Springer Optimization and Its Applications

Volume 137

Aims and Scope
Optimization has been expanding in all directions at an astonishing rate during the last few decades. New algorithmic and theoretical techniques have been developed, the diffusion into other disciplines has proceeded at a rapid pace, and our knowledge of all aspects of the field has grown even more profound. At the same time, one of the most striking trends in optimization is the constantly increasing emphasis on the interdisciplinary nature of the field. Optimization has been a basic tool in all areas of applied mathematics, engineering, medicine, economics and other sciences.

The series *Springer Optimization and Its Applications* publishes undergraduate and graduate textbooks, monographs and state-of-the-art expository works that focus on algorithms for solving optimization problems and also study applications involving such problems. Some of the topics covered include nonlinear optimization (convex and nonconvex), network flow problems, stochastic optimization, optimal control, discrete optimization, multi-objective programming, description of software packages, approximation techniques and heuristic approaches.

More information about this series at http://www.springer.com/series/7393

Yurii Nesterov

Lectures on Convex Optimization

Second Edition

Springer

Yurii Nesterov
CORE/INMA
Catholic University of Louvain
Louvain-la-Neuve, Belgium

ISSN 1931-6828 ISSN 1931-6836 (electronic)
Springer Optimization and Its Applications
ISBN 978-3-319-91577-7 ISBN 978-3-319-91578-4 (eBook)
https://doi.org/10.1007/978-3-319-91578-4

Library of Congress Control Number: 2018949149

Mathematics Subject Classification (2010): 49M15, 49M29, 49N15, 65K05, 65K10, 90C25, 90C30, 90C46, 90C51, 90C52, 90C60

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my wife Svetlana

Preface

The idea of writing this book came from the editors of Springer, who suggested that the author should think about a renewal of the book

Introductory Lectures on Convex Optimization: Basic Course,

which was published by Kluwer in 2003 [39]. In fact, the main part of this book was written in the period 1997–1998, so its material is at least twenty years old. For such a lively field as Convex Optimization, this is indeed a long time.

However, having started to work with the text, the author realized very quickly that this modest goal was simply unreachable. The main idea of [39] was to present a *short* one-semester course (12 lectures) on Convex Optimization, which reflected the main algorithmic achievements in the field at the time. Therefore, some important notions and ideas, especially related to all kinds of Duality Theory, were eliminated from the contents without any remorse. In some sense, [39] still remains the *minimal course* representing the basic concepts of algorithmic Convex Optimization. Any enlargements to this text would require difficult explanations as to why the selected material is more important than the many other interesting candidates which have been left on the shelf.

Thus, the author came to a hard decision to write a *new book*, which includes all of the material of [39], along with the most important advances in the field during the last two decades. From the chronological point of view, this book covers the period up to the year 2012.[1] Therefore, the newer results on random coordinate descent methods and universal methods, complexity results on zero-order algorithms and methods for solving huge-scale problems are still missing. However, in our opinion, these very interesting topics have not yet matured enough for a monographic presentation, especially in the form of lectures.

From the methodological point of view, the main novelty of this book consists in the wide presence of duality. Now the reader can see the story from both sides,

[1]Well, just for consistency, we added the results from several last-minute publications, which are important for the topics discussed in the book.

primal and dual. As compared to [39], the size of the book is doubled, which looks to be a reasonable price to pay for a comprehensive presentation. Clearly, this book is too big now to be taught during one-semester. However, it fits well a two-semester term. Alternatively, different parts of it can be used in diverse educational programs on modern optimization. We discuss possible variants at the end of the Introduction.

In this book we include three topics, which are new to the monographic literature.

- **The smoothing technique.** This approach has completely changed our understanding of complexity of nonsmooth optimization problems, which arise in the vast majority of applications. It is based on the *algorithmic possibility* of approximating a non-differentiable convex function by a smooth one, and minimizing the new objective by Fast Gradient Methods. As compared with standard subgradient methods, the complexity of each iteration of the new schemes does not change. However, the estimate for the number of iterations of these schemes becomes proportional to the *square root* of this number for the standard methods. Since in practice, these numbers are usually of the order of many thousands, or even millions, the gain in computational time becomes spectacular.
- **Global complexity bounds for second-order methods.** Second-order methods, and their most famous representative, the Newton's Method, are among the oldest schemes in Numerical Analysis. However, their global complexity analysis has only recently been carried out, after the discovery of the Cubic Regularization of Newton's Method. For this new variant of classical scheme, we can write down the global complexity bounds for different problem classes. Consequently, we can now compare global efficiency of different second-order methods and develop *accelerated schemes*. A completely new feature of these methods is the accumulation of some model of the objective function during the minimization process. At the same time, we can derive for them lower complexity bounds and develop optimal second-order methods. Similar modifications can be made for methods solving systems of nonlinear equations.
- **Optimization in relative scale.** The standard way of defining an approximate solution of an optimization problem consists in introducing absolute accuracy. However, in many engineering applications, it is natural to measure the quality of solution in a *relative scale* (percent). To adjust minimization methods toward this goal, we introduce a special model of objective function and apply efficient preprocessing algorithms for computing an appropriate metric, compatible with the topology of the objective. As a result, we get very efficient optimization methods with a weak dependence of their complexity bounds in the size of input data.

We hope that this book will be useful for a wide audience, including students with mathematical, economical, and engineering specializations, practitioners of different fields, and researchers in Optimization Theory, Operations Research, and Computer Science. The main lesson of the development of our field in the last few decades is that efficient optimization methods can be developed only by intelligently employing the structure of particular instances of problems. In order to do this, it is always useful to look at successful examples. We believe that this book will provide the interested reader with a great deal of information of this type.

Louvain-la-Neuve, Belgium Yurii Nesterov
January 2018

Acknowledgements

Through my scientific career, I have had an extraordinary opportunity of being able to have regular scientific discussions with Arkady Nemirovsky. His remarkable mathematical intuition and profound mathematical culture helped me enormously in my scientific research. Boris Polyak has remained my scientific adviser starting from the time of my PhD, for almost four decades. His scientific longevity has set a very stimulating example. I am very thankful to my colleagues A. d'Aspremont, A. Antipin, V. Blondel, O. Burdakov, C. Cartis, F. Glineur, C. Gonzaga, R. Freund, A. Juditsky, H.-J. Lüthi, B. Mordukhovich, M. Overton, R. Polyak, V. Protasov, J. Renegar, P. Richtarik, R. Sepulchre, K. Scheinberg, A. Shapiro, S.Shpirko, Y. Smeers, L. Tuncel, P. Vandooren, J.-Ph. Vial, and R. Weismantel for our regular scientific discussions resulting from time to time in a joint paper. In the recent years, my contact with young researchers P. Dvurechensky, N. Doikov, A. Gasnikov, G. Grapiglia, R. Hildebrand, A. Rodomanov, and V.Shikhman has been very interesting and stimulating. At the same time, I am convinced that the excellent conditions for research, provided me by Université Catholique de Louvain (UCL), is a result of continuous support (over several decades!) from the patriarchs of UCL Jacques Dreze, Michele Gevers, and Laurence Wolsey. To all these people, I express my sincere gratitude.

The contents of this book have already been presented in several educational courses. I am very thankful to C. Helmberg, R. Freund, B. Legat, J. Renegar, H. Sendov, A. Tits, M. Todd, L. Tuncel, and P. Weiss for reporting to me a number of misprints in [39]. In the period 2011–2017 I had the very useful opportunity of presenting some parts of the new material in several advanced courses on Modern Convex Optimization at different universities over the world (University of Liege, ENSAE (ParisTech), University of Vienna, Max Planck Institute (Saarbrucken), FIM (ETH Zurich), Ecole Polytechnique, Higher School of Economics (Moscow), Korea Advanced Institute of Science Technology (Daejeon), Chinese Academy of Sciences (Beijing)). I am very thankful to all these people and institutions for their interest in my research.

Finally, only the patience and professionalism of Springer editors Anne-Kathrin Birchley-Brun and Rémi Lodh has made the publication of this book possible.

Introduction

Optimization problems arise naturally in many different fields. Very often, at some point we get a craving to arrange things in the best possible way. This intention, converted into a mathematical formulation, becomes an optimization problem of a certain type. Depending on the field of interest, it could be an optimal design problem, an optimal control problem, an optimal location problem, an optimal diet problem, etc. However, the next step, consisting in finding a solution to the mathematical model, is far from being trivial. At first glance, everything looks very simple: many commercial optimization packages are easily available and any user can get a "solution" to the model just by clicking at an icon on the desktop of a personal computer. However, the question is, what do we actually get? How much can we trust the answer?

One of the goals of this course is to show that, despite their easy availability, the proposed "solutions" of general optimization problems very often cannot satisfy the expectations of a naive user. In our opinion, the main fact, which should be known to any person dealing with optimization models, is that in general, *optimization problems are unsolvable*. This statement, which is usually missing in standard optimization courses, is very important for understanding optimization theory and the logic of its development in the past and in the future.

In many practical applications, the process of creating a model can take a lot of time and effort. Therefore, the researchers should have a clear understanding of the properties of the model they are constructing. At the stage of modelling, many different ideas can be applied to represent a real-life situation, and it is absolutely necessary to understand the computational consequences of each step in this process. Very often, we have to choose between a "perfect" model, which we cannot solve,[2] and a "sketchy" model, which can be solved for sure. What is better?

In fact, computational practice provides us with an answer. Up to now, the most widespread optimization models have been the models of *Linear Optimization*. It is very unlikely that such models can describe our nonlinear world very well. Hence,

[2]More precisely, which we can only *try* to solve.

the main reason for their popularity is that practitioners prefer to deal with solvable models. Of course, very often the linear approximations are poor. However, usually it is possible to predict the consequences of such a choice and make a correction in interpretation of the obtained solution. This is much better than trying to solve an overcomplicated model without any guarantee of success.

Another goal of this course consists in discussing numerical methods for *solvable* nonlinear models, namely the problems of *Convex Optimization*. The development of Convex Optimization in the last decades has been very rapid and exciting. Now it consists of several competing branches, each of which has some strong and some weak points. We will discuss their features in detail, taking into account the historical aspect. More precisely, we will try to understand the internal logic of the development of each branch of the field. Up to now, the main results of these developments could only be found in specialized journals. However, in our opinion, many of these theoretical achievements are ready to be understood by the final users: computer scientists, industrial engineers, economists, and students of different specializations. We hope that this book will be interesting even for experts in optimization theory since it contains many results which have never been published in a monograph.

In this book, we will try to convince the reader that, in order to work with optimization formulations successfully, it is necessary to be aware of some theory, which explains what we can and what we cannot do with optimization problems. The elements of this simple theory can be found in almost every chapter of the first part of the book, dealing with the standard black-box model of the objective function. We will see that Black-Box Convex Optimization is an excellent example of a *comprehensive* application theory, which is simple, easy to learn and which can be very useful in practical applications. On the other hand, in the second part of the book, we will see how much we can gain from a proper use of a problem's structure. This enormous increase of our abilities does not discard the results of the first part. On the contrary, most of the achievements in Structural Optimization are firmly supported by the fundamental methods of Black-Box Convex Optimization.

In this book, we discuss the most efficient modern optimization schemes and establish for them global efficiency bounds. Our presentation is self-contained; we prove all necessary results. Nevertheless, the proofs and reasonings should not be a problem, even for a second-year undergraduate student.

The structure of the book is as follows. It consists of seven relatively independent chapters. Each chapter includes three or four sections. Most of them correspond approximately to a two-hour lecture. Thus, the contents of the book can be directly used for a standard two-semester course on Convex Optimization. Of course, different subsets of the chapters can be useful for a smaller course.

The whole contents is divided into two parts. Part I, which includes Chaps. 1–4, contains all the material related to the Black-Box model of optimization problem. In this framework, additional information on the given problem can be obtained only by request, which corresponds to a particular set of values of the decision variables. Typically, the result of this request is either the value of the objective function, or this value and the gradient, etc. This framework is the most advanced part of Convex Optimization Theory.

Chapter 1 is devoted to *general optimization* problems. In Sect. 1.1, we introduce the terminology, the notions of oracle, black box, functional model of an optimization problem and the complexity of general iterative schemes. We prove that global optimization problems are “unsolvable” and discuss the main features of different fields of optimization theory. In Sect. 1.2, we discuss two main local unconstrained minimization schemes: the gradient method and the Newton’s method. We establish their local rates of convergence and discuss the possible difficulties (divergence, convergence to a saddle point). In Sect. 1.3, we compare the formal structures of the gradient and the Newton’s method. This analysis leads to the idea of a variable metric. We describe quasi-Newton methods and conjugate gradient schemes. We conclude this section with an analysis of different methods for constrained minimization: Lagrangian relaxation with a certificate for global optimality, the penalty function method, and the barrier approach.

In Chap. 2, we consider methods of *smooth convex optimization*. In Sect. 2.1, we analyze the main reason for difficulties encountered in the previous chapter. From this analysis, we *derive* two good functional classes, the classes of smooth convex and smooth strongly convex functions. For corresponding unconstrained minimization problems, we establish the lower complexity bounds. We conclude this section with an analysis of a gradient scheme, which demonstrates that this method is not optimal. The optimal schemes for smooth convex minimization problems, so-called Fast Gradient Methods, are discussed in Sect. 2.2. We start by presenting a special technique for convergence analysis, based on estimating sequences. Initially, it is introduced for problems of Unconstrained Minimization. After that, we introduce convex sets and define a notion of gradient mapping for a problem with simple set constraints. We show that the gradient mapping can formally replace a gradient step in the optimization schemes. In Sect. 2.3, we discuss more complicated problems, which involve several smooth convex functions, namely, the minimax problem and the constrained minimization problem. For both problems we use a notion of gradient mapping and present the optimal schemes.

Chapter 3 is devoted to the theory of *nonsmooth convex optimization*. Since we do not assume that the reader has a background in Convex Analysis, the chapter begins with Sect. 3.1, which contains a compact presentation of all the necessary facts. The final goal of this section is to justify the rules for computing the subgradients of a convex function. At the same time, we also discuss optimality conditions, Fenchel duality and Lagrange multipliers. At the end of the section, we prove several minimax theorems and explain the basic notions justifying the primal-dual optimization schemes. This is the biggest section in the book and it can serve as a basis for a mini-course on Convex Analysis.

The next Sect. 3.2 starts from the lower complexity bounds for nonsmooth optimization problems. After that, we present a general scheme for the complexity analysis of the corresponding methods. We use this scheme in order to establish a convergence rate for the simplest subgradient method and for its switching variant, treating the problems with functional constraints. For the latter scheme, we justify the possibility of approximating optimal Lagrange multipliers. In the remaining part of the section, we consider the two most important finite-dimensional methods: the center-of-gravity method and the ellipsoid method. At the end, we briefly discuss some other cutting plane schemes. Section 3.3 is devoted to the minimization schemes, which employ a piece-wise linear model of a convex function. We describe Kelley's method and show that it can be extremely slow. After that, we introduce the so-called Level Method. We justify its efficiency estimates for unconstrained minimization problems and for problems with functional constraints.

Part I is concluded by Chap. 4, devoted to a global complexity analysis of second-order methods. In Sect. 4.1, we introduce cubic regularization of the Newton method and study its properties. We show that the auxiliary optimization problem in this scheme can be efficiently solved even if the Hessian of the objective function is not positive semidefinite. We study global and local convergence of the Cubic Newton Method in convex and non-convex cases. In Sect. 4.2, we show that this method can be accelerated using the estimating sequences technique.

In Sect. 4.3, we derive lower complexity bounds for second-order methods and present a conceptual optimal scheme. At each iteration of this method, it is necessary to perform a potentially expensive search procedure. Therefore, we conclude that the problem of constructing an efficient optimal second-order scheme remains open.

In the last Sect. 4.4, we consider a modification of the standard Gauss-Newton method for solving systems of nonlinear equations. This modification is also based on an overestimating principle as applied to the norm of the residual of the system. Both global and local convergence results are justified.

In Part II, we include results related to Structural Optimization. In this framework, we have direct access to the elements of optimization problems. We can work with the input data at the preliminary stage, and modify it, if necessary, to make the problem simpler. We show that such a freedom can significantly increase our computational abilities. Very often, we are able to get optimization methods which go far beyond the limits prescribed by the lower complexity bounds of Black-Box Optimization Theory.

In the first chapter of this part, Chap. 5, we present theoretical foundations for polynomial-time interior-point methods. In Sect. 5.1, we discuss a certain contradiction in the Black Box concept as applied to a convex optimization model. We introduce a *barrier model* of an optimization problem, which is based on the notion of a *self-concordant function*. For such functions, the second-order oracle is not local. Moreover, they can easily be minimized by the standard Newton's method. We study the properties of these functions and their dual counterparts.

In the next Sect. 5.2, we study the complexity of minimization of self-concordant functions by different variants of Newton's method. The efficiency of direct minimization is compared with that of a path-following scheme, and it is proved that the latter method is much better.

In Sect. 5.3, we introduce *self-concordant barriers*, a subclass of standard self-concordant functions, which is suitable for sequential unconstrained minimization schemes. We study the properties of such barriers and prove the efficiency estimate of the path-following scheme.

In Sect. 5.4, we consider several examples of optimization problems, for which we can construct a self-concordant barrier. Consequently, these problems can be solved by a polynomial-time path-following scheme. We consider linear and quadratic optimization problems, problems of semidefinite optimization, separable optimization and geometrical optimization, problems with extremal ellipsoids, and problems of approximation in ℓ_p-norms. A special subsection is devoted to a general technique for constructing self-concordant barriers for particular convex sets, which is provided with several application examples. We conclude Chap. 5 with a comparative analysis of performance of an interior-point scheme with a nonsmooth optimization method as applied to a particular problem class.

In Chap. 6, we present different approaches based on the direct use of a primal-dual model of the objective function. First of all, we study a possibility of approximating nonsmooth functions by smooth functions. In the previous chapters, it was shown that in the Black-Box framework smooth optimization problems are much easier than nonsmooth problems. However, any non-differentiable function can be approximated with arbitrary accuracy by a differentiable function. We pay for the better quality of approximation by a higher curvature of the smooth function. In Sect. 6.1, we show how to balance the accuracy of approximation and its curvature in an optimal way. As a result, we develop a technique for creating computable smoothed versions of non-differentiable functions and minimizing them by Fast Gradient Methods described in Chap. 2. The number of iterations of the resulting methods is proportional to the square root of the number of iterations of the standard subgradient scheme. At the same time, the complexity of each iteration does not change. In Sect. 6.2, we show that this technique can also be used in a symmetric primal-dual form. In the next Sect. 6.3, we give an example of application of the smoothing technique to the problems of Semidefinite Programming.

This chapter concludes with Sect. 6.4, where we analyze methods based on minimization of a local model of the objective function. Our optimization problem has a composite objective function equipped with a linear optimization oracle. For this problem, we justify global complexity bounds for two versions of the Conditional Gradient method (the Frank–Wolfe algorithm). It is shown that these methods can compute approximations of the primal-dual problem. At the end of this section, we analyze a new version of the Trust-Region second-order method, for which we obtain the worst-case global complexity guarantee.

In the last Chap. 7, we collect optimization methods which are able to solve problems with a certain relative accuracy. Indeed, in many applications, it is difficult to relate the number of iterations of an optimization scheme with a desired accuracy of the solution since the corresponding inequality contains unknown parameters (Lipschitz constants, distance to the optimum). However, in many cases the required level of relative accuracy is quite understandable. For developing methods which compute solutions with relative accuracy, we need to employ internal structure of the problem. In this chapter, we start from problems of minimizing homogeneous objective functions over a convex set separated from the origin (Sect. 7.1). The availability of a subdifferential of this function at zero provides us with a good metric, which can be used in optimization schemes and in the smoothing technique. If this subdifferential is polyhedral, then the metric can be computed by a cheap preliminary rounding process (Sect. 7.2).

In the next Sect. 7.3, we present a barrier subgradient method, which computes an approximate maximum of a positive convex function with a certain relative accuracy. We show how to apply this method for solving problems of fractional covering, maximal concurrent flow, semidefinite relaxation, online optimization, portfolio management, and others.

We conclude this chapter with Sect. 7.4.1, where we study the possibility of finding good relative approximations to a special class of convex functions, which we call *strictly positive*. For these functions, it is possible to introduce a new notion of *mixed accuracy* (absolute/relative) and develop a quasi-Newton scheme for its efficient approximation. We derive global complexity bounds for this method and show that they are monotone in the dimension of the problem. This means that small dimensions always help.

The book concludes with Bibliographical Comments and an Appendix, where we analyze efficiency of some methods for solving auxiliary optimization problems.

Let us conclude this Introduction by describing some possible combinations of chapters suitable for a course. The most classical one-semester course can be composed by Chaps. 1, 2, 3, and 5. It corresponds more or less to the contents of monograph [39]. The only difference is that in the present book Sect. 3.1 is much bigger and it will be reasonable to restrict the student's attention only to the necessary parts. Chapter 3 can be replaced by Chap. 4, which will yield a course devoted only to differentiable optimization.

All three chapters of Part II are completely independent. At the same time, they can be unified in an advanced one-semester course on Modern Convex Optimization.

Contents