FEV range. Sto-SQP decreases more slowly but remains consistent. The SVR-SQP method with $b = 16$ becomes more competitive in the later stage of the run and approaches the accuracy of Sto-SQP, which suggests a certain sensitivity of variance-reduced methods to the statistical profile of the dataset. The variant with $b = 128$ remains the least effective in this comparison.

Figure 7.8d) corresponds to the *Heart* dataset. In this case, ASPEN delivers the fastest initial decrease, reaching a distance below 10^{-1} after only a few hundred scalar products. However, once this rapid initial progress is achieved, the method enters a prolonged plateau. By contrast, SVR-SQP with $b = 16$ improves more gradually, but continues its descent and eventually surpasses ASPEN in terms of final solution accuracy. Sto-SQP and SVR-SQP with $b = 128$ remain less competitive during the whole run.

Figure 7.8e) presents the results for the *Splice* dataset. The same general phenomenon appears here as in the previous case, but in a more pronounced way. ASPEN shows a clear advantage in the early phase of the run, but then exhibits a long interval of stagnation before progress resumes. Meanwhile, SVR-SQP with $b = 16$ decreases more

steadily and ultimately attains the best final accuracy among all tested methods. Sto-SQP stays moderately effective, and the larger-batch SVR-SQP variant remains the slowest. These two datasets therefore illustrate that ASPEN is particularly strong in fast initial reduction of the error, while on some problems a variance-reduced SQP scheme may eventually become preferable in the long run.

Figure 7.8f) corresponds to the *MNIST* dataset. In this experiment, ASPEN ultimately achieves the best final accuracy, although it requires a somewhat larger number of scalar products before clearly outperforming the competing methods. Sto-SQP performs best at the beginning of the run, while both SVR-SQP variants become more favorable than Sto-SQP for larger values of *FEV*. Overall, the MNIST experiment confirms that ASPEN remains highly competitive even on the largest and most demanding dataset in the test collection.

Taken together, these comparisons indicate that ASPEN is especially effective in the regime where fast progress under limited computational budgets is essential. Even in cases where another method eventually reaches a slightly better final point, ASPEN frequently attains a good solution much earlier, which is often the practically relevant criterion in large-scale applications.

The comparison with stochastic SQP-type methods places ASPEN in a broader methodological context and shows that its favorable performance is not limited to simple baselines. Still, real-data experiments combine many effects at once, including geometry, scaling, and dataset-specific structure.

7.1.4 Noise sensitivity and adaptive sample size growth

We conclude the ASPEN experiments with a controlled study aimed at illustrating the relation between data heterogeneity and sample size

adaptation. For this purpose, we consider the academic HS24 problem from the CUTEst collection [27] and introduce Gaussian perturbations into the local objective functions. In this way, we obtain a synthetic stochastic model in which the level of similarity among the local components can be tuned explicitly.

More precisely, we consider the perturbed problem

$$\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{N} \sum_{i=1}^N (\tilde{f}(x) + \varepsilon_i^2 \|x\|^2) \quad \text{s.t.} \quad \|x\|_2^2 = 1, \quad (7.2)$$

where

$$\tilde{f}(x) = (x_1 - 2)^4 + (x_1 - 2x_2)^2$$

is the original objective function of HS24, and the random variables ε_i are drawn from the Gaussian distribution $\mathcal{N}(0, \sigma^2)$. The parameter σ controls the level of perturbation: small values of σ correspond to local functions that are more similar to one another, while larger values of σ produce a more heterogeneous collection of local objectives.

The results reported in Figure 7.9 confirm two important properties of ASPEN. On the one hand, the method remains robust with respect to the optimality gap even as the level of noise increases. On the other hand, the evolution of the sample size clearly reflects the level of heterogeneity in the data. Larger values of σ require more rapid or more substantial sample size growth, which is consistent with the intuitive expectation that more dissimilar local functions demand larger mini-batches in order to provide a reliable approximation of the global objective.

This experiment is particularly valuable because it isolates the effect of heterogeneity from the many confounding factors present in real datasets. It shows directly that the adaptive mechanism in ASPEN is not arbitrary: the algorithm enlarges the sample precisely when the local models become less representative of one another, thereby con-

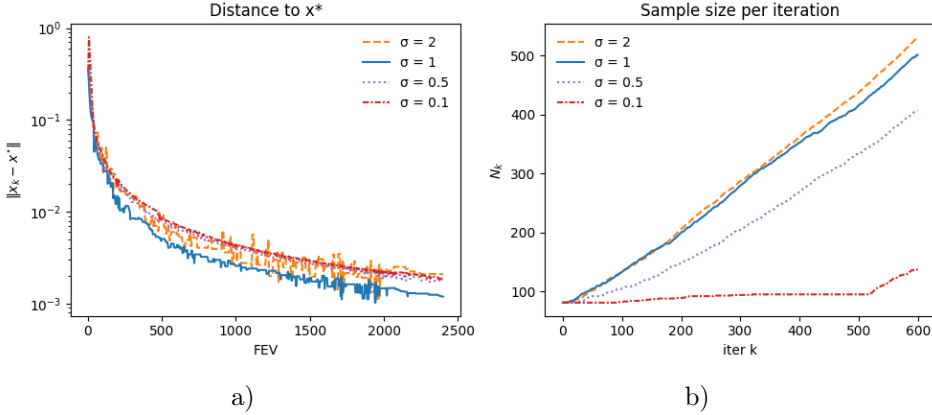


Figure 7.9: Influence of dissimilarity of local cost functions (modeled by $\sigma \in \{0.1, 0.5, 1, 2\}$) on the behavior of ASPEN algorithm applied on problem (7.2): a) optimality gap; b) sample size increase.

firming the intended role of additional sampling as a data-sensitive control mechanism.

With this final experiment, the numerical study of ASPEN is rounded off from three complementary angles: comparison with basic references, comparison with established stochastic constrained methods, and sensitivity to controlled variation in local-model similarity.

7.2 Box-Constrained Problems: AS-BOX

This part of the numerical study is devoted to the box-constrained framework introduced in the previous section. The goal is to assess the practical behavior of the AS-BOX method on representative convex and nonconvex learning problems, and to compare it with algorithms that are specifically designed for stochastic optimization under bound constraints. In contrast to the equality-constrained ex-

periments, where the emphasis was on penalty-based behavior and adaptive sampling under nonlinear constraints, the present subsection focuses on the interaction between projected steps, box feasibility, and additional sampling.

In particular, the numerical study below is organized so as to highlight three aspects of AS-BOX: its efficiency on standard box-constrained classification models, the way its adaptive sample size mechanism evolves in practice, and its robustness when the objective function is nonconvex. This provides a complementary experimental perspective to the theoretical results established in the previous chapter.

7.2.1 Experimental setting and benchmark methods

We now describe the computational setting used for the AS-BOX experiments. The proposed algorithm is tested against representative methods from the literature for smooth stochastic optimization with box constraints. Our main comparison is carried out with the stochastic gradient-based interior-point method (SIPM) introduced in [20], which is specifically designed for smooth box-constrained problems. As a second benchmark, we consider PSGM, the projected stochastic gradient method used in [20]. In PSGM, each iteration is based on a stochastic gradient step followed by an explicit projection onto the feasible box. In order to ensure a fair comparison, all parameters related to SIPM and PSGM are taken exactly as reported in the numerical section of [20].

The choice of these two benchmark methods is natural. SIPM represents an interior-point philosophy, where feasibility is handled implicitly through barrier-type mechanisms, while PSGM is a direct projection-based approach. Since AS-BOX also relies on projected

directions but augments them with additional sampling and adaptive sample size control, comparing these three methods makes it possible to isolate the practical effect of the proposed adaptive strategy.

The experiments include two binary classification models:

1. logistic regression, which gives rise to a convex optimization problem;
2. a single-hidden-layer neural network trained with cross-entropy loss, which yields a nonconvex optimization problem.

All experiments are performed on datasets from the LIBSVM repository, namely *Mushrooms* (8124 samples, 112 features) and *IJCNN1* (49990 samples, 22 features). These two datasets are sufficiently different in size and structure to provide a meaningful test bed for the considered methods. In all cases, labels are encoded in the binary format $\{-1, 1\}$.

The initial point x_0 is generated independently for each problem by sampling its components uniformly from the interval $[-0.01, 0.01]$. For AS-BOX we use the parameter values

$$D_k = 1, \quad C = 1, \quad \beta = 0.1, \quad \eta = 10^{-4}, \quad c = 10^{-4}, \quad \varepsilon_k = \frac{1}{k^{1.1}}.$$

These values are consistent with the standard settings used in earlier work on additional-sampling methods [40, 42, 46].

As in the ASPEN experiments, the use of a common computational metric is important here because the compared methods do not spend the same amount of work per iteration. For that reason, all numerical comparisons below are reported against the number of scalar products rather than against the raw number of iterations.

7.2.2 Logistic regression under box constraints

We first consider a convex benchmark model in the form of box-constrained logistic regression. This problem is a natural starting point because it is standard in large-scale classification and, at the same time, sufficiently structured to allow a clear interpretation of the observed algorithmic behavior.

The optimization problem is given by

$$\min_{x \in [-1, 1]^n} \frac{1}{N} \sum_{i=1}^N \log(1 + \exp(-b_i a_i^\top x)),$$

where (a_i, b_i) are the training samples and $b_i \in \{-1, 1\}$ are the binary labels. This model is convex and widely used as a baseline in large-scale classification. The box constraints $[-1, 1]^n$ are imposed in order to match the setup used in [20] and to enable a direct comparison with existing stochastic box-constrained methods. As before, computational effort is measured by FEV_k , the number of scalar products needed by a given method to compute the iterate x_k , starting from the initial point x_0 .

To assess performance, we plot the distance to the reference solution,

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

against the computational budget FEV_k . In all reported comparisons, the stopping rule is not based on a stationarity tolerance, but on a prescribed budget of scalar products. Hence, all methods are evaluated under the same computational limitation.

Figure 7.10 displays the results on the *Mushrooms* dataset for four algorithms: AS-BOX, SIPM, PSGM, and the full sample variant of AS-BOX, that is, the version in which $N_k = N$ for all iterations. The plot shows the Euclidean distance $\|x_k - x^*\|$ as a function of

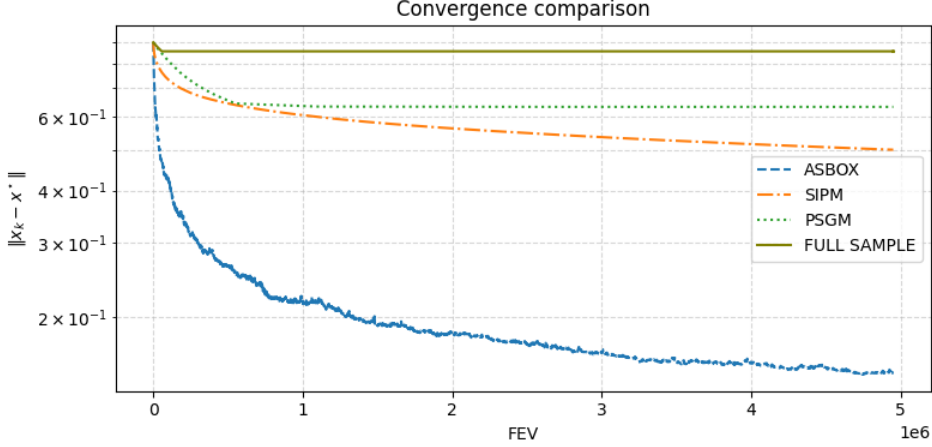


Figure 7.10: Distance to the solution $\|x_k - x^*\|$ versus FEV_k for logistic regression on the *Mushrooms* dataset.

the number of scalar products. Among the compared methods, AS-BOX reaches the reference solution neighborhood the fastest. SIPM decreases more slowly, but in a stable and consistent manner, whereas PSGM exhibits clear signs of stagnation. On this dataset, the full sample variant performs the worst, which suggests that always using the complete dataset leads to an unnecessarily high per-iteration cost without compensating gains in progress.

The same general picture appears on the *IJCNN1* dataset; see Figure 7.11. AS-BOX again provides the best performance and achieves the smallest distance to the reference solution. In particular, it reaches an accuracy level of approximately 10^{-1} between 150,000 and 200,000 scalar products. SIPM follows with a visibly slower reduction of the distance, while PSGM remains the least competitive among the sampled methods. The full sample strategy performs slightly better than PSGM on this dataset, but it still remains clearly inferior to AS-BOX and SIPM.

These two experiments indicate that the advantage of AS-BOX is not tied to a single problem size or data regime. On both a medium-scale dataset and a significantly larger one, the method benefits from using only a relatively small portion of the data at each iteration while still producing more rapid reduction of the solution error.

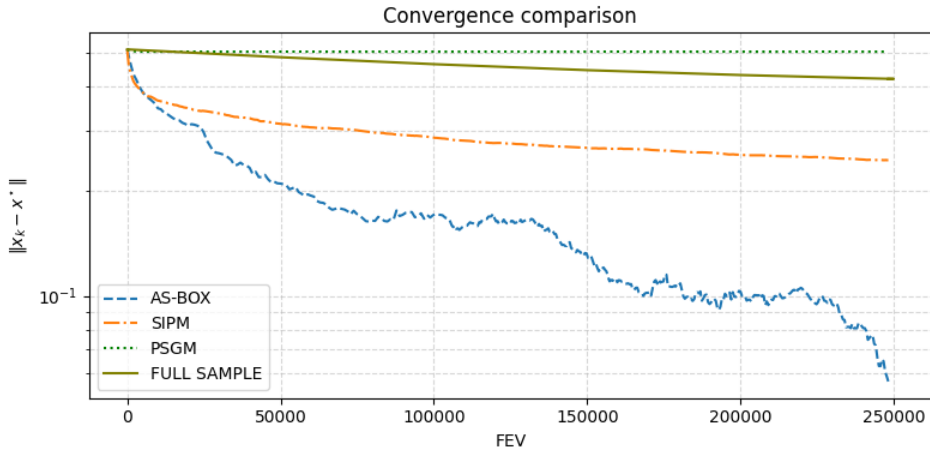


Figure 7.11: Distance to the solution $\|x_k - x^*\|$ versus FEV_k for logistic regression on the *IJCNN1* dataset.

7.2.3 Adaptive subsample growth

We now turn to the adaptive behavior of the sample size sequence generated by AS-BOX. Since one of the central ingredients of the proposed method is the possibility of increasing the sample size only when necessary, it is natural to inspect how this mechanism behaves in practice.

The following plots are particularly informative because they reveal whether the sample size adaptation is merely a formal ingredient of

the algorithm or whether it actually governs the optimization process in a meaningful way. In particular, they show whether AS-BOX tends to drift toward the full sample regime or whether it is able to remain efficient while using only small subsamples.

Figure 7.12 reports the evolution of the subsample size N_k as a function of the computational budget FEV_k for the two datasets. The observed growth is monotone and staircase-like: the sample size is increased only when the additional-sampling acceptance test fails, and it remains unchanged otherwise. On the *Mushrooms* dataset ($m = 8,124$), the largest observed sample size is 168, which is about 2.1% of the full dataset. On the *IJCNN1* dataset ($m = 49,990$), the maximum sample size is 504, which corresponds to approximately 1.0% of the available data. In neither case does the method reach the full sample.

This is an important observation. It shows that AS-BOX is capable of making stable progress without ever resorting to a complete evaluation of the objective function. Consequently, its practical efficiency comes not only from the use of stochastic directions, but also from the fact that the adaptive strategy prevents the sample size from growing unnecessarily.

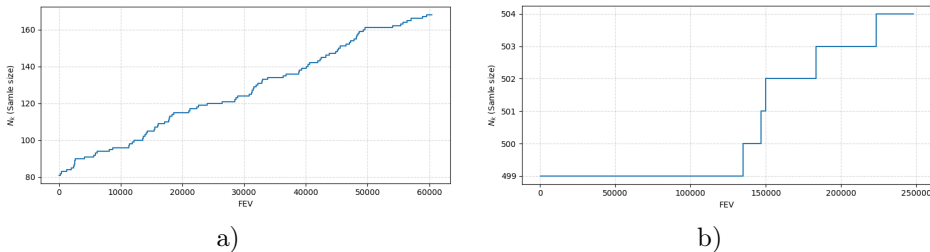


Figure 7.12: AS-BOX: evolution of the subsample size N_k as a function of FEV_k . Part a): N_k versus FEV_k for the *Mushrooms* dataset. Part b): N_k versus FEV_k for the *IJCNN1* dataset.

From a practical standpoint, these sample size trajectories sup-

port the intended interpretation of additional sampling: the method requests a larger batch only when the current one ceases to be sufficiently representative. It is also worth noting that the full sample was not reached on either dataset during the reported runs. Although reaching the full sample is not a required objective of the method, this behavior illustrates that the adaptive sampling rule may keep the iteration in a subsampled regime while still allowing visible progress.

7.2.4 Neural network training under box constraints

We finally consider a nonconvex benchmark in order to evaluate the robustness of AS-BOX beyond convex logistic regression. The model is a single-hidden-layer neural network trained for binary classification. This experiment is significantly more challenging because the optimization landscape is nonconvex and the solution is not unique in general.

Including this experiment is important for two reasons. First, it tests whether the favorable behavior of AS-BOX persists outside the convex setting. Second, it provides a setting in which the projected stationarity measure is more informative than the distance to a fixed reference solution, because the training problem is nonconvex and different runs may converge to different stationary points with comparable objective values.

Let $\tanh(\cdot)$ be the activation function in the hidden layer, while the output layer uses a sigmoid activation. The network output is given by

$$\sigma(W_2 \tanh(W_1 a + b_1) + b_2),$$

where W_1 and W_2 are weight matrices, and b_1 and b_2 are bias vectors.

The associated training objective is the average cross-entropy loss

$$\min_{x \in [-1, 1]^d} \frac{1}{N} \sum_{i=1}^N \left[-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i) \right],$$

where the vector x collects all parameters (W_1, W_2, b_1, b_2) , and d denotes the total number of parameters in the network. As in the logistic regression setting, all parameters are restricted to lie in the interval $[-1, 1]$.

Solutions of this nonconvex problem are not unique in general, and therefore the most informative quantity is not the distance to a single reference point, but the projected stationarity measure $\|d(x_k)\|$. For this reason, we report both the evolution of the cross-entropy loss and the evolution of the stationarity measure as functions of the computational cost FEV_k .

Figure 7.13a) shows the loss trajectory on the *Mushrooms* dataset. AS-BOX again provides the best overall behavior, reducing the loss from approximately 2×10^{-1} to below 10^{-2} within about 10^5 function-evaluation units. SIPM decreases more gradually and stabilizes around 5×10^{-2} . PSGM starts with a competitive initial decrease and even appears slightly better than AS-BOX at the very beginning, but it soon stagnates around 3×10^{-2} . This suggests that while plain projected stochastic gradient steps may be effective at the start, they are less reliable in sustaining long-term progress.

Figure 7.13b) presents the corresponding stationarity measure on the same dataset. The behavior is consistent with the loss curves. AS-BOX reduces $\|d(x_k)\|$ from roughly 10^{-1} to about 10^{-2} and does so in a relatively stable way, despite the stochastic fluctuations inherent in the method. SIPM continues to improve but remains above 3×10^{-2} , whereas PSGM levels off near 4×10^{-2} much earlier.

The same type of comparison is given for the *IJCNN1* dataset in Figure 7.14. In part a), all methods begin from an initial loss

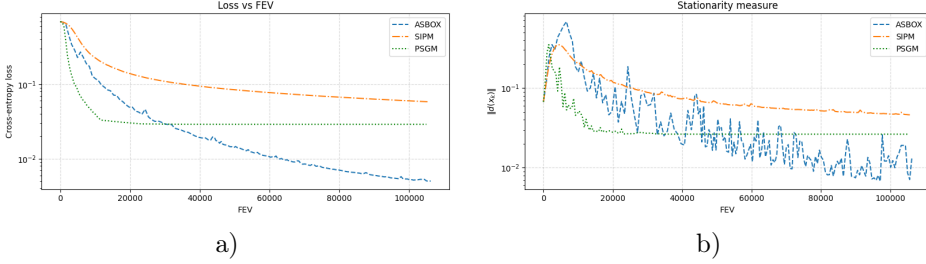


Figure 7.13: Part a): Cross-entropy loss versus FEV_k for the *Mushrooms* dataset. Part b): Stationarity measure $\|d(x_k)\|$ versus FEV_k for the *Mushrooms* dataset.

of approximately 4.2×10^{-1} . AS-BOX then exhibits the fastest and most persistent decrease, reaching a loss below 3.2×10^{-1} after roughly 4×10^5 function-evaluation units. SIPM improves more slowly, while PSGM decreases at the beginning but does not maintain that rate of progress.

Part b) of Figure 7.14 shows the stationarity measure $\|d(x_k)\|$. Once again, AS-BOX displays the most pronounced reduction, eventually going below 2×10^{-2} . SIPM stabilizes around 4×10^{-2} , and PSGM also levels off near that value after an initially faster decrease. Hence, in terms of both objective reduction and projected stationarity, AS-BOX provides the most favorable long-run performance on this more difficult dataset.

Comparing the two datasets leads to a consistent overall conclusion. On both *Mushrooms* and *IJCNN1*, AS-BOX achieves the largest decrease in cross-entropy loss and in the projected stationarity measure within the prescribed computational budget. At the same time, the speed of convergence is not identical across datasets, which is natural given their structural differences. The *Mushrooms* dataset is smaller and more easily separable, so all methods improve more quickly there, and AS-BOX reaches near-optimal behavior already within 10^5

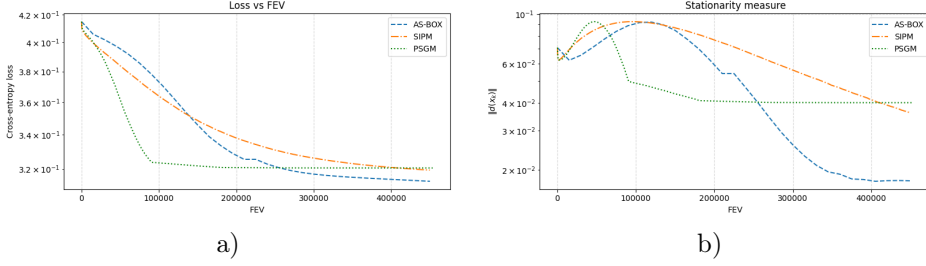


Figure 7.14: Part a): Cross-entropy loss versus FEV_k for the *IJCNN1* dataset. Part b): Stationarity measure $\|d(x_k)\|$ versus FEV_k for the *IJCNN1* dataset.

scalar products. By contrast, *IJCNN1* is larger and less separable, which slows down all methods; nevertheless, AS-BOX preserves a clear advantage over SIPM and PSGM and attains roughly twice the reduction in stationarity by the end of the run.

Taken together, the box-constrained experiments provide empirical evidence that AS-BOX performs well on the considered test problems. In the convex logistic regression experiments, it achieves the best error reduction under a fixed computational budget, while in the neural network experiment it produces the strongest decrease of the projected stationarity measure. Since the numerical study is limited to a finite set of benchmark problems and one neural-network training example, these results should be interpreted as experimental evidence rather than as a universal performance claim. Nevertheless, they suggest that the proposed combination of projected sampled directions, non-monotone line search, and additional-sampling-based sample control can be beneficial beyond the convex setting.

Chapter 8

Conclusion

In this dissertation, first-order stochastic optimization methods based on the additional sampling framework have been presented and analyzed to address computational aspects of large-scale constrained optimization. Driven by applications in machine learning, data analysis, and statistical estimation, contemporary optimization models frequently feature objective functions defined as finite sums or mathematical expectations over large datasets. This thesis addresses this limitation by developing adaptive sampling methodologies that balance approximation accuracy against computational cost. By employing smaller subsamples during the initial iterations and controlling sample expansion through independent validation, the algorithms reduce the computational workload while maintaining asymptotic convergence guarantees. Throughout this work, computational effort has been evaluated using the cumulative number of scalar products (FEV_k), providing a metric to compare the efficiency of the developed methods against deterministic baselines and fixed heuristic stochastic strategies.

The first part of this research addresses optimization problems with nonlinear equality constraints through the formulation of the

ASPEN (Additional Sampling Penalty method for Equality Nonlinear constraints) framework. This methodology combines a quadratic penalty function approach and a nonmonotone Armijo-type line search with an acceptance test performed on an independent auxiliary sample. This auxiliary step evaluates the progress of the candidate iterations, regulating the growth of the main sample size. At the theoretical level, this dissertation establishes the almost sure convergence of the generated iterate sequence to Karush–Kuhn–Tucker points under standard smoothness assumptions and the Linear Independence Constraint Qualification. Furthermore, this framework has been extended to a weighted stochastic finite sum environment, where individual component functions are drawn according to non-uniform probability distributions reflecting their weights.

Parallel to the equality-constrained setting, the dissertation analyzes nonnegativity and general box constraints via the AS-NC and AS-BOX algorithms. These projection-based methods utilize binary and ternary indicator mappings to monitor the structural consistency of the bounding constraints between the main and auxiliary samples. This additional sampling test evaluates whether the current sample size is sufficient to identify the active constraints and the coordinates that remain in the interior of the feasible domain. The theoretical analysis demonstrates that every accumulation point of the generated sequence is almost surely a first-order stationary point of the bound-constrained problem. Additionally, a worst-case expected computational complexity bound of order $\mathcal{O}(\nu^{-2})$ is derived (where, $\nu > 0$ denotes the prescribed stationarity tolerance, for instance for the norm of the projected gradient mapping.), which corresponds to the theoretical rate of classical deterministic projected gradient schemes. Under the assumption of strong convexity, almost sure convergence of the entire sequence of iterates to the unique global solution is established.

The theoretical considerations are evaluated through numerical experiments on real and synthetic datasets from the LIBSVM, MNIST,

and CUTEst repositories. The computational framework encompasses convex models, such as bounded logistic regression, and nonconvex architectures, including single-hidden-layer artificial neural networks, alongside noise-sensitivity analyses. Evaluated against criteria such as objective loss reduction, distance to the reference solution, and the norm of the projected gradient mapping, the empirical results show reduced resource consumption for the adaptive mechanisms. The proposed schemes are compared with standard heuristic stochastic gradient methods, full sample deterministic baselines, as well as stochastic sequential quadratic programming and interior-point algorithms.

The results achieved in this dissertation suggest several possible research directions, particularly the potential generalization of the additional sampling paradigm to optimization problems involving general nonlinear inequality constraints. This extension could be analyzed by combining the penalty architecture of ASPEN with the projected active-set tracking of AS-BOX. By introducing slack variables, general inequality constraints $g(x) \leq 0$ could be transformed into a system of nonlinear equalities with a nonnegativity requirement. In such a framework, the penalty mechanisms might be used to manage the nonlinear equations, while the ternary indicator mappings could be extended to track whether a constraint is active, inactive, or violated on independent auxiliary samples. Investigating the almost sure convergence properties, establishing mixed-constraint complexity bounds, and applying these schemes to large-scale risk-averse decision models may represent possible future objectives of this research.

Bibliography

- [1] S. BELLAVIA, G. GURIOLI, B. MORINI, & P. L. TOINT, (2022). Trust-region algorithms: Probabilistic complexity and intrinsic noise with applications to subsampling techniques. *EURO Journal on Computational Optimization*, 10, 100043.
- [2] A. S. BERAHAS, R. BOLLAPRAGADA, & B. ZHOU, (2022). An adaptive sampling sequential quadratic programming method for equality constrained stochastic optimization. Available at: *arXiv:2206.00712*.
- [3] A. S. BERAHAS, F. E. CURTIS, M. J. O'NEILL, & D. P. ROBINSON, (2024). A stochastic sequential quadratic optimization algorithm for nonlinear-equality-constrained optimization with rank-deficient Jacobians. *Mathematics of Operations Research*, 49(4), 2212-2248.
- [4] A. S. BERAHAS, F. E. CURTIS, D. ROBINSON, & B. ZHOU, (2021). Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM Journal on Optimization*, 31(2), 1352-1379.
- [5] A. S. BERAHAS, J. SHI, Z. YI, B. ZHOU, (2023). Accelerating stochastic sequential quadratic programming for equality con-

- strained optimization using predictive variance reduction, *Computational Optimization and Applications*, 86(1), 79-116.
- [6] E. G. BIRGIN, L. F. BUENO, & J. M. MARTÍNEZ, (2016). Sequential equality-constrained optimization for non-linear programming. *Computational Optimization and Applications*, 65(3), 699-721.
- [7] E. G. BIRGIN, & J. M. MARTÍNEZ, (2012). Augmented Lagrangian method with nonmonotone penalty parameters for constrained optimization. *Computational Optimization and Applications*, 51(3), 941-965.
- [8] E. G. BIRGIN, & J. M. MARTÍNEZ, (2014). Practical augmented Lagrangian methods for constrained optimization. *Society for Industrial and Applied Mathematics*.
- [9] E. G. BIRGIN, & J. M. MARTINEZ, (2019). A Newton-like method with mixed factorizations and cubic regularization for unconstrained minimization. *Computational Optimization and Applications*, 73(3), 707-753.
- [10] E. G. BIRGIN, J. M. MARTÍNEZ, & M. RAYDAN, (2000). Non-monotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10(4), 1196-1211.
- [11] J. F. BONNANS, J. C. GILBERT, C. LEMARÉCHAL, & C. A. SAGASTIZÁBAL, (2006). Numerical optimization: theoretical and practical aspects. *Berlin, Heidelberg: Springer Berlin Heidelberg*.
- [12] R. H. BYRD, S. L. HANSEN, J. NOCEDAL, & Y. SINGER, (2016). A stochastic quasi-Newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2), 1008-1031.

- [13] P. CASCARANO, G. FRANCHINI, E. KOBLER, F. PORTA, & A. SEBASTIANI (2024). A variable metric proximal stochastic gradient method: An application to classification problems. *EURO Journal on Computational Optimization*, 12, 100088-100088.
- [14] C. C. CHANG, C. J. LIN, (2011). LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.
- [15] W. CHENG, D.H. LI, (2008). A derivative-free nonmonotone line search and its applications to the spectral residual method. *IMA Journal of Numerical Analysis*, 29, pp. 814-825.
- [16] E. K. CHONG, W. S. LU, & S. H. ZAK, (2023). An introduction to optimization: With applications to machine learning. *John Wiley & Sons*.
- [17] F. E. CURTIS, (2012). A penalty-interior-point algorithm for nonlinear constrained optimization. *Mathematical Programming Computation*, 4(2), 181-209.
- [18] F. E. CURTIS, V. KUNGURTSEV, D. P. ROBINSON, & Q. WANG, (2025). A stochastic-gradient-based interior-point algorithm for solving smooth bound-constrained optimization problems. *SIAM Journal on Optimization*, 35(2), 1030-1059.
- [19] F. E. CURTIS, S. LIU, & D. P. ROBINSON, (2024). Fair machine learning through constrained stochastic optimization and an ϵ -constraint method. *Optimization Letters*, 18(9), 1975-1991.
- [20] F. E. CURTIS, D. P. ROBINSON, & B. ZHOU, (2021). Inexact sequential quadratic optimization for minimizing a stochastic objective function subject to deterministic nonlinear equality constraints. Available at: *arXiv: 2107.03512*.

- [21] F. E. CURTIS, M. J. O'NEILL, & D. P. ROBINSON, (2024). Worst-case complexity of an SQP method for nonlinear equality constrained stochastic optimization. *Mathematical Programming*, 205(1), 431-483.
- [22] Y. H. DAI, (2002). On the nonmonotone line search. *Journal of Optimization Theory and Applications*, 112(2), 315-330.
- [23] D. DI SERAFINO, N. KREJIĆ, N. KRKLEC JERINKIĆ, M. VIOLA, (2023). LSOS: line search Second-Order Stochastic optimization methods for nonconvex finite sums, *Mathematics of Computation*, 92(341), 1273-1299.
- [24] Y. FANG, S. NA, M. W. MAHONEY, & M. KOLAR, (2024). Fully stochastic trust-region sequential quadratic programming for equality-constrained optimization problems. *SIAM Journal on Optimization*, 34(2), 2007-2037.
- [25] A. GALÁNTAI, (2003). Projectors and projection methods (Vol. 6). *Springer Science & Business Media*.
- [26] A. GENEVAY, M. CUTURI, G. PEYRÉ, & F. BACH, F. (2016). Stochastic optimization for large-scale optimal transport. *Advances in Neural Information Processing Systems*, 29.
- [27] N. I. M. GOULD, D. ORBAN, & P. L. TOINT, (2015). CUTEst: a Constrained and Unconstrained Testing Environment with safe threads for mathematical optimization. *Computational Optimization and Applications*, 60, 545-557.
- [28] I. GRIVA, S. G. NASH, & A. SOFER, (2008). Linear and nonlinear optimization 2nd edition. *Society for Industrial and Applied Mathematics*.

- [29] L. GRIPPO, F. LAMPARIELLO, & S. LUCIDI, (1986). A non-monotone line search technique for Newton's method. *SIAM journal on Numerical Analysis*, 23(4), 707-716.
- [30] L. GRIPPO, F. LAMPARIELLO, & S. LUCIDI, (1991). A class of nonmonotone stabilization methods in unconstrained optimization. *Numerische Mathematik*, 59(1), 779-805.
- [31] W. W. HAGER, (2021). Applied numerical linear algebra. *Society for Industrial and Applied Mathematics*.
- [32] T. HOMEM-DE-MELLO, (2003). Variable-sample methods for stochastic optimization. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 13(2), 108-133.
- [33] W. HUYER, & A. NEUMAIER, (2003). A new exact penalty function. *SIAM Journal on Optimization*, 13(4), 1141-1158.
- [34] A. N. IUSEM, A. JOFRÉ, R. I. OLIVEIRA, & P. THOMPSON, (2019). Variance-based extragradient methods with line search for stochastic variational inequalities. *SIAM Journal on Optimization*, 29(1), 175-206.
- [35] A. N. IUSEM, A. JOFRÉ, R. I. OLIVEIRA, & P. THOMPSON, (2017). Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2), 686-724.
- [36] N. KREJIĆ, N., & N. KRKLEC, (2013). Line search methods with variable sample size for unconstrained optimization. *Journal of Computational and Applied Mathematics*, 245, 213-231.
- [37] N. KREJIĆ, N. KRKLEC, & A. ROŽNJIĆ, (2018). Variable sample size method for equality constrained optimization problems. *Optimization Letters*, 12, 485-497.

- [38] N. KREJIĆ & N. KRKLEC JERINKIĆ, (2015). Nonmonotone line search methods with variable sample size. *Numerical Algorithms*, 68(4), 711-739.
- [39] N. KREJIĆ, N. KRKLEC JERINKIĆ, A. MARTINEZ, & M. YOUSEFI, (2024). A non-monotone trust-region method with noisy oracles and additional sampling. *Computational Optimization and Applications*, 89(1), 247-278.
- [40] N. KREJIĆ, N. KRKLEC JERINKIĆ, S. RAPAJIĆ, & L. RUTEŠIĆ, (2025). IPAS: An adaptive sample size method for weighted finite sum problems with linear equality constraints. Available at: *arXiv: 2504.19629*.
- [41] N. KREJIĆ, N. KRKLEC JERINKIĆ, T. OSTOJIĆ, & N. VUČIĆEVIĆ, (2025). AS-BOX: Additional sampling method for weighted sum problems with box constraints. *Numerical Algorithms*, 1-26.
- [42] N. KREJIĆ, N. K. JERINKIĆ, T. OSTOJIĆ, & N. VUČIĆEVIĆ, (2025). ASPEN: An additional sampling penalty method for finite sum optimization problems with nonlinear equality constraints. *arXiv preprint arXiv:2508.02299*.
- [43] N. KRKLEC JERINKIĆ, (2014). Line search methods with variable sample size (Doctoral dissertation, University of Novi Sad Serbia).
- [44] N. KRKLEC JERINKIĆ, & T. OSTOJIĆ, (2024). AN-SPS: Adaptive sample size nonmonotone line search spectral projected sub-gradient method for convex constrained optimization problems, *Optimization Methods and Software*, 39(5), 1143-1167.

- [45] N. KRKLEC JERINKIĆ, F. PORTA, V. RUGGIERO, & I. TROMBINI, (2025). Variable metric proximal stochastic gradient methods with additional sampling, *Computational Optimization and Applications*, <https://doi.org/10.1007/s10589-025-00720-w>.
- [46] N. KRKLEC JERINKIĆ, V. RUGGIERO, & I. TROMBINI, (2025). Spectral stochastic gradient method with additional sampling for finite and infinite sums, *Computational Optimization and Applications*, 91 (2), 717–758.
- [47] N. KRKLEC JERINKIĆ, & A. ROŽNJK, (2020). Penalty variable sample size method for solving optimization problems with equality constraints in a form of mathematical expectation. *Numerical Algorithms*, 83(2), 701–718.
- [48] W. LA CRUZ, J. MARTÍNEZ, & M. RAYDAN, (2006). Spectral residual method without gradient information for solving large-scale nonlinear systems of equations. *Mathematics of Computation*, 75(255), 1429–1448.
- [49] G. LAN, (2020). First-order and stochastic optimization methods for machine learning, *Cham: Springer International Publishing*, (Vol. 1), 21–51.
- [50] G. LAN, & Z. ZHOU, (2016). Algorithms for stochastic optimization with expectation constraints. *arXiv preprint arXiv:1604.03887*.
- [51] D. H. LI, & M. FUKUSHIMA, (2000). A derivative-free line search and global convergence of Broyden-like method for nonlinear equations. *Optimization Methods and Software*, 13(3), 181–201.
- [52] M. LICHMAN, (2013). UCI machine learning repository. Available at: <https://archive.ics.uci.edu/ml/index.php>

- [53] Y. LECUN, C. CORTES, & C. J. C. BURGESS, (1998). The MNIST database of handwritten digits. Available at: <http://yann.lecun.com/exdb/mnist/>
- [54] S. NA, M. ANITESCU, & M. KOLAR, (2023). An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *Mathematical Programming*, 199(1), 721-791.
- [55] S. NA, M. ANITESCU, & M. KOLAR, (2023). Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. *Mathematical Programming*, 202(1), 279-353.
- [56] Y. NANDWANI, A. PATHAK, & P. SINGLA, (2019). A primal dual formulation for deep learning with constraints. *Advances in neural information processing systems*, 32, 12157-12168.
- [57] G. NÉGIAR, G. DRESDNER, A. TSAI, L. EL GHAOU, F. LOCATELLO, R. FREUND, & F. PEDREGOSA, (2020). Stochastic Frank-Wolfe for constrained finite sum minimization. *Proceedings of the 37th international conference on machine learning*, PMLR, 119, 7253-7262.
- [58] J. NOCEDAL, & S. J. WRIGHT, (2006). Numerical optimization. New York, NY: Springer New York.
- [59] E. POLAK, & J. O. ROYSET, (2008). Efficient sample sizes in stochastic nonlinear programming. *Journal of Computational and Applied Mathematics*, 217(2), 301-310.
- [60] H. ROBBINS, & S. MONRO, (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 400-407.

- [61] S. K. ROY, Z. MHAMMEDI, & M. HARANDI, (2018). Geometry aware constrained optimization techniques for deep learning. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 4460-4469.
- [62] S. SANGALLI, E. ERDIL, A. HÖTKER, O. DONATI, & E. KONUKOGLU, (2021). Constrained optimization to train neural networks on critical and under-represented classes. *Advances in Neural Information Processing Systems*, 34, 25400-25411.
- [63] B. SEN, (2018). A gentle introduction to empirical process theory and applications. *Lecture Notes, Columbia University*, 11, 28-29.
- [64] A. SHAPIRO, (1989). Asymptotic properties of statistical estimators in stochastic programming. *The Annals of Statistics*, 841-858.
- [65] A. SHAPIRO, & Y. WARDI, (1996). Convergence analysis of stochastic algorithms. *Mathematics of Operations Research*, 21(3), 615-628.
- [66] J. C. SPALL, (2005). Introduction to stochastic search and optimization: estimation, simulation, and control. *John Wiley & Sons*.
- [67] G. STRANG, (2022). Introduction to linear algebra. *Wellesley-Cambridge Press*.
- [68] R. Y. SUN, (2020). Optimization for deep learning: An overview. *Journal of the Operations Research Society of China*, 8(2), 249-294.
- [69] L. N. TREFETHEN, & D. BAU, (2022). Numerical linear algebra. *Society for Industrial and Applied Mathematics*.

- [70] Y. WARDI, (1988). A stochastic steepest-descent algorithm. *Journal of Optimization Theory and Applications*, 59(2), 307-323.
- [71] Y. WARDI, (1990). Stochastic algorithms with Armijo stepsizes for minimization of functions. *Journal of Optimization Theory and Applications*, 64(2), 399-417.
- [72] W. WANG, & S. AHMED, (2008). Sample average approximation of expected value constrained stochastic programs. *Operations Research Letters*, 36(5), 515-519.
- [73] X. WANG, S. MA, & Y. X. YUAN, (2017). Penalty methods with stochastic approximation for stochastic nonlinear programming. *Mathematics of Computation*, 86(306), 1793-1820.
- [74] Y. XIA, H. LEUNG, & J. WANG, (2002). A projection neural network and its application to constrained optimization problems. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 49(4), 447-458.
- [75] H. ZHANG, & W. W. HAGER, (2004). A nonmonotone line search technique and its application to unconstrained optimization. *SIAM journal on Optimization*, 14(4), 1043-1056.
- [76] B. ZHOU, L. GAO, & Y. H. DAI, (2006). Gradient methods with adaptive step-sizes. *Computational Optimization and Applications*, 35(1), 69-86.

Short Biography



Nemanja Vučićević was born on July 27, 1996, in Kragujevac, Serbia. He completed his elementary education at the "Sveti Sava" school in Toponica, graduating as the valedictorian ("Vukova diploma" recipient). He graduated from the First Kragujevac Gymnasium, attending the specialized department for mathematically gifted students, where he actively participated and won awards in mathematics, physics, astronomy, and geography, ranging from municipal to national levels.

In 2015, he enrolled in the Bachelor's program in Mathematics, specializing in Theoretical Mathematics, at the Faculty of Science, University of Kragujevac, which he completed in 2019 with a GPA of 10.00. He earned his Master's degree from the same faculty in 2020, also with a GPA of 10.00. The same year, he enrolled in the PhD program at the Faculty of Sciences, University of Novi Sad. He has passed all curriculum exams with a GPA of 10.00.

Since 2018, Nemanja has been employed at the Faculty of Science, University of Kragujevac, within the Institute for Mathematics and Informatics. He initially served as a student demonstrator, then as a teaching assistant associate, and currently holds the position of Teaching Assistant, specializing in the field of Mathematical Analysis with Applications.

During his PhD studies, Nemanja has actively participated in scientific research, as well as various national and international conferences and seminars, further enriching his academic and research experience.

Novi Sad, May 2026

Nemanja Vučićević

Овај Образац чини саставни део докторске дисертације, односно докторског уметничког пројекта који се брани на Универзитету у Новом Саду. Попуњен Образац укорицити иза текста докторске дисертације, односно докторског уметничког пројекта.

План третмана података

Назив пројекта/истраживања
Методе додатног узорковања за проблеме коначних сума са ограничењима Additional Sampling Methods for Finite Sum Constrained 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 академске 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