Springer Optimization and Its Applications 137

Yurii Nesterov

# Lectures on Convex Optimization

*Second Edition*

Springer

# 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 $\ell_p$-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

**Part II Structural Optimization**

# Part I
# Black-Box Optimization

# Chapter 1
# Nonlinear Optimization

In this chapter, we introduce the main notations and concepts used in Continuous Optimization. The first theoretical results are related to Complexity Analysis of the problems of Global Optimization. For these problems, we start with a very pessimistic lower performance guarantee. It implies that for any method there exists an optimization problem in $\mathbb{R}^n$ which needs at least $O\left(\frac{1}{\epsilon^n}\right)$ computations of the function values in order to approximate its global solution up to accuracy $\epsilon$. Therefore, in the next section we pass to local optimization, and consider two main methods, the Gradient Method and the Newton Method. For both of them, we establish some local rates of convergence. In the last section, we present some standard methods in General Nonlinear Optimization: the conjugate gradient methods, quasi-Newton methods, theory of Lagrangian relaxation, barrier methods and penalty function methods. For some of them, we prove global convergence results.

## 1.1 The World of Nonlinear Optimization

(General formulation of the problem; Important examples; Black box and iterative methods; Analytical and arithmetical complexity; The Uniform Grid Method; Lower complexity bounds; Lower bounds for global optimization; Identity cards of the fields.)

### *1.1.1 General Formulation of the Problem*

Let us start by fixing the mathematical form of our main problem and the standard terminology. Let $x$ be an $n$-dimensional real vector:

$$x = (x^{(1)}, \dots, x^{(n)})^T \in \mathbb{R}^n,$$

Y. Nesterov, *Lectures on Convex Optimization*, Springer Optimization and Its Applications 137, https://doi.org/10.1007/978-3-319-91578-4_1

and $f_0(\cdot), \ldots, f_m(\cdot)$ be some real-valued functions defined on a set $Q \subseteq \mathbb{R}^n$. In this book, we consider different variants of the following general minimization problem:

$$\begin{aligned} &\min \; f_0(x), \\ &\text{s.t. } f_j(x) \;\&\; 0, \quad j = 1 \ldots m, \\ &x \in Q, \end{aligned} \tag{1.1.1}$$

where the sign & can be $\leq$, $\geq$, or $=$.

We call $f_0(\cdot)$ the *objective* function of our problem, the vector function

$$f(x) = (f_1(x), \ldots, f_m(x))^T$$

is called the vector of *functional constraints*, the set $Q$ is called the *basic feasible set*, and the set

$$\mathscr{F} = \{x \in Q \mid \; f_j(x) \leq 0, \; j = 1 \ldots m\}$$

is called the *(entire) feasible set* of problem (1.1.1). It is just a convention to consider minimization problems. Instead, we could consider maximization problems with the objective function $-f_0(\cdot)$.

There exists a natural classification of the *types* of minimization problems.

- *Constrained problems*: $\mathscr{F} \subsetneq \mathbb{R}^n$.
- *Unconstrained problems*: $\mathscr{F} = \mathbb{R}^n$.[1]
- *Smooth problems*: all $f_j(\cdot)$ are differentiable.
- *Nonsmooth problems*: there are several nondifferentiable components $f_k(\cdot)$.
- *Linearly constrained problems*: the functional constraints are affine:

$$f_j(x) = \sum_{i=1}^{n} a_j^{(i)} x^{(i)} + b_j \equiv \langle a_j, x \rangle + b_j, \; j = 1 \ldots m,$$

  (here $\langle \cdot, \cdot \rangle$ stands for the *inner (or scalar) product* in $\mathbb{R}^n$: $\langle a, x \rangle = a^T x$), and $Q$ is a polyhedron. If $f_0(\cdot)$ is also affine, then (1.1.1) is a *linear optimization problem*. If $f_0(\cdot)$ is quadratic, then (1.1.1) is a *quadratic optimization problem*. If all the functions $f_0(\cdot), \cdot, f_m(\cdot)$ are quadratic, then this is a quadratically constrained quadratic problem.

There is also a classification based on properties of the feasible set.

[1] Sometimes, problems with a "simple" basic feasible set $Q$ and no functional constraints are also treated as "unconstrained" problems. In this case, we need to know how to solve some auxiliary optimization problems over the set $Q$ in a closed form.

- Problem (1.1.1) is called *feasible* if $\mathscr{F} \neq \emptyset$.
- Problem (1.1.1) is called *strictly feasible* if there exists an $x \in Q$ such that $f_j(x) < 0$ (or $> 0$) for all inequality constraints and $f_j(x) = 0$ for all equality constraints. (*Slater condition*.)

Finally, we distinguish different types of solutions to (1.1.1):

- A point $x^* \in \mathscr{F}$ is called the optimal *global solution* to (1.1.1) if $f_0(x^*) \leq f_0(x)$ for all $x \in \mathscr{F}$ (*global minimum*). In this case, $f_0(x^*)$ is called the (global) *optimal value* of the problem.
- A point $x^* \in \mathscr{F}$ is called a *local solution* to (1.1.1) if there exists a set $\hat{\mathscr{F}} \subseteq \mathscr{F}$ such that $x^* \in \mathrm{int}\hat{\mathscr{F}}$ and $f_0(x^*) \leq f_0(x)$ for all $x \in \hat{\mathscr{F}}$ (*local minimum*). If $f_0(x^*) < f_0(x)$ for all $x \in \hat{\mathscr{F}} \setminus \{x^*\}$, then $x^*$ is called *strict* (or *isolated*) local minimum.

Let us consider now several examples representing the main sources of optimization problems.

*Example 1.1.1* Let $x^{(1)}, \ldots, x^{(n)}$ be our *design variables*. Then we can fix some functional *characteristics* of our decision vector $x$: $f_0(x), \ldots, f_m(x)$. For example, we can consider a price of the project, amount of required resources, reliability of the system, etc. We fix the most important characteristic, $f_0(x)$, as our *objective*. For all others, we impose some bounds: $a_j \leq f_j(x) \leq b_j$. Thus, we come to the problem

$$\min_{x \in Q} \; f_0(x),$$

$$\text{s.t. } a_j \leq f_j(x) \leq b_j, \;\; j = 1 \ldots m,$$

where $Q$ stands for the *structural* constraints like nonnegativity, boundedness of some variables, etc. □

*Example 1.1.2* Let our initial problem be as follows:

$$\text{Find } x \in \mathbb{R}^n \text{ such that } f_j(x) = a_j, \;\; j = 1 \ldots m, \tag{1.1.2}$$

where $a_j \in \mathbb{R}$, $j = 1, \ldots, m$. Then we can consider the problem

$$\min_{x \in \mathbb{R}^n} \sum_{j=1}^{m} (f_j(x) - a_j)^2,$$

perhaps even with some additional constraints on $x$. If the optimal value of the latter problem is zero, we conclude that our initial problem (1.1.2) has a solution.

Note that in Nonlinear Analysis the problem (1.1.2) is almost *universal*. It covers ordinary differential equations, partial differential equations, problems arising in Game Theory, and many others. □

*Example 1.1.3* Sometimes our decision variables $x^{(1)}, \dots, x^{(n)}$ must be *integer*. This can be described by the following constraint:

$$\sin(\pi x^{(i)}) = 0, \quad i = 1 \dots n.$$

Thus, we can also treat *integer optimization* problems:

$$\begin{aligned} &\min_{x \in Q} \; f_0(x), \\ &\text{s.t. } a_j \leq f_j(x) \leq b_j, \;\; j = 1 \dots m, \\ &\sin(\pi x^{(i)}) = 0, \;\; i = 1 \dots n. \qquad \square \end{aligned}$$

Looking at these examples, we can easily understand the optimism of the pioneers of Nonlinear Optimization, which can be easily seen in the papers of the 1950s and 1960s. Our first impression should be, of course, as follows:

> Nonlinear Optimization is a very important and promising application theory. It covers almost ALL needs of Operations Research and Numerical Analysis.

However, by looking at the same list of examples, especially at Examples 1.1.2 and 1.1.3, a more experienced (or suspicious) reader could come to the following conjecture.

> *In general, optimization problems should be* UNSOLVABLE (?)

Indeed, from our real-life experience, it is difficult to believe in the existence of a universal tool which is able to solve all problems in the world.

However, suspicions are not the legal instruments of science. It is a question of personal taste how much we can trust them. Therefore, it was definitely one of the most important events in Optimization Theory when, in the middle of 1970s, this conjecture was *proved* in a *strict* mathematical sense. This proof is so important and simple that we cannot avoid it in our course. But first of all, we should introduce a special language which is required for speaking about such things.

### 1.1.2 Performance of Numerical Methods

Let us imagine the following situation. We are going to solve a problem $P$, and we know that there exist many different numerical methods for doing so. Of course, we want to find a scheme which is the best for our $P$. However, it appears that we are looking for something which does not exist. In fact, maybe it does, but it is definitely not recommended to ask the winner for help. Indeed, consider a method for solving problem (1.1.1), which does nothing except report that $x^* = 0$. Of course, this method does not work properly for any problems *except those* which have the optimal solution exactly at the origin, in which case the "performance" of this method is unbeatable.

Hence, we cannot speak about the best method for a particular problem $P$, but we can do so for a *class* of problems $\mathscr{P} \ni P$. Indeed, numerical methods are usually developed to solve many different problems with similar characteristics. Thus, the *performance* of a method $\mathscr{M}$ on the whole class $\mathscr{P}$ can be a natural measure of its efficiency.

Since we are going to speak about the performance of $\mathscr{M}$ on a class $\mathscr{P}$, we should assume that the method $\mathscr{M}$ does not have *complete* information about a particular problem $P$.

> The *known* (to a numerical scheme) "part" of problem $P$ is called the *model* of the problem.

We denote the model by $\Sigma$. Usually the model consists of the formulation of the problem, description of classes of functional components, etc.

In order to recognize the problem $P$ (and solve it), the method should be able to collect specific information about $P$. It is convenient to describe the process of collecting this data via the notion of an *oracle*. An oracle $\mathscr{O}$ is just a unit which answers the successive questions of the methods. The method $\mathscr{M}$ is trying to solve the problem $P$ by collecting and handling the answers.

In general, each problem can be described by different models. Moreover, for each problem we can develop different types of oracles. But let us fix $\Sigma$ and $\mathscr{O}$. In this case, it is natural to define the performance of $\mathscr{M}$ on $(\Sigma, \mathscr{O})$ as its performance on the *worst* $P_{\rm w}$ from $(\Sigma, \mathscr{O})$. Note that this $P_{\rm w}$ can be bad only for $\mathscr{M}$.

Further, what is the *performance* of $\mathscr{M}$ on $P$ ? Let us start from an intuitive definition.

> The performance of $\mathscr{M}$ on $P$ is the total amount of *computational effort* required by method $\mathscr{M}$ to *solve the problem* $P$.

In this definition, there are two additional notions to be specified. First of all, what does "*to solve the problem*" mean? In some situations it could mean finding an *exact* solution. However, in many areas of Numerical Analysis this is impossible (and in Optimization this is definitely the case). Therefore, we accept a relaxed goal.

> Solving the problem means finding an *approximate solution* to $\mathscr{P}$ with some accuracy $\epsilon > 0$.

Again, the meaning of the expession "*with some accuracy* $\epsilon > 0$" is very important for our definitions. However, it is too early to speak about this now. We just introduce the notation $\mathscr{T}_\epsilon$ for a stopping criterion. Its meaning will always be clear for particular problem classes. Now we have a formal description of the problem class:

$$\mathscr{P} \equiv (\Sigma, \mathscr{O}, \mathscr{T}_\epsilon).$$

In order to solve a problem $P$ from $\mathscr{P}$, we apply to it an *iterative process*, which is a natural form of any method which works with an oracle.

**General Iterative Scheme**

**Input:** Starting point $x_0$ and accuracy $\epsilon > 0$.
**Initialization.** Set $k = 0$, $\mathscr{I}_{-1} = \emptyset$. Here $k$ is the iteration counter and $\mathscr{I}_k$ is the accumulated *informational set.*

**Main loop:**

1. Call oracle $\mathscr{O}$ at point $x_k$.
2. Update the informational set: $\mathscr{I}_k = \mathscr{I}_{k-1} \bigcup (x_k, \mathscr{O}(x_k))$.
3. Apply the rules of method $\mathscr{M}$ to $\mathscr{I}_k$ and generate a new point $x_{k+1}$.
4. Check criterion $\mathscr{T}_\epsilon$. If **yes** then form an output $\bar{x}$. Otherwise set $k := k + 1$ and go to Step 1.

(1.1.3)

Now we can specify the meaning of *computational effort* in our definition of performance. In the scheme (1.1.3), we can see two potentially expensive steps. The first one is Step 1, where we call the oracle. The second one is Step 3, where we form the new test point. Thus, we can introduce two measures of *complexity* of problem $P$ for method $\mathscr{M}$:

*Analytical complexity:* The number of calls of the oracle which is necessary to solve problem $P$ up to accuracy $\epsilon$.
*Arithmetical complexity:* The total number of arithmetic operations (including the work of oracle and work of method), which is necessary for solving problem $P$ up to accuracy $\epsilon$.

Comparing the notions of analytical and arithmetical complexity, we can see that the second one is more realistic. However usually, for a particular method $\mathscr{M}$ as applied to problem $P$, arithmetical complexity can be easily obtained from the analytical complexity and complexity of the oracle. Therefore, in Part I of this course we speak mainly about bounds on the analytical complexity for some problem classes. Arithmetical complexity will be treated in Part II, where we consider methods of Structural Optimization.

There is one standard assumption on the oracle which allows us to obtain the majority of results on analytical complexity for optimization schemes. This assumption, called the *Local Black Box Concept*, is as follows.

**Local Black Box**

1. The only information available for the numerical scheme is the answer of the oracle.
2. The oracle is *local*: A small variation of the problem far enough from the test point $x$, which is compatible with the description of the problem class, does not change the answer at $x$.

This concept is very useful in the complexity analysis. Of course, its first part looks like an artificial wall between the method and the oracle. It seems natural to give methods full access to the internal structure of the problem. However, we will see that for problems with a complicated or implicit structure this access is almost useless. For more simple problems it could help. We will see this in the second part of this book.

To conclude the section, let us mention that the standard formulation (1.1.1) is called a *functional model* of optimization problems. Usually, for such models the standard assumptions are related to the level of smoothness of functional components. According to the degree of smoothness we can apply different types of oracle:

- *Zero-order* oracle: returns the function value $f(x)$.
- *First-order* oracle: returns the function value $f(x)$ and the gradient $\nabla f(x)$.
- *Second-order* oracle: returns $f(x)$, $\nabla f(x)$, and the Hessian $\nabla^2 f(x)$.

### *1.1.3 Complexity Bounds for Global Optimization*

Let us try to apply the formal language of the previous section to a particular problem class. Consider the following problem:

$$\min_{x \in B_n} f(x). \tag{1.1.4}$$

In our terminology, this is a constrained minimization problem with no functional constraints. The basic feasible set of this problem is $B_n$, an $n$-dimensional box in $\mathbb{R}^n$:

$$B_n = \{x \in \mathbb{R}^n \mid 0 \leq x^{(i)} \leq 1, \; i = 1 \dots n\}.$$

Let us measure distances in $\mathbb{R}^n$ by the $\ell_\infty$-norm:

$$\| x \|_{(\infty)} = \max_{1 \leq i \leq n} | x^{(i)} |.$$

Assume that, with respect to this norm,

> the objective function $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ is *Lipschitz continuous* on $B_n$:
>
> $$| f(x) - f(y) | \leq L \| x - y \|_{(\infty)} \quad \forall x, y \in B_n,$$
>
> with some constant $L$ (*Lipschitz constant*).

(1.1.5)

Let us consider a very simple method for solving (1.1.4), which is called the *Uniform Grid Method*. This method $\mathscr{G}(p)$ has one integer input parameter $p \geq 1$.

**Method** $\mathscr{G}(p)$

**1.** Form $p^n$ points

$$x_\alpha = \left(\tfrac{2i_1-1}{2p}, \tfrac{2i_2-1}{2p}, \dots, \tfrac{2i_n-1}{2p}\right)^T,$$

where $\alpha \equiv (i_1, \dots, i_n) \in \{1, \dots, p\}^n$.

**2.** Among all points $x_\alpha$, find the point $\bar{x}$ with the minimal value of the objective function.

**3.** The pair $(\bar{x}, f(\bar{x}))$ is the output of the method.

(1.1.6)

Thus, this method forms a uniform grid of the test points inside the box $B_n$, computes the best value of the objective over this grid, and returns this value as an approximate solution to problem (1.1.4). In our terminology, this is a zero-order iterative method without any influence from the accumulated information on the sequence of test points. Let us find its efficiency estimate.

**Theorem 1.1.1** *Let $f^*$ be a global optimal value of problem (1.1.4). Then*

$$f(\bar{x}) - f^* \le \tfrac{L}{2p}.$$

*Proof* For a multi-index $\alpha = (i_1, \dots, i_n)$, define

$$X_\alpha = \{x \in \mathbb{R}^n : \|x - x_\alpha\|_{(\infty)} \le \tfrac{1}{2p}\}.$$

Clearly, $\bigcup\limits_{\alpha \in \{1,\dots,p\}^n} X_\alpha = B_n$.

Let $x_*$ be a global solution of our problem. Then there exists a multi-index $\alpha^*$ such that $x^* \in X_{\alpha^*}$. Note that $\|x^* - x_{\alpha^*}\|_{(\infty)} \le \frac{1}{2p}$. Therefore,

$$f(\bar{x}) - f(x^*) \le f(x_{\alpha^*}) - f(x^*) \overset{(1.1.5)}{\le} \tfrac{L}{2p}. \qquad \square$$

Let us conclude with the definition of our problem class. We fix our goal as follows:

$$\text{Find } \bar{x} \in B_n : \quad f(\bar{x}) - f^* \leq \epsilon. \tag{1.1.7}$$

Then we immediately get the following result.

**Corollary 1.1.1** *The analytical complexity of problem class (1.1.4), (1.1.5), (1.1.7) for method $\mathscr{G}$ is at most*

$$\mathscr{A}(\mathscr{G}) = \left( \left\lfloor \tfrac{L}{2\epsilon} \right\rfloor + 1 \right)^n,$$

(here and in the sequel, $\lfloor a \rfloor$ is the integer part of $a \in \mathbb{R}$).

*Proof* Take $p = \left\lfloor \frac{L}{2\epsilon} \right\rfloor + 1$. Then $p \geq \frac{L}{2\epsilon}$, and, in view of Theorem 1.1.1, we have $f(\bar{x}) - f^* \leq \frac{L}{2p} \leq \epsilon$. Note that we need to call the oracle at $p^n$ points. □

Thus, $\mathscr{A}(\mathscr{G})$ justifies an *upper* complexity bound for our problem class.

This result is quite informative. However, we still have some questions. Firstly, it may happen that our proof is too rough and the real performance of method $\mathscr{G}(p)$ is much better. Secondly, we still cannot be sure that $\mathscr{G}(p)$ is a reasonable method for solving (1.1.4). There could exist other schemes with much higher performance.

In order to answer these questions, we need to derive *lower complexity bounds* for the problem class (1.1.4), (1.1.5), (1.1.7). The main features of such bounds are as follows.

- They are based on the *Black Box Concept*.
- These bounds are valid for all reasonable iterative schemes. Thus, they provide us with a lower estimate for the *analytical complexity* of the problem class.
- Very often such bounds employ the idea of a *resisting* oracle.

For us, only the concept of a resisting oracle is new. Therefore, let us present it in more detail.

A resisting oracle tries to create the *worst possible* problem for each particular method. It starts from an "empty" function and it tries to answer each call of the method in the worst possible way. However, the answers must be *compatible* with the previous answers and with description of the problem class. Then, after termination of the method it is possible to *reconstruct* a problem which perfectly fits the final informational set accumulated by the algorithm. Moreover, if we run the method on this newborn problem, it will reproduce the same sequence of test points since it will have the same sequence of answers from the oracle.

Let us show how this works for problem (1.1.4). Consider the class of problems $\mathscr{P}_\infty$ defined as follows.

| | |
|---|---|
| **Model** : | $\min\limits_{x \in B_n} f(x)$, where $f(\cdot)$ is $\ell_\infty$-Lipschitz continuous on $B_n$. |
| **Oracle** : | Zero-order Local Black Box. |
| **Approximate solution** : | Find $\bar{x} \in B_n : f(\bar{x}) - f^* < \epsilon$. |

**Theorem 1.1.2** *For $\epsilon < \frac{1}{2}L$, the analytical complexity of problem class $\mathscr{P}_\infty$ is at least $\left\lfloor \frac{L}{2\epsilon} \right\rfloor^n$ calls of the oracle.*

*Proof* Let $p = \left\lfloor \frac{L}{2\epsilon} \right\rfloor$ $(\geq 1)$. Assume that there exists a method which needs $N < p^n$ calls of oracle to solve any problem from $\mathscr{P}$. Let us apply this method to the following resisting strategy:

$$\boxed{\text{Return } f(x) = 0 \text{ at any test point } x.}$$

Therefore this method can find only $\bar{x} \in B_n$ with $f(\bar{x}) = 0$.

However, since $N < p^n$, there exists a multi-index $\hat{\alpha}$ such that there were no test points in the box $X_{\hat{\alpha}}$ (see the notation of Theorem 1.1.1). Define $x_* = x_{\hat{\alpha}}$, and consider the function

$$\bar{f}(x) = \min\{0, L\|x - x_*\|_{(\infty)} - \epsilon\}.$$

Clearly, this function is $\ell_\infty$-Lipschitz continuous with constant $L$, and its global optimal value is $-\epsilon$. Moreover, $\bar{f}(\cdot)$ differs from zero only inside the box $X_{\hat{\alpha}}$. Thus, $\bar{f}(\cdot)$ is equal to zero *at all test points* of our method.

Since the accuracy of the output of our method is $\epsilon$, we come to the following conclusion:

> If the number of calls of the oracle is less than $p^n$,then the accuracy of the result cannot be better than $\epsilon$.

Thus, the desired statement is proved. □

Now we can say much more about the performance of the Uniform Grid Method. Let us compare its efficiency estimate with the lower bound:

$$\mathscr{G} : \left( \left\lfloor \frac{L}{2\epsilon} \right\rfloor + 1 \right)^n \quad \Leftrightarrow \quad \text{Lower bound: } \left\lfloor \frac{L}{2\epsilon} \right\rfloor^n .$$

If $\epsilon \leq O(\frac{L}{n})$, then the lower and upper bounds coincide up to an absolute constant multiplicative factor. This means that, for such level of accuracy, $\mathscr{G}(\cdot)$ is *optimal* for the problem class $\mathscr{P}_\infty$.

At the same time, Theorem 1.1.2 supports our initial claim that the general optimization problems are unsolvable. Let us look at the following illustrative example.

*Example 1.1.4* Consider the problem class $\mathscr{P}_\infty$ defined by the following parameters:

$$L = 2, \quad n = 10, \quad \epsilon = 0.01.$$

Note that the size of these problems is very small and we ask only for a moderate $1\%$ accuracy.

The lower complexity bound for this class is $\left\lfloor \frac{L}{2\epsilon} \right\rfloor^n$ calls of the oracle. Let us compute this value for our example.

| | |
|---|---|
| **Lower bound** : | $10^{20}$ calls of the oracle |
| **Oracle complexity** : | at least $n$ arithmetic operations (a.o.) |
| **Total complexity** : | $10^{21}$ a.o. |
| **Processor performance** : | $10^6$ a.o. per second |
| **Total time** : | $10^{15}$s |
| **One year** : | less than $3.2 \cdot 10^7$s |
| **We need** : | 31,250,000 years |

This estimate is so disappointing that we cannot maintain any hope that such problems may become solvable in the future. Let us just play with the parameters of the problem class.

- If we change $n$ to $n + 1$, then the estimate is multiplied by one hundred. Thus, for $n = 11$ our lower bound is valid for a much more powerful computer.
- On the contrary, if we multiply $\epsilon$ by two, we reduce the complexity by a factor of a thousand. For example, if $\epsilon = 8\%$, then we need only two weeks.[2] □

[2] We keep this calculation unchanged from the first version of this book [39]. In this example, the processor performance corresponds to a Sun Station, which was the most powerful personal computer at the beginning of the 1990s. Now, after twenty five years of intensive progress in the abilities of hardware, modern personal computers have reached a speed level of $10^8$ a.o. per second. Thus indeed, our time estimate remains valid for $n = 11$.

We should note that the lower complexity bounds for problems with smooth functions, or for high-order methods, are not much better than the bound of Theorem 1.1.2. This can be proved using the same arguments and we leave the proof as an exercise for the reader. Comparison of the above results with the *upper* bounds for NP-hard problems, which are considered as classical examples of very difficult problems in Combinatorial Optimization, is also quite disappointing. To find the exact solution, the hardest combinatorial problems need only $2^n$ a.o. !

To conclude this section, let us compare our observations with some other fields of Numerical Analysis. It is well known that the uniform grid approach is a standard tool in many domains. For example, if we need to compute numerically the value of the integral of a univariate function

$$\mathscr{S} = \int\limits_0^1 f(x)dx,$$

the standard way to proceed is to form a discrete sum

$$S_N = \tfrac{1}{N}\sum_{i=1}^{n} f(x_i), \quad x_i = \tfrac{i}{N},\ i = 1\ldots N.$$

If $f(\cdot)$ is Lipschitz continuous, then this value is a good approximation to $\mathscr{I}$:

$$N = L/\epsilon \quad \Rightarrow \quad | \mathscr{S} - S_N | \leq \epsilon.$$

Note that in our terminology this is exactly a uniform grid approach. Moreover, this is a standard way for approximating integrals. The reason why it works here is related to the *dimension* of the problem. For integration, the standard dimensions are very small (up to three). However, in Optimization, sometimes we need to solve problems with several million variables.

### *1.1.4 Identity Cards of the Fields*

After the pessimistic results of the previous section, we should try to find a reasonable target in the theoretical analysis of optimization schemes. It seems that everything is clear with general Global Optimization. However, maybe the goals of this field are too ambitious? In some practical problems could we be satisfied by much less "optimal" solutions? Or, are there some interesting problem classes which are not as dangerous as the class of general continuous functions?

In fact, each of these questions can be answered in different ways, each of which define the style of research (or rules of the game) in different fields of Nonlinear

Optimization. If we try to classify these fields, we can easily see that they differ one from another in the following aspects:

- Goals of the methods.
- Classes of functional components.
- Description of the oracle.

These aspects naturally define the list of desired properties of the optimization methods. Let us present the "identity cards" of the fields which we are going to consider in this book.

1. General Global Optimization (Sect. 1.1)

- **Goals:** Find a global minimum.
- **Functional class:** Continuous functions.
- **Oracle:** 0–1–2 order Black Box.
- **Desired properties:** Convergence to a global minimum.
- **Features:** From theoretical point of view, this game is too short.
- **Problem sizes:** Sometimes, we can solve problems with many variables. No guarantee of success even for small problems.
- **History** Starts from 1955. Several local peaks of interest related to new heuristic ideas (simulated annealing, genetic algorithms).

2. General Nonlinear Optimization (Sects. 1.2, 1.3)

- **Goals:** Find a local minimum.
- **Functional class:** Differentiable functions.
- **Oracle:** First- and second-order Black Box.
- **Desired properties:** Fast convergence to a local minimum.
- **Features:** Variability of approaches. Most widespread software. The goals are not always acceptable and reachable.
- **Problem sizes:** Up to several thousand variables.
- **History:** Starts from 1955. Peak period: 1965 – 1985. Theoretical activity now is rather low.

3. Black Box Convex Optimization (Chaps. 2, 3, and 4)

- **Goals:** Find a global minimum.
- **Functional class:** Convex sets and functions.
- **Oracle:** First- and second-order Black Box.
- **Desired properties:** Convergence to a global minimum. The rate of convergence may depend on dimension.
- **Features:** Very interesting and rich complexity theory. Efficient practical methods. The problem class is sometimes restrictive.
- **Problem sizes:** Several thousand variables for the second-order methods, and several million for the first-order schemes.
- **History:** Starts from 1970. Peak period: 1975–1985. Theoretical activity now is high due to the interest to Structural Optimization and global complexity analysis of second-order methods (2006).

4. Structural Optimization (Part II)

- **Goals:** Find a global minimum.
- **Functional class:** Simple convex sets and functions with explicit minimax structure.
- **Oracle:** Second-order Black Box for special barrier functions (Chap. 5), and modified first-order Black Box (Chaps. 6, 7).
- **Desired properties:** Fast convergence to a global minimum. The rate of convergence depends on the structure of the problem.
- **Features:** Very new and perspective theory rejecting the Black Box Concept. The problem class is practically the same as in Convex Optimization.
- **Problem sizes:** Sometimes up to several million variables.
- **History:** Starts from 1984. Peak period: 1990–2000 for Interior-Point Methods. The first accelerated first-order method for problems with explicit structure was developed in 2005. Very high theoretical activity right now.

## 1.2 Local Methods in Unconstrained Minimization

(Relaxation and approximation; Necessary optimality conditions; Sufficient optimality conditions; The class of differentiable functions; The class of twice differentiable functions; The Gradient Method; Rate of convergence; Newton's Method.)

### 1.2.1 Relaxation and Approximation

The simplest goal in general Nonlinear Optimization consists in finding a local minimum of a differentiable function. However, even to reach such a restricted goal, it is necessary to follow some special principles which guarantee convergence of the minimization process.

The majority of methods in general Nonlinear Optimization are based on the idea of *relaxation.*

A sequence of real numbers $\{a_k\}_{k=0}^{\infty}$ is called a *relaxation sequence* if

$$a_{k+1} \le a_k \quad \forall k \ge 0.$$

In this section we consider several methods for solving the following unconstrained minimization problem:

$$\min_{x \in \mathbb{R}^n} f(x), \tag{1.2.1}$$

where $f(\cdot)$ is a smooth function. In order to do so, these methods generate a relaxation sequence of function values $\{f(x_k)\}_{k=0}^{\infty}$:

$$f(x_{k+1}) \le f(x_k), \quad k = 0, 1, \ldots \quad .$$

This rule has the following important advantages.

1. If $f(\cdot)$ is bounded below on $\mathbb{R}^n$, then the sequence $\{f(x_k)\}_{k=0}^{\infty}$ converges.
2. In any case, we improve the initial value of the objective function.

However, it is impossible to implement the idea of relaxation without employing another fundamental element of Numerical Analysis, *approximation*. In general,

To approximate means to replace an initial complex object by a simpler one which is close to the original in terms of its properties.

In Nonlinear Optimization, we usually apply *local approximations* based on derivatives of nonlinear functions. These are the first- and second-order approximations (or, the linear and quadratic approximations).

Let the function $f(\cdot)$ be differentiable at $\bar{x} \in \mathbb{R}^n$. Then, for any $y \in \mathbb{R}^n$ we have

$$f(y) = f(\bar{x}) + \langle \nabla f(\bar{x}), y - \bar{x} \rangle + o(\| y - \bar{x} \|),$$

where $o(\cdot) : [0, \infty) \to \mathbb{R}$ is a function of $r \geq 0$ satisfying the conditions

$$\lim_{r \downarrow 0} \tfrac{1}{r} o(r) = 0, \quad o(0) = 0.$$

In the remaining part of this chapter, unless stated otherwise, we use the notation $\| \cdot \|$ for the standard *Euclidean* norm in $\mathbb{R}^n$:

$$\|x\| = \left[ \sum_{i=1}^{n} \left(x^{(i)}\right)^2 \right]^{1/2} = (x^T x)^{1/2} = \langle x, x \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product in the corresponding coordinate space. Note that for any $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, and matrix $A \in \mathbb{R}^{m \times n}$ we have

$$\langle Ax, y \rangle \equiv \langle x, A^T y \rangle. \tag{1.2.2}$$

The linear function $f(\bar{x}) + \langle \nabla f(\bar{x}), y - \bar{x} \rangle$ is called the *linear approximation* of $f$ at $\bar{x}$. Recall that the vector $\nabla f(\bar{x})$ is called the *gradient* of function $f$ at $\bar{x}$. Considering the points $y_i = \bar{x} + \epsilon e_i$, where $e_i$ is the $i$th coordinate vector in $\mathbb{R}^n$, and taking the limit as $\epsilon \to 0$, we obtain the following coordinate representation of the gradient:

$$\nabla f(\bar{x}) = \left( \tfrac{\partial f(\bar{x})}{\partial x^{(1)}}, \dots, \tfrac{\partial f(\bar{x})}{\partial x^{(n)}} \right)^T. \tag{1.2.3}$$

Let us mention two important properties of the gradient. Denote by $\mathscr{L}_f(\alpha)$ the *(sub)level set* of $f(\cdot)$:

$$\mathscr{L}_f(\alpha) = \{ x \in \mathbb{R}^n \mid f(x) \leq \alpha \}.$$

Consider the set of directions that are *tangent* to $\mathscr{L}_f(f(\bar{x}))$ at $\bar{x}$:

$$S_f(\bar{x}) = \left\{ s \in \mathbb{R}^n \mid s = \lim_{k \to \infty} \tfrac{y_k - \bar{x}}{\|y_k - \bar{x}\|}, \text{ for some } \{y_k\} \to \bar{x} \text{ with } f(y_k) = f(\bar{x}) \; \forall k \right\}.$$

**Lemma 1.2.1** *If $s \in S_f(\bar{x})$, then $\langle \nabla f(\bar{x}), s \rangle = 0$.*

*Proof* Since $f(y_k) = f(\bar{x})$, we have

$$f(y_k) = f(\bar{x}) + \langle \nabla f(\bar{x}), y_k - \bar{x} \rangle + o(\| y_k - \bar{x} \|) = f(\bar{x}).$$

Therefore $\langle \nabla f(\bar{x}), y_k - \bar{x} \rangle + o(\| y_k - \bar{x} \|) = 0$. Dividing this equation by $\| y_k - \bar{x} \|$ and taking the limit as $y_k \to \bar{x}$, we obtain the result. □

Let $s$ be a direction in $\mathbb{R}^n$, $\| \ s \ \| = 1$. Consider the local decrease of the function $f(\cdot)$ along direction $s$:

$$\Delta(s) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha}[f(\bar{x} + \alpha s) - f(\bar{x})].$$

Note that $f(\bar{x} + \alpha s) - f(\bar{x}) = \alpha \langle \nabla f(\bar{x}), s \rangle + o(\alpha)$. Therefore

$$\Delta(s) = \langle \nabla f(\bar{x}), s \rangle.$$

Using the Cauchy–Schwarz inequality,

$$- \| \ x \ \| \cdot \| \ y \ \| \leq \langle x, y \rangle \ \ \leq \| \ x \ \| \cdot \| \ y \ \|,$$

we obtain $\Delta(s) = \langle \nabla f(\bar{x}), s \rangle \geq - \| \ \nabla f(\bar{x}) \ \|$. Let us take

$$\bar{s} = -\nabla f(\bar{x}) / \| \ \nabla f(\bar{x}) \ \| \ .$$

Then

$$\Delta(\bar{s}) = -\langle \nabla f(\bar{x}), \nabla f(\bar{x}) \rangle / \| \ \nabla f(\bar{x}) \ \| = - \| \ \nabla f(\bar{x}) \ \| \ .$$

Thus, the direction $-\nabla f(\bar{x})$ (the *antigradient*) is the direction of the *fastest local decrease* of the function $f(\cdot)$ at point $\bar{x}$.

The next statement is probably the most fundamental fact in Optimization Theory.

**Theorem 1.2.1 (First-Order Optimality Condition)** *Let $x^*$ be a local minimum of a differentiable function $f(\cdot)$. Then*

$$\nabla f(x^*) = 0. \tag{1.2.4}$$

*Proof* Since $x^*$ is a local minimum of $f(\cdot)$, there exists an $r > 0$ such that for all $y \in \mathbb{R}^n$, $\|y - x^*\| \leq r$, we have $f(y) \geq f(x^*)$. Since $f$ is differentiable, this implies that

$$f(y) = f(x^*) + \langle \nabla f(x^*), y - x^* \rangle + o(\| \ y - x^* \ \|) \geq f(x^*).$$

Thus, for all $s \in \mathbb{R}^n$, we have $\langle \nabla f(x^*), s \rangle \geq 0$. By taking $s = -\nabla f(x^*)$, we get $-\|\nabla f(x^*)\|^2 \geq 0$. Hence, $\nabla f(x^*) = 0$. □

In what follows the notation $B \succeq 0$, where $B$ is a symmetric $(n \times n)$-matrix, means that $B$ is *positive semidefinite*:

$$\langle Bx, x \rangle \geq 0 \quad \forall x \in \mathbb{R}^n.$$

The notation $B \succ 0$ means that $B$ is *positive definite* (in this case, the inequality above must be strict for all $x \neq 0$).

**Corollary 1.2.1** *Let $x^*$ be a local minimum of a differentiable function $f(\cdot)$ subject to the linear equality constraints*

$$x \in \mathscr{L} \equiv \{x \in \mathbb{R}^n |\ Ax = b\} \neq \emptyset,$$

*where $A$ is an $m \times n$-matrix with full row rank, and $b \in \mathbb{R}^m$, $m < n$. Then there exists a vector of multipliers $\lambda^* \in R^m$ such that*

$$\nabla f(x^*) = A^T \lambda^*. \tag{1.2.5}$$

*Proof* Let us assume that $\nabla f(x^*) \neq 0$. Consider the following optimization problem:

$$g^* = \min_{\lambda \in \mathbb{R}^m} \left\{ g(\lambda) = \tfrac{1}{2} \|\nabla f(x^*) - A^T \lambda\|^2 \right\}. \tag{1.2.6}$$

Assume that $g^* > 0$. Note that

$$g(\lambda) = \tfrac{1}{2} \|\nabla f(x^*)\|^2 - \langle \nabla f(x^*), A^T \lambda \rangle + \tfrac{1}{2} \langle B\lambda, \lambda \rangle,$$

where $B = AA^T \succeq \lambda_{\min}(B) I_n$ and $\lambda_{\min}(B) > 0$ denotes the smallest eigenvalue of matrix $B$. Hence, the level sets of this function are bounded, and therefore the problem (1.2.6) has a solution $\lambda^*$ satisfying the first-order optimality condition:

$$0 \overset{(1.2.4)}{=} \nabla g(\lambda^*) \ = B\lambda^* - A \nabla f(x^*).$$

Thus, $\lambda^* = B^{-1} A \nabla f(x^*)$. Let $s^* = (I_n - A^T B^{-1} A) \nabla f(x^*)$. Note that $As^* = 0$. Then,

$$\langle \nabla f(x^*), s^* \rangle = \|\nabla f(x^*)\|^2 - \langle B^{-1} A \nabla f(x^*), A \nabla f(x^*) \rangle \ = \ 2g^* \ > \ 0.$$

Therefore, the optimal value of the function $g$ can be reduced along the ray $\{x^* - \alpha s^* : \ \alpha \geq 0\}$. This contradiction proves that $g^* = 0$. □

Note that we have proved only a *necessary* condition for a local minimum. The points satisfying this condition are called the *stationary points* of the function $f$. In order to see that such points are not always local minima, it is enough to look at the function $f(x) = x^3$, $x \in \mathbb{R}$, at the point $x = 0$.

Now let us introduce second-order approximation. Let the function $f(\cdot)$ be twice differentiable at $\bar{x}$. Then

$$f(y) = f(\bar{x}) + \langle \nabla f(\bar{x}), y - \bar{x} \rangle + \frac{1}{2} \langle \nabla^2 f(\bar{x})(y - \bar{x}), y - \bar{x} \rangle + o(\| \ y - \bar{x} \ \|^2).$$

The quadratic function

$$f(\bar{x}) + \langle \nabla f(\bar{x}), y - \bar{x} \rangle + \frac{1}{2} \langle \nabla^2 f(\bar{x})(y - \bar{x}), y - \bar{x} \rangle$$

is called the *quadratic* (or *second-order)* approximation of the function $f$ at $\bar{x}$. Recall that $\nabla^2 f(\bar{x})$ is an $(n \times n)$-matrix with the following entries:

$$(\nabla^2 f(\bar{x}))^{(i,j)} = \frac{\partial^2 f(\bar{x})}{\partial x^{(i)} \partial x^{(j)}}, \quad i, j = 1, \ldots, n.$$

It is called the *Hessian* of function $f$ at $\bar{x}$. Note that the Hessian is a symmetric matrix:

$$\nabla^2 f(\bar{x}) = \left[ \nabla^2 f(\bar{x}) \right]^T.$$

The Hessian can be regarded as a derivative of the vector function $\nabla f(\cdot)$:

$$\nabla f(y) = \nabla f(\bar{x}) + \nabla^2 f(\bar{x})(y - \bar{x}) + \mathbf{o}(\| y - \bar{x} \|) \in \mathbb{R}^n, \tag{1.2.7}$$

where $\mathbf{o}(\cdot) : [0, \infty) \to \mathbb{R}^n$ is a continuous vector function satisfying the condition

$$\lim_{r \downarrow 0} \frac{1}{r} \| \mathbf{o}(r) \| = 0.$$

Using the second-order approximation, we can write down the second-order optimality conditions.

**Theorem 1.2.2 (Second-Order Optimality Condition)** *Let $x^*$ be a local minimum of a twice differentiable function $f(\cdot)$. Then*

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

*Proof* Since $x^*$ is a local minimum of the function $f(\cdot)$, there exists an $r > 0$ such that for all $y$, $\|y - x^*\| \leq r$, we have

$$f(y) \geq f(x^*).$$

In view of Theorem 1.2.1, $\nabla f(x^*) = 0$. Therefore, for any such $y$,

$$f(y) = f(x^*) + \langle \nabla^2 f(x^*)(y - x^*), y - x^* \rangle + o(\| y - x^* \|^2) \geq f(x^*).$$

Thus, $\langle \nabla^2 f(x^*)s, s \rangle \geq 0$, for all $s$, $\| s \| = 1$. □

Again, the above theorem is a *necessary* (second-order) characteristic of a local minimum. Let us prove now a sufficient condition.

**Theorem 1.2.3** *Let a function $f(\cdot)$ be twice differentiable on $\mathbb{R}^n$ and let $x^* \in \mathbb{R}^n$ satisfy the following conditions:*

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

*Then $x^*$ is a strict local minimum of $f(\cdot)$.*

*Proof* Note that in a small neighborhood of a point $x^*$ the function $f(\cdot)$ can be represented as

$$f(y) = f(x^*) + \frac{1}{2}\langle \nabla^2 f(x^*)(y - x^*), y - x^* \rangle + o(\| y - x^* \|^2).$$

Since $\frac{o(r^2)}{r^2} \to 0$ as $r \downarrow 0$, there exists a value $\bar{r} > 0$ such that for all $r \in [0, \bar{r}]$ we have

$$| o(r^2) | \le \tfrac{r^2}{4} \lambda_{\min}(\nabla^2 f(x^*)).$$

In view of our assumption, this eigenvalue is positive. Therefore, for any $y \in \mathbb{R}^n$, $0 <\| y - x^* \| \le \bar{r}$, we have

$$\begin{aligned} f(y) &\ge f(x^*) + \tfrac{1}{2}\lambda_{\min}(\nabla^2 f(x^*)) \| y - x^* \|^2 + o(\| y - x^* \|^2) \\ &\ge f(x^*) + \tfrac{1}{4}\lambda_{\min}(\nabla^2 f(x^*)) \| y - x^* \|^2 > f(x^*). \quad \square \end{aligned}$$

## 1.2.2 Classes of Differentiable Functions

It is well known that any continuous function can be approximated by a smooth function with arbitrarily small accuracy. Therefore, assuming only differentiability of the objective function, we cannot ensure any reasonable properties of minimization processes. For that, we need to impose some additional assumptions on the magnitude of some derivatives. Traditionally, in Optimization such assumptions are presented in the form of a *Lipschitz condition* for a derivative of certain degree.

Let $Q$ be a subset of $\mathbb{R}^n$. We denote by $C_L^{k,p}(Q)$ the class of functions with the following properties:

- any $f \in C_L^{k,p}(Q)$ is $k$ times continuously differentiable on $Q$.
- Its $p$th derivative is Lipschitz continuous on $Q$ with constant $L$:

$$\|\nabla^p f(x) - \nabla^p f(y)\| \le L\|x - y\|$$

for all $x, y \in Q$. In this book, we usually work with $p = 1$ and $p = 2$.

Clearly, we always have $p \leq k$. If $q \geq k$, then $C_L^{q,p}(Q) \subseteq C_L^{k,p}(Q)$. For example, $C_L^{2,1}(Q) \subseteq C_L^{1,1}(Q)$. Note also that these classes possess the following property:

*If* $f_1 \in C_{L_1}^{k,p}(Q)$, $f_2 \in C_{L_2}^{k,p}(Q)$ and $\alpha_1, \alpha_2 \in \mathbb{R}$, *then for*

$$L_3 = \mid \alpha_1 \mid L_1 + \mid \alpha_2 \mid L_2$$

*we have* $\alpha_1 f_1 + \alpha_2 f_2 \in C_{L_3}^{k,p}(Q)$.

We use the notation $f \in C^k(Q)$ for function $f$ which is $k$ times continuously differentiable on $Q$.

One of the most important classes of differentiable functions is $C_L^{1,1}(\mathbb{R}^n)$, the class of functions with Lipschitz continuous gradient. By definition the inclusion $f \in C_L^{1,1}(\mathbb{R}^n)$ means that

$$\| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \| \tag{1.2.8}$$

for all $x, y \in \mathbb{R}^n$. Let us give a sufficient condition for this inclusion.

**Lemma 1.2.2** *A function* $f(\cdot)$ *belongs to the class* $C_L^{2,1}(\mathbb{R}^n) \subset C_L^{1,1}(\mathbb{R}^n)$ *if and only if for all* $x \in \mathbb{R}^n$ *we have*

$$\| \nabla^2 f(x) \| \leq L. \tag{1.2.9}$$

*Proof* Indeed, for any $x, y \in \mathbb{R}^n$ we have

$$\begin{aligned} \nabla f(y) &= \nabla f(x) + \int_0^1 \nabla^2 f(x + \tau(y-x))(y-x) d\tau \\ &= \nabla f(x) + \left( \int_0^1 \nabla^2 f(x + \tau(y-x)) d\tau \right) \cdot (y-x). \end{aligned}$$

Therefore, if condition (1.2.9) is satisfied, then

$$\begin{aligned} \| \nabla f(y) - \nabla f(x) \| &= \left\| \left( \int_0^1 \nabla^2 f(x + \tau(y-x)) d\tau \right) \cdot (y-x) \right\| \\ &\leq \left\| \int_0^1 \nabla^2 f(x + \tau(y-x)) d\tau \right\| \cdot \| y - x \| \\ &\leq \int_0^1 \| \nabla^2 f(x + \tau(y-x)) \| d\tau \cdot \| y - x \| \\ &\leq L \| y - x \| . \end{aligned}$$

On the other hand, if $f \in C_L^{2,1}(\mathbb{R}^n)$, then for any $s \in \mathbb{R}^n$ and $\alpha > 0$, we have

$$\left\| \left( \int\limits_0^\alpha \nabla^2 f(x+\tau s) d\tau \right) \cdot s \right\| = \| \nabla f(x+\alpha s) - \nabla f(x) \| \leq \alpha L \| s \| .$$

Dividing this inequality by $\alpha$ and taking $\alpha \downarrow 0$, we obtain (1.2.9). □

Note that the condition (1.2.9) can be written in the form of a matrix inequality:

$$-LI_n \preceq \nabla^2 f(x) \preceq LI_n, \quad \forall x \in \mathbb{R}^n. \tag{1.2.10}$$

Lemma 1.2.2 provides us with many examples of functions with Lipschitz continuous gradient.

*Example 1.2.1*

1. The linear function $f(x) = \alpha + \langle a, x \rangle \in C_0^{1,1}(\mathbb{R}^n)$ since

$$\nabla f(x) = a, \quad \nabla^2 f(x) = 0.$$

2. For a quadratic function $f(x) = \alpha + \langle a, x \rangle + \frac{1}{2}\langle Ax, x \rangle$ with $A = A^T$, we have

$$\nabla f(x) = a + Ax, \quad \nabla^2 f(x) = A.$$

Therefore $f(\cdot) \in C_L^{1,1}(\mathbb{R}^n)$ with $L = \| A \|$.

3. Consider the function of one variable $f(x) = \sqrt{1+x^2}$, $x \in \mathbb{R}$. We have

$$\nabla f(x) = \frac{x}{\sqrt{1+x^2}}, \quad \nabla^2 f(x) = \frac{1}{(1+x^2)^{3/2}} \leq 1.$$

Therefore, $f(\cdot) \in C_1^{1,1}(\mathbb{R})$. □

The next statement is important for the geometric interpretation of functions in $C_L^{1,1}(\mathbb{R}^n)$.

**Lemma 1.2.3** *Let $f \in C_L^{1,1}(\mathbb{R}^n)$. Then, for any $x, y$ from $\mathbb{R}^n$, we have*

$$| f(y) - f(x) - \langle \nabla f(x), y - x \rangle | \leq \frac{L}{2} \| y - x \|^2 . \tag{1.2.11}$$

*Proof* For all $x, y \in \mathbb{R}^n$, we have

$$\begin{aligned} f(y) &= f(x) + \int\limits_0^1 \langle \nabla f(x + \tau(y-x)), y - x \rangle d\tau \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \int\limits_0^1 \langle \nabla f(x+\tau(y-x)) - \nabla f(x), y - x \rangle d\tau. \end{aligned}$$

Therefore,

$$\begin{aligned}
&| \ f(y) - f(x) - \langle \nabla f(x), y - x \rangle \ | \\
&= | \int\limits_0^1 \langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x \rangle d\tau \ | \\
&\leq \int\limits_0^1 | \ \langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x \rangle \ | \ d\tau \\
&\leq \int\limits_0^1 \| \ \nabla f(x + \tau(y - x)) - \nabla f(x) \ \| \cdot \| \ y - x \ \| \ d\tau \\
&\leq \int\limits_0^1 \tau L \ \| \ y - x \ \|^2 \ d\tau = \tfrac{L}{2} \ \| \ y - x \ \|^2 \ . \quad \Box
\end{aligned}$$

Geometrically, we have the following picture. Consider a function $f \in C_L^{1,1}(\mathbb{R}^n)$. Let us fix a point $x_0 \in \mathbb{R}^n$, and define two quadratic functions

$$\begin{aligned}
\phi_1(x) &= f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \tfrac{L}{2} \ \| \ x - x_0 \ \|^2, \\
\phi_2(x) &= f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \tfrac{L}{2} \ \| \ x - x_0 \ \|^2 \ .
\end{aligned}$$

Then the graph of the function $f$ lies between the graphs of $\phi_1$ and $\phi_2$:

$$\phi_1(x) \leq f(x) \leq \phi_2(x), \quad \forall x \in \mathbb{R}^n.$$

Let us prove similar results for the class of twice differentiable functions. The main class of functions of this type is $C_M^{2,2}(\mathbb{R}^n)$, the class of twice differentiable functions with Lipschitz continuous Hessian. Recall that for $f \in C_M^{2,2}(\mathbb{R}^n)$, we have

$$\| \ \nabla^2 f(x) - \nabla^2 f(y) \ \| \leq \ M \ \| \ x - y \ \|, \quad \forall x, y \in \mathbb{R}^n. \tag{1.2.12}$$

**Lemma 1.2.4** *Let $f \in C_M^{2,2}(\mathbb{R}^n)$. Then for all $x, y \in \mathbb{R}^n$ we have*

$$\| \ \nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x) \ \| \leq \tfrac{M}{2} \ \| \ y - x \ \|^2, \tag{1.2.13}$$

$$\begin{aligned}
&|f(y) - f(x) - \langle \nabla f(x), y - x \rangle - \tfrac{1}{2}\langle \nabla^2 f(x)(y - x), y - x \rangle| \\
&\leq \tfrac{M}{6} \ \| \ y - x \ \|^3 \ .
\end{aligned} \tag{1.2.14}$$

*Proof* Let us fix some $x, y \in \mathbb{R}^n$. Then

$$\nabla f(y) = \nabla f(x) + \int\limits_0^1 \nabla^2 f(x + \tau(y - x))(y - x) d\tau$$

$$= \nabla f(x) + \nabla^2 f(x)(y - x) + \int\limits_0^1 (\nabla^2 f(x + \tau(y - x)) - \nabla^2 f(x))(y - x) d\tau.$$

Therefore,

$$\begin{aligned} &\| \nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x) \| \\ &= \| \int\limits_0^1 (\nabla^2 f(x + \tau(y - x)) - \nabla^2 f(x))(y - x) d\tau \| \\ &\le \int\limits_0^1 \| (\nabla^2 f(x + \tau(y - x)) - \nabla^2 f(x))(y - x) \| d\tau \\ &\le \int\limits_0^1 \| \nabla^2 f(x + \tau(y - x)) - \nabla^2 f(x) \| \cdot \| y - x \| d\tau \\ &\le \int\limits_0^1 \tau M \| y - x \|^2 d\tau \; = \; \tfrac{M}{2} \| y - x \|^2 . \end{aligned}$$

Inequality (1.2.14) can be proved in a similar way. □

**Corollary 1.2.2** *Let $f \in C^{2,2}_M(\mathbb{R}^n)$ and $x, y \in \mathbb{R}^n$ with $\| y - x \| = r$. Then*

$$\nabla^2 f(x) - MrI_n \preceq \nabla^2 f(y) \preceq \nabla^2 f(x) + MrI_n.$$

(Recall that for matrices $A$ and $B$ we write $A \succeq B$ if $A - B \succeq 0$.)

*Proof* Let $G = \nabla^2 f(y) - \nabla^2 f(x)$. Since $f \in C^{2,2}_M(\mathbb{R}^n)$, we have $\| G \| \le Mr$. This means that the eigenvalues of the symmetric matrix $G$, $\lambda_i(G)$, satisfy the following inequality:

$$| \lambda_i(G) | \le Mr, \quad i = 1 \dots n.$$

Hence, $-MrI_n \preceq G \equiv \nabla^2 f(y) - \nabla^2 f(x) \preceq MrI_n$. □

### 1.2.3 The Gradient Method

Now we are ready to sudy the rate of convergence of unconstrained minimization schemes. Let us start with the simplest method. As we have already seen, the antigradient is the direction of locally steepest descent of a differentiable function. Since we are going to find a local minimum, the following strategy is the first to be tried.

**Gradient Method**

(1.2.15)

**Choose** $x_0 \in \mathbb{R}^n$.
**Iterate** $x_{k+1} = x_k - h_k \nabla f(x_k),\ k = 0, 1, \dots$.

We will refer to this scheme as the *Gradient Method*. The scalar factors for the gradients, $h_k$, are called the *step sizes*. Of course, they must be positive.

There are many variants of this method, which differ one from another by the *step-size strategy*. Let us consider the most important examples.

1. The sequence $\{h_k\}_{k=0}^{\infty}$ is chosen *in advance*. For example,

$$h_k = h > 0, \text{ (constant step)}$$

$$h_k = \frac{h}{\sqrt{k+1}}.$$

2. *Full relaxation*:

$$h_k = \arg\min_{h \geq 0} f(x_k - h\nabla f(x_k)).$$

3. The *Armijo* rule: Find $x_{k+1} = x_k - h\nabla f(x_k)$ with $h > 0$ such that

$$\alpha\langle \nabla f(x_k), x_k - x_{k+1}\rangle \leq f(x_k) - f(x_{k+1}), \tag{1.2.16}$$

$$\beta\langle \nabla f(x_k), x_k - x_{k+1}\rangle \geq f(x_k) - f(x_{k+1}), \tag{1.2.17}$$

where $0 < \alpha < \beta < 1$ are some fixed parameters.

Comparing these strategies, we see that the first strategy is the simplest one. It is often used in the context of Convex Optimization. In this framework, the behavior of functions is much more predictable than in the general nonlinear case.

The second strategy is completely theoretical. It is never used in practice since even in one-dimensional case we cannot find the exact minimum in finite time.

The third strategy is used in the majority of practical algorithms. It has the following geometric interpretation. Let us fix $x \in \mathbb{R}^n$ assuming that $\nabla f(x) \neq 0$. Consider the following function of one variable:

$$\phi(h) = f(x - h\nabla f(x)), \quad h \geq 0.$$

Then the step-size values acceptable for this strategy belong to the part of the graph of $\phi$ which is located between two linear functions:

$$\phi_1(h) = f(x) - \alpha h \parallel \nabla f(x) \parallel^2, \quad \phi_2(h) = f(x) - \beta h \parallel \nabla f(x) \parallel^2 .$$

Note that $\phi(0) = \phi_1(0) = \phi_2(0)$ and $\phi'(0) < \phi_2'(0) < \phi_1'(0) < 0$. Therefore, the acceptable values exist unless $\phi(\cdot)$ is not bounded below. There are several very fast one-dimensional procedures for finding a point satisfying the Armijo conditions. However, their detailed description is not important for us now.

Let us estimate the performance of the Gradient Method. Consider the problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{1.2.18}$$

with $f \in C_L^{1,1}(\mathbb{R}^n)$, and assume that $f(\cdot)$ is bounded below on $\mathbb{R}^n$.

Let us evaluate the result of one gradient step. Consider $y = x - h\nabla f(x)$. Then, in view of (1.2.11), we have

$$\begin{aligned} f(y) &\leq f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{L}{2} \parallel y - x \parallel^2 \\ &= f(x) - h \parallel \nabla f(x) \parallel^2 + \tfrac{h^2}{2} L \parallel \nabla f(x) \parallel^2 \\ &= f(x) - h(1 - \tfrac{h}{2}L) \parallel \nabla f(x) \parallel^2 . \end{aligned} \tag{1.2.19}$$

Thus, in order to get the best upper bound for the possible decrease of the objective function, we have to solve the following one-dimensional problem:

$$\Delta(h) = -h\left(1 - \tfrac{h}{2}L\right) \to \min_h .$$

Computing the derivative of this function, we conclude that the optimal step size must satisfy the equation $\Delta'(h) = hL - 1 = 0$. Thus, $h^* = \frac{1}{L}$, which is a minimum of $\Delta(h)$ since $\Delta''(h) = L > 0$.

Thus, our considerations prove that one step of the Gradient Method decreases the value of the objective function at least as follows:

$$f(y) \leq f(x) - \tfrac{1}{2L} \parallel \nabla f(x) \parallel^2 .$$

Let us check what is going on with the other step-size strategies.

Let $x_{k+1} = x_k - h_k \nabla f(x_k)$. Then for the constant step strategy, $h_k = h$, we have

$$f(x_k) - f(x_{k+1}) \geq h(1 - \frac{1}{2}Lh) \parallel \nabla f(x_k) \parallel^2 .$$

Therefore, if we choose $h_k = \frac{2\alpha}{L}$ with $\alpha \in (0, 1)$, then

$$f(x_k) - f(x_{k+1}) \geq \tfrac{2}{L}\alpha(1-\alpha) \parallel \nabla f(x_k) \parallel^2 .$$

Of course, the optimal choice is $h_k = \frac{1}{L}$.

For the full relaxation strategy we have

$$f(x_k) - f(x_{k+1}) \geq \tfrac{1}{2L} \parallel \nabla f(x_k) \parallel^2$$

since the maximal decrease is not worse than the decrease attained by $h_k = \frac{1}{L}$.

Finally, for the Armijo rule, in view of (1.2.17), we have

$$f(x_k) - f(x_{k+1}) \leq \beta \langle \nabla f(x_k), x_k - x_{k+1} \rangle = \beta h_k \parallel \nabla f(x_k) \parallel^2 .$$

From (1.2.19), we obtain

$$f(x_k) - f(x_{k+1}) \geq h_k \left(1 - \tfrac{h_k}{2} L\right) \parallel \nabla f(x_k) \parallel^2 .$$

Therefore, $h_k \geq \frac{2}{L}(1-\beta)$. Further, using (1.2.16), we have

$$f(x_k) - f(x_{k+1}) \geq \alpha \langle \nabla f(x_k), x_k - x_{k+1} \rangle = \alpha h_k \parallel \nabla f(x_k) \parallel^2 .$$

Combining this inequality with the previous one, we conclude that

$$f(x_k) - f(x_{k+1}) \geq \tfrac{2}{L}\alpha(1-\beta) \parallel \nabla f(x_k) \parallel^2 .$$

Thus, we have proved that in *all* cases we have

$$f(x_k) - f(x_{k+1}) \geq \tfrac{\omega}{L} \parallel \nabla f(x_k) \parallel^2, \tag{1.2.20}$$

where $\omega$ is some positive constant.

Now we are ready to estimate the performance of Gradient Method. Summing up the inequalities (1.2.20) for $k = 0 \ldots N$, we obtain

$$\tfrac{\omega}{L} \sum_{k=0}^{N} \parallel \nabla f(x_k) \parallel^2 \leq f(x_0) - f(x_{N+1}) \leq f(x_0) - f^*, \tag{1.2.21}$$

where $f^*$ is a lower bounds for the values of objective function in the problem (1.2.1). As a simple consequence of the bound (1.2.21), we have

$$\| \nabla f(x_k) \| \to 0 \quad \text{as} \quad k \to \infty.$$

However, we can also say something about the *rate of convergence*. Indeed, define

$$g_N^* = \min_{0 \le k \le N} \| \nabla f(x_k) \| .$$

Then, in view of (1.2.21), we come to the following inequality:

$$g_N^* \le \frac{1}{\sqrt{N+1}} \left[ \frac{1}{\omega} L(f(x_0) - f^*) \right]^{1/2}. \tag{1.2.22}$$

The right-hand side of this inequality describes the *rate of convergence* of the sequence $\{g_N^*\}$ to zero. Note that we cannot say anything about the rate of convergence of the sequences $\{f(x_k)\}$ and $\{x_k\}$.

Recall that in general Nonlinear Optimization, our current goal is quite modest: we only want to approach a local minimum of the optimization problem (1.2.18). Nevertheless, in general, even this goal is unreachable for the Gradient Method. Let us consider the following example.

*Example 1.2.2* Consider the following function of two variables:

$$f(x) \equiv f(x^{(1)}, x^{(2)}) = \tfrac{1}{2}(x^{(1)})^2 + \tfrac{1}{4}(x^{(2)})^4 - \tfrac{1}{2}(x^{(2)})^2.$$

The gradient of this function is $\nabla f(x) = (x^{(1)}, (x^{(2)})^3 - x^{(2)})^T$. Therefore, there are only three points which can pretend to be a local minimum of this function:

$$x_1^* = (0,0), \quad x_2^* = (0,-1), \quad x_3^* = (0,1).$$

Computing the Hessian of this function,

$$\nabla^2 f(x) = \begin{pmatrix} 1 & 0 \\ 0 & 3(x^{(2)})^2 - 1 \end{pmatrix},$$

we conclude that $x_2^*$ and $x_3^*$ are isolated local minima,[3] but $x_1^*$ is only a *stationary point* of our function. Indeed, $f(x_1^*) = 0$ and $f(x_1^* + \epsilon e_2) = \frac{\epsilon^4}{4} - \frac{\epsilon^2}{2} < 0$ for $\epsilon$ small enough.

Let us consider now the trajectory of the Gradient Method which starts at $x_0 = (1,0)$. Note that the second coordinate of this point is zero. Therefore, the second coordinate of $\nabla f(x_0)$ is also zero. Consequently, the second coordinate of $x_1$ is

[3] In fact, in our example they are global solutions.

zero, etc. Thus, the entire sequence of points generated by the Gradient Method will have the second coordinate equal to zero. This means that this sequence converges to $x_1^*$.

To conclude our example, note that this situation is typical for all first-order unconstrained minimization methods. Without additional rather restrictive assumptions, it is impossible to guarantee their global convergence to a local minimum. Only a stationary point can be approached by these schemes. □

Note that inequality (1.2.22) provides us with an example of a new notion, that is, the *rate of convergence* of a minimization process. How can we use this information in the complexity analysis? The rate of convergence delivers an *upper* complexity bound for the corresponding problem class. Such a bound is always justified by some numerical method. A method for which the upper complexity bound is proportional to the *lower* complexity bound of the problem class is said to be *optimal*. Recall that in Sect. 1.1.3 we have already seen an optimal method for the problem class $\mathscr{P}_\infty$.

Let us now present a formal description of our result. Consider the following problem class $\mathscr{G}_*$.

| | |
|---|---|
| **Model** : | 1. Unconstrained minimization.<br>2. $f \in C_L^{1,1}(\mathbb{R}^n)$.<br>3. $f(\cdot)$ is bounded below by the value $f^*$. |
| **Oracle** : | First-order Black Box. |
| $\varepsilon$-**solution** : | $f(\bar{x}) \leq f(x_0), \ \parallel \nabla f(\bar{x}) \parallel \leq \epsilon.$ |

(1.2.23)

Note that inequality (1.2.22) can be used in order to obtain an upper bound for the number of steps (= calls of the oracle), which is necessary to find a point where the norm of the gradient is small. For that, let us write down the following inequality:

$$g_N^* \leq \frac{1}{\sqrt{N+1}} \left[ \frac{1}{\omega} L(f(x_0) - f^*) \right]^{1/2} \leq \epsilon. \tag{1.2.24}$$

Therefore, if $N + 1 \geq \frac{L}{\omega \epsilon^2}(f(x_0) - f^*)$, then we necessarily have $g_N^* \leq \epsilon$.

Thus, we can use the value $\frac{L}{\omega \epsilon^2}(f(x_0) - f^*)$ as an *upper complexity bound* for our problem class. Comparing this estimate with the result of Theorem 1.1.2, we can see that it is much better. At least it does not depend on $n$. The lower complexity bound for the class $\mathscr{G}_*$ is unknown.

Let us see, what can be said about the *local* convergence of the Gradient Method. Consider the unconstrained minimization problem

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

under the following assumptions.

1. $f \in C_M^{2,2}(\mathbb{R}^n)$.
2. There exists a local minimum $x^* \in \mathbb{R}^n$ of function $f$ at which the Hessian is *positive definite*.
3. We know some bounds $0 < \mu \leq L < \infty$ for the Hessian at $x^*$:

$$\mu I_n \preceq \nabla^2 f(x^*) \preceq L I_n. \tag{1.2.25}$$

4. Our starting point $x_0$ is close enough to $x^*$.

Consider the process: $x_{k+1} = x_k - h_k \nabla f(x_k)$. Note that $\nabla f(x^*) = 0$. Hence,

$$\nabla f(x_k) = \nabla f(x_k) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*))(x_k - x^*) d\tau$$

$$= G_k(x_k - x^*),$$

where $G_k = \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau$. Therefore,

$$x_{k+1} - x^* = x_k - x^* - h_k G_k(x_k - x^*) = (I_n - h_k G_k)(x_k - x^*).$$

There is a standard technique for analyzing processes of this type, which is based on *contraction mappings*. Let the sequence $\{a_k\}$ be defined as follows:

$$a_0 \in \mathbb{R}^n, \quad a_{k+1} = A_k a_k,$$

where $A_k$ are $(n \times n)$-matrices such that $\| A_k \| \leq 1 - q$ for all $k \geq 0$ with $q \in (0, 1)$. Then we can estimate the rate of convergence of the sequence $\{a_k\}$ to zero:

$$\| a_{k+1} \| \leq (1-q) \| a_k \| \leq (1-q)^{k+1} \| a_0 \| \to 0.$$

In our case, we need to estimate $\| I_n - h_k G_k \|$. Let $r_k = \| x_k - x^* \|$. In view of Corollary 1.2.2, we have

$$\nabla^2 f(x^*) - \tau M r_k I_n \preceq \nabla^2 f(x^* + \tau(x_k - x^*)) \preceq \nabla^2 f(x^*) + \tau M r_k I_n.$$

Therefore, using assumption (1.2.25), we obtain

$$(\mu - \tfrac{r_k}{2}M)I_n \preceq G_k \;\preceq\; (L + \tfrac{r_k}{2}M)I_n.$$

Hence, $(1-h_k(L+\frac{r_k}{2}M))I_n \preceq I_n - h_k G_k \preceq (1-h_k(\mu-\frac{r_k}{2}M))I_n$, and we conclude that

$$\| \; I_n - h_k G_k \; \| \leq \max\{a_k(h_k), b_k(h_k)\}, \tag{1.2.26}$$

where $a_k(h) = 1 - h(\mu - \frac{r_k}{2}M)$ and $b_k(h) = h(L + \frac{r_k}{2}M) - 1$.

Note that $a_k(0) = 1$ and $b_k(0) = -1$. Therefore, if $0 < r_k < \bar{r} \equiv \frac{2\mu}{M}$, then $a_k(\cdot)$ is a strictly decreasing function and we can ensure

$$\| \; I_n - h_k G_k \; \| < 1$$

for $h_k$ small enough. In this case, we will have $r_{k+1} < r_k$.

As usual, many step-size strategies are available. For example, we can choose $h_k = \frac{1}{L}$. Let us consider the "optimal" strategy consisting in minimizing the right-hand side of (1.2.26):

$$\max\{a_k(h), b_k(h)\} \to \min_h .$$

Assume that $r_0 < \bar{r}$. Then, if we form the sequence $\{x_k\}$ using the optimal strategy, we can be sure that $r_{k+1} < r_k < \bar{r}$. Further, the optimal step size $h_k^*$ can be found from the equation

$$a_k(h) = b_k(h) \quad \Leftrightarrow \quad 1 - h(\mu - \tfrac{r_k}{2}M) = h(L + \tfrac{r_k}{2}M) - 1.$$

Hence

$$h_k^* = \tfrac{2}{L+\mu}. \tag{1.2.27}$$

(Surprisingly enough, the optimal step size does not depend on $M$.) Under this choice, we obtain

$$r_{k+1} \leq \tfrac{(L-\mu)r_k}{L+\mu} + \tfrac{Mr_k^2}{L+\mu}.$$

Let us estimate the rate of convergence of the process. Let $q = \frac{2\mu}{L+\mu}$ and $a_k = \frac{M}{L+\mu} r_k$ $(< q)$. Then

$$a_{k+1} \leq (1-q)a_k + a_k^2 = a_k(1 + (a_k - q)) = \tfrac{a_k(1-(a_k-q)^2)}{1-(a_k-q)} \leq \tfrac{a_k}{1+q-a_k}.$$

Therefore $\frac{1}{a_{k+1}} \geq \frac{1+q}{a_k} - 1$, or

$$\frac{q}{a_{k+1}} - 1 \geq \frac{q(1+q)}{a_k} - q - 1 = (1+q)\left(\frac{q}{a_k} - 1\right).$$

Hence,

$$\begin{aligned} \frac{q}{a_k} - 1 &\geq (1+q)^k \left(\frac{q}{a_0} - 1\right) = (1+q)^k \left(\frac{2\mu}{L+\mu} \cdot \frac{L+\mu}{r_0 M} - 1\right) \\ &= (1+q)^k \left(\frac{\bar{r}}{r_0} - 1\right). \end{aligned}$$

Thus,

$$a_k \leq \frac{q r_0}{r_0 + (1+q)^k (\bar{r} - r_0)} \leq \frac{q r_0}{\bar{r} - r_0} \left(\frac{1}{1+q}\right)^k.$$

This proves the following theorem.

**Theorem 1.2.4** *Let the function $f(\cdot)$ satisfy our assumptions and let the starting point $x_0$ be close enough to a strict local minimum $x^*$:*

$$r_0 = \| x_0 - x^* \| < \bar{r} = \frac{2\mu}{M}.$$

*Then the Gradient Method with step size (1.2.27) converges as follows:*

$$\| x_k - x^* \| \leq \frac{\bar{r} r_0}{\bar{r} - r_0} \left(1 - \frac{2\mu}{L+3\mu}\right)^k.$$

This type of rate of convergence is called *linear*.

### 1.2.4 Newton's Method

Newton's Method is widely known as a technique for finding a root of a univariate function. Let $\phi(\cdot) : \mathbb{R} \to \mathbb{R}$. Consider the equation

$$\phi(t^*) = 0.$$

Newton's rule can be obtained by linear approximation. Assume that we know some $t \in \mathbb{R}$ which is close enough to $t^*$. Note that

$$\phi(t + \Delta t) = \phi(t) + \phi'(t)\Delta t + o(| \Delta t |).$$

Therefore, the solution of the equation $\phi(t + \Delta t) = 0$ can be approximated by the solution of the following *linear* equation:

$$\phi(t) + \phi'(t)\Delta t = 0.$$

Under some conditions, we can expect the displacement $\Delta t$ to be a good approximation to the optimal displacement $\Delta t^* = t^* - t$. Converting this idea into an algorithm, we get the process

$$t_{k+1} = t_k - \tfrac{\phi(t_k)}{\phi'(t_k)}.$$

This scheme can be naturally extended to the problem of finding a solution to a system of nonlinear equations,

$$F(x) = 0,$$

where $x \in \mathbb{R}^n$ and $F(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$. In this case, we need to define the displacement $\Delta x$ as a solution to the following system of linear equations:

$$F(x) + F'(x)\Delta x = 0$$

(called the *Newton system*). If the Jacobian $F'(x)$ is nondegenerate, we can compute the displacement $\Delta x = -[F'(x)]^{-1}F(x)$. The corresponding iterative scheme is as follows:

$$x_{k+1} = x_k - [F'(x_k)]^{-1}F(x_k).$$

Finally, in view of Theorem 1.2.1, we can replace the unconstrained minimization problem (1.2.1) by the problem of finding a root of the nonlinear system

$$\nabla f(x) = 0. \tag{1.2.28}$$

(This replacement is not completely equivalent, but it works in nondegenerate situations.) Further, to solve (1.2.28) we can apply the standard Newton Method for the system of nonlinear equations. In this case, the Newton system is as follows:

$$\nabla f(x) + \nabla^2 f(x)\Delta x = 0.$$

Hence, the Newton's Method for optimization problems can be written in the following form:

$$\boxed{x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1}\nabla f(x_k).} \tag{1.2.29}$$

Note that we can obtain the process (1.2.29) using the idea of quadratic approximation. Consider this approximation, computed with respect to the point $x_k$:

$$\phi(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle.$$

Assume that $\nabla^2 f(x_k) \succ 0$. Then we can choose $x_{k+1}$ as the minimizer of the quadratic function $\phi(\cdot)$. This means that

$$\nabla\phi(x_{k+1}) = \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k) = 0,$$

and we come again to Newton's process (1.2.29).

We will see that the convergence of the Newton's Method in a neighborhood of a strict local minimum is very fast. However, this method has two serious drawbacks. Firstly, it can break down if $\nabla^2 f(x_k)$ is degenerate. Secondly, Newton's process can diverge. Let us look at the following example.

*Example 1.2.3* Let us apply the Newton's Method for finding a root of the following univariate function:

$$\phi(t) = \tfrac{t}{\sqrt{1+t^2}}.$$

Clearly, $t^* = 0$. Note that

$$\phi'(t) = \tfrac{1}{[1+t^2]^{3/2}}.$$

Therefore Newton's process is as follows:

$$t_{k+1} = t_k - \tfrac{\phi(t_k)}{\phi'(t_k)} = t_k - \tfrac{t_k}{\sqrt{1+t_k^2}} \cdot [1 + t_k^2]^{3/2} = -t_k^3.$$

Thus, if $\mid t_0 \mid < 1$, then this method converges and the convergence is extremely fast. The points $\pm 1$ are oscillation points of this scheme. If $\mid t_0 \mid > 1$, then the method diverges. □

In order to avoid a possible divergence, in practice we can apply the *damped Newton's method*:

$$\boxed{x_{k+1} = x_k - h_k[\nabla^2 f(x_k)]^{-1}\nabla f(x_k),}$$

where $h_k > 0$ is a step size parameter. At the initial stage of the method we can use the same step size strategies as for the gradient scheme. At the final stage, it is reasonable to choose $h_k = 1$. Another possibility for ensuring the global convergence of this scheme consists in using Cubic Regularization. This approach will be studied in detail in Chap. 4.

Let us derive the local rate of convergence of the Newton's Method. Consider the problem

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

under the following assumptions:

1. $f \in C_M^{2,2}(\mathbb{R}^n)$.
2. There exists a local minimum of the function $f$ with *positive definite* Hessian:

$$\nabla^2 f(x^*) \succeq \mu I_n, \quad \mu > 0. \tag{1.2.30}$$

3. Our starting point $x_0$ is close enough to $x^*$.

Consider the process $x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$. Then, using the same reasoning as for the Gradient Method, we obtain the following representation:

$$\begin{aligned} x_{k+1} - x^* &= x_k - x^* - [\nabla^2 f(x_k)]^{-1}\nabla f(x_k) \\ &= x_k - x^* - [\nabla^2 f(x_k)]^{-1} \int\limits_0^1 \nabla^2 f(x^* + \tau(x_k - x^*))(x_k - x^*)d\tau \\ &= [\nabla^2 f(x_k)]^{-1} G_k(x_k - x^*), \end{aligned}$$

where $G_k = \int\limits_0^1 [\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*))]d\tau$.

Let $r_k =\parallel x_k - x^* \parallel$. Then

$$\begin{aligned} \parallel G_k \parallel &= \parallel \int\limits_0^1 [\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*))]d\tau \parallel \\ &\le \int\limits_0^1 \parallel \nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)) \parallel d\tau \\ &\le \int\limits_0^1 M(1-\tau)r_k d\tau = \tfrac{r_k}{2}M. \end{aligned}$$

In view of Corollary 1.2.2, and relation (1.2.30), we have

$$\nabla^2 f(x_k) \succeq \nabla^2 f(x^*) - Mr_k I_n \succeq (\mu - Mr_k)I_n.$$

Therefore, if $r_k < \frac{\mu}{M}$, then $\nabla^2 f(x_k)$ is positive definite and

$$\| [\nabla^2 f(x_k)]^{-1} \| \leq (\mu - Mr_k)^{-1}.$$

Hence, for $r_k$ small enough ($r_k \leq \frac{2\mu}{3M}$), we have

$$r_{k+1} \leq \frac{Mr_k^2}{2(\mu - Mr_k)} \quad (\leq r_k).$$

The rate of convergence of this type is called *quadratic*.

Thus, we have proved the following theorem.

**Theorem 1.2.5** *Let the function $f(\cdot)$ satisfy our assumptions. Suppose that the initial starting point $x_0$ is close enough to $x^*$:*

$$\| x_0 - x^* \| \leq \bar{r} = \frac{2\mu}{3M}.$$

*Then $\| x_k - x^* \| \leq \bar{r}$ for all $k$ and the Newton's Method converges quadratically:*

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

Comparing this result with the local rate of convergence of the Gradient Method, we see that the Newton's Method is much faster. Surprisingly enough, the *region of quadratic convergence* of the Newton's Method is almost the same as the region of linear convergence of the Gradient Method. This justifies the standard recommendation to use the Gradient Method only at the initial stage of the minimization process in order to get close to a local minimum. The final job should be performed by Newton's scheme. However, we will come back to a detailed comparison of the performance of these two methods in Chap. 4.

In this section, we have seen several examples of convergence rate. Let us find a correspondence between these rates and the complexity bounds. As we have already seen (for example, in the case of the problem class $\mathscr{G}_*$ (1.2.23)), the upper bound for the analytical complexity of a problem class is an inverse function of the rate of convergence.

1. *Sublinear rate.* This rate is described in terms of a power function of the iteration counter. For example, suppose that for some method we can prove the rate of convergence $r_k \leq \frac{c}{\sqrt{k}}$. In this case, the upper complexity bound justified by this scheme for the corresponding problem class is $\left(\frac{c}{\epsilon}\right)^2$.

   The sublinear rate is rather slow. In terms of complexity, each new right digit of the answer takes a number of iterations *comparable* with the total amount of the previous work. Note also, that the constant $c$ plays a significant role in the corresponding complexity bound.

2. *Linear rate.* This rate is given in terms of an exponential function of the iteration counter. For example,

$$r_k \le c(1-q)^k \le ce^{-qk}, \quad 0 < q \le 1.$$

Note that the corresponding complexity bound is $\frac{1}{q}(\ln c + \ln \frac{1}{\epsilon})$.

This rate is fast: Each new right digit of the answer takes a constant number of iterations. Moreover, the dependence of the complexity estimate on the constant $c$ is very weak.

3. *Quadratic rate.* This rate has a double exponential dependence in the iteration counter. For example,

$$r_{k+1} \le cr_k^2.$$

The corresponding complexity estimate depends on the double logarithm of the desired accuracy: $\ln \ln \frac{1}{\epsilon}$.

This rate is extremely fast: Each iteration doubles the number of right digits in the answer. The constant $c$ is important only for the starting moment of the quadratic convergence ($cr_k < 1$). For example, after the moment $cr_k \le \frac{1}{2}$, we can guarantee a fast convergence rate $r_{k+1} \le \frac{1}{2}r_k$, which does not depend on $c$ at all.

## 1.3 First-Order Methods in Nonlinear Optimization

(The Gradient Method and Newton's Method: What is different? The idea of a variable metric; Variable metric methods; Conjugate gradient methods; Constrained minimization; Lagrangian relaxation; A sufficient condition for zero duality gap; Penalty functions and penalty function methods; Barrier functions and barrier function methods.)

### *1.3.1 The Gradient Method and Newton's Method: What Is Different?*

In the previous section, we considered two local methods for finding a local minimum of the simplest minimization problem

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

with $f \in C_M^{2,2}(\mathbb{R}^n)$. Namely, the Gradient Method

$$x_{k+1} = x_k - h_k \nabla f(x_k), \quad h_k > 0.$$

and the Newton's Method:

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k).$$

Recall that the local rate of convergence of these methods is different. We have seen that the Gradient Method has a linear rate and the Newton's method converges quadratically. What is the reason for this difference?

If we look at the analytical form of these methods, we can see at least the following formal difference: In the Gradient Method, the search direction is the antigradient, while in the Newton's method we multiply the antigradient by some matrix, the inverse Hessian. Let us try to derive these directions using some "universal" reasoning.

Let us fix a point $\bar{x} \in \mathbb{R}^n$. Consider the following approximation of the function $f(\cdot)$:

$$\phi_1(x) = f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \tfrac{1}{2h} \parallel x - \bar{x} \parallel^2,$$

where the parameter $h$ is positive. The first-order optimality condition provides us with the following equation for $x_1^*$, the unconstrained minimum of this function:

$$\nabla \phi_1(x_1^*) = \nabla f(\bar{x}) + \tfrac{1}{h}(x_1^* - \bar{x}) = 0.$$

Thus, $x_1^* = \bar{x} - h\nabla f(\bar{x})$. This is exactly the iterate of the Gradient Method. Note that if $h \in (0, \frac{1}{L}]$, then the function $\phi_1(\cdot)$ is a *global upper* approximation of $f(\cdot)$:

$$f(x) \le \phi_1(x), \quad \forall x \in \mathbb{R}^n,$$

(see Lemma 1.2.3). This fact is responsible for the global convergence of the Gradient Method.

Further, consider a quadratic approximation of the function $f(\cdot)$:

$$\phi_2(x) = f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \tfrac{1}{2}\langle \nabla^2 f(\bar{x})(x - \bar{x}), x - \bar{x} \rangle.$$

We have already seen that the minimum of this function is

$$x_2^* = \bar{x} - [\nabla^2 f(\bar{x})]^{-1} \nabla f(\bar{x}),$$

and this is exactly the iterate of the Newton's Method.

Thus, we can try to use some quadratic approximations of the function $f(\cdot)$, which are better than $\phi_1(\cdot)$ and which are less expensive than $\phi_2(\cdot)$.

Let $G$ be a symmetric positive definite $n \times n$-matrix. Define

$$\phi_G(x) = f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{1}{2}\langle G(x - \bar{x}), x - \bar{x} \rangle.$$

Computing the minimizer of $\phi_G(\cdot)$ from the equation

$$\nabla \phi_G(x_G^*) = \nabla f(\bar{x}) + G(x_G^* - \bar{x}) = 0,$$

we obtain

$$x_G^* = \bar{x} - G^{-1}\nabla f(\bar{x}). \tag{1.3.1}$$

The first-order methods, which form a sequence of matrices

$$\{G_k\} : \; G_k \to \nabla^2 f(x^*)$$

(or $\{H_k\} : \; H_k \equiv G_k^{-1} \to [\nabla^2 f(x^*)]^{-1}$), are called *variable metric* methods. (Sometimes the name *quasi-Newton* methods is used.) In these methods, only the gradients are involved in the process of generating the sequences $\{G_k\}$ or $\{H_k\}$.

The updating rule (1.3.1) is very common in Optimization. Let us provide it with one more interpretation.

Note that the gradient and Hessian of a nonlinear function $f(\cdot)$ are defined *with respect to* the standard Euclidean inner product on $\mathbb{R}^n$:

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x^{(i)} y^{(i)}, \; x, y \in \mathbb{R}^n, \quad \| x \| = \langle x, x \rangle^{1/2}.$$

Indeed, the definition of the gradient is as follows:

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + o(\| h \|).$$

From this equation, we derive its coordinate representation:

$$\nabla f(x) = \left( \frac{\partial f(x)}{\partial x^{(1)}}, \ldots, \frac{\partial f(x)}{\partial x^{(n)}} \right)^T .$$

Let us now introduce a new inner product. Consider a symmetric positive definite $(n \times n)$-matrix $A$. For $x, y \in \mathbb{R}^n$ define

$$\langle x, y \rangle_A = \langle Ax, y \rangle, \quad \| x \|_A = \langle Ax, x \rangle^{1/2}.$$

The function $\| \cdot \|_A$ is treated as a new *norm* on $\mathbb{R}^n$. Note that topologically this new norm is equivalent to the old one:

$$\lambda_{\min}(A)^{1/2} \| x \| \leq \| x \|_A \leq \lambda_{\max}(A)^{1/2} \| x \|,$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the smallest and the largest eigenvalues of the matrix $A$. However, the gradient and the Hessian, computed with respect to the new inner product, are different:

$$\begin{aligned} f(x+h) &= f(x) + \langle \nabla f(x), h\rangle + \tfrac{1}{2}\langle \nabla^2 f(x)h, h\rangle + o(\| h \|) \\ &= f(x) + \langle A^{-1}\nabla f(x), h\rangle_A + \tfrac{1}{2}\langle A^{-1}\nabla^2 f(x)h, h\rangle_A + o(\| h \|_A). \end{aligned}$$

Hence, $\nabla f_A(x) = A^{-1}\nabla f(x)$ is the new gradient and $\nabla^2 f_A(x) = A^{-1}\nabla^2 f(x)$ is the new Hessian.

Thus, the direction used in the Newton's method can be seen as a gradient direction computed with respect to the inner product defined by $A = \nabla^2 f(x) \succ 0$. Note that the Hessian of $f(\cdot)$ at $x$ computed with respect to $A = \nabla^2 f(x)$ is $I_n$.

*Example 1.3.1* Consider the quadratic function

$$f(x) = \alpha + \langle a, x\rangle + \frac{1}{2}\langle Ax, x\rangle,$$

where $A = A^T \succ 0$. Note that $\nabla f(x) = Ax + a$, $\nabla^2 f(x) = A$ and

$$\nabla f(x^*) = Ax^* + a = 0$$

for $x^* = -A^{-1}a$. Let us compute the Newton's direction at some $x \in \mathbb{R}^n$:

$$d_N(x) = [\nabla^2 f(x)]^{-1}\nabla f(x) = A^{-1}(Ax + a) = x + A^{-1}a.$$

Therefore for any $x \in \mathbb{R}^n$ we have $x - d_N(x) = -A^{-1}a = x^*$. Thus, for a quadratic function, Newton's method converges in one step. Note also that

$$\begin{aligned} f(x) &= \alpha + \langle A^{-1}a, x\rangle_A + \tfrac{1}{2} \| x \|_A^2, \\ \nabla f_A(x) &= A^{-1}\nabla f(x) = d_N(x), \\ \nabla^2 f_A(x) &= A^{-1}\nabla^2 f(x) = I_n. \qquad \square \end{aligned}$$

Let us look at the general scheme of the *variable metric* methods.

```
Variable metric method

0. Choose $x_0 \in \mathbb{R}^n$. Set $H_0 = I_n$. Compute $f(x_0)$ and $\nabla f(x_0)$.

1. $k$th iteration ($k \geq 0$).
    (a) Set $p_k = H_k \nabla f(x_k)$.
    (b) Find $x_{k+1} = x_k - h_k p_k$
        (see Section 1.2.3 for step size rules).
    (c) Compute $f(x_{k+1})$ and $\nabla f(x_{k+1})$.
    (d) Update the matrix $H_k$ to $H_{k+1}$.
```

The variable metric schemes differ from one to another only in the implementation of Step 1(d), which updates the matrix $H_k$. For that, they use new information, accumulated at Step 1(c), namely the gradient $\nabla f(x_{k+1})$. This update is justified by the following property of quadratic functions. Let

$$f(x) = \alpha + \langle a, x \rangle + \frac{1}{2}\langle Ax, x \rangle, \quad \nabla f(x) = Ax + a.$$

Then, for any $x, y \in \mathbb{R}^n$ we have $\nabla f(x) - \nabla f(y) = A(x - y)$. This identity explains the origin of the so-called *quasi-Newton rule*.

**Quasi-Newton rule**

Choose $H_{k+1} = H_{k+1}^T \succ 0$ such that

$$H_{k+1}(\nabla f(x_{k+1}) - \nabla f(x_k)) = x_{k+1} - x_k.$$

Actually, there are many ways to satisfy this relation. Below, we present several examples of schemes which are usually recommended as the most efficient.

Define

$$\Delta H_k = H_{k+1} - H_k, \quad \gamma_k = \nabla f(x_{k+1}) - \nabla f(x_k), \quad \delta_k = x_{k+1} - x_k.$$

Then the quasi-Newton relation is satisfied by the following updating rules.

1. *Rank-one correction scheme*: $\Delta H_k = \frac{(\delta_k - H_k\gamma_k)(\delta_k - H_k\gamma_k)^T}{\langle \delta_k - H_k\gamma_k, \gamma_k\rangle}$.
2. *Davidon–Fletcher–Powell scheme (DFP)*: $\Delta H_k = \frac{\delta_k \delta_k^T}{\langle \gamma_k, \delta_k\rangle} - \frac{H_k\gamma_k\gamma_k^T H_k}{\langle H_k\gamma_k, \gamma_k\rangle}$.
3. *Broyden–Fletcher–Goldfarb–Shanno scheme (BFGS)*:

$$\Delta H_k = \beta_k \frac{\delta_k \delta_k^T}{\langle \gamma_k, \delta_k\rangle} - \frac{H_k\gamma_k\delta_k^T + \delta_k\gamma_k^T H_k}{\langle \gamma_k, \delta_k\rangle},$$

where $\beta_k = 1 + \langle H_k\gamma_k, \gamma_k\rangle / \langle \gamma_k, \delta_k\rangle$.

Clearly, there are many other possibilities. From the computational point of view, BFGS is considered to be the most stable scheme.

Note that for quadratic functions, the variable metric methods usually terminate in at most $n$ iterations. In a neighborhood of a strict local minimum $x^*$ they demonstrate a *superlinear* rate of convergence: for any $x_0 \in \mathbb{R}^n$ close enough to $x^*$ there exists a number $N$ such that for all $k \geq N$ we have

$$\| x_{k+1} - x^* \| \leq \text{const} \cdot \| x_k - x^* \| \cdot \| x_{k-n} - x^* \|$$

(the proofs are very long and technical). As far as the worst-case global convergence is concerned, these methods are not better than the Gradient Method.

In the variable metric schemes it is necessary to store and update a symmetric $(n \times n)$-matrix. Thus, each iteration needs $O(n^2)$ auxiliary arithmetic operations. This feature is considered as one of the main drawbacks of the variable metric methods. It stimulated the interest in *conjugate gradient* schemes which have a much lower complexity of each iteration. We discuss these schemes in Sect. 1.3.2.

### 1.3.2 Conjugate Gradients

Conjugate gradient methods were initially proposed for minimizing quadratic functions. Consider the problem

$$\min_{x \in \mathbb{R}^n} f(x) \tag{1.3.2}$$

with $f(x) = \alpha + \langle a, x\rangle + \frac{1}{2}\langle Ax, x\rangle$ and $A = A^T \succ 0$. We have already seen that the solution of this problem is $x^* = -A^{-1}a$. Therefore, our objective function can be written in the following form:

$$f(x) = \alpha + \langle a, x\rangle + \tfrac{1}{2}\langle Ax, x\rangle \;=\; \alpha - \langle Ax^*, x\rangle + \tfrac{1}{2}\langle Ax, x\rangle$$

$$= \alpha - \tfrac{1}{2}\langle Ax^*, x^*\rangle + \tfrac{1}{2}\langle A(x - x^*), x - x^*\rangle.$$

Thus, $f^* = \alpha - \frac{1}{2}\langle Ax^*, x^*\rangle$ and $\nabla f(x) = A(x - x^*)$.

Suppose we are given a starting point $x_0 \in \mathbb{R}^n$. Consider the linear *Krylov* subspaces

$$\mathscr{L}_k = \mathrm{Lin}\{A(x_0 - x^*), \ldots, A^k(x_0 - x^*)\}, \quad k \geq 1,$$

where $A^k$ is the $k$th power of matrix $A$. A sequence of points $\{x_k\}$ is generated by the *Conjugate Gradient Method* in accordance with the following rule.

$$\boxed{x_k = \arg\min\{f(x) \mid \; x \in x_0 + \mathscr{L}_k\}, \; k \geq 1.} \tag{1.3.3}$$

This definition looks quite artificial. However, later we will see that this method can be written in a pure “algorithmic” form. We need representation (1.3.3) only for theoretical analysis.

**Lemma 1.3.1** *For any $k \geq 1$ we have $\mathscr{L}_k = \mathrm{Lin}\{\nabla f(x_0), \ldots, \nabla f(x_{k-1})\}$.*

*Proof* For $k = 1$, the statement is true since $\nabla f(x_0) = A(x_0 - x^*)$. Suppose that it is valid for some $k \geq 1$. Consider a point

$$x_k = x_0 + \sum_{i=1}^{k} \lambda^{(i)} A^i(x_0 - x^*) \in x_0 + \mathscr{L}_k$$

with some $\lambda \in \mathbb{R}^k$. Then

$$\nabla f(x_k) = A(x_0 - x^*) + \sum_{i=1}^{k} \lambda^{(i)} A^{i+1}(x_0 - x^*) = y + \lambda^{(k)} A^{k+1}(x_0 - x^*),$$

for a certain $y$ from $\mathscr{L}_k$. Thus,

$$\mathscr{L}_{k+1} \equiv \mathrm{Lin}\{\mathscr{L}_k \bigcup A^{k+1}(x_0 - x^*)\} = \mathrm{Lin}\{\mathscr{L}_k \bigcup \nabla f(x_k)\}$$

$$= \mathrm{Lin}\{\nabla f(x_0), \ldots, \nabla f(x_k)\}. \qquad \square$$

The next result helps us to understand the behavior of the sequence $\{x_k\}$.

**Lemma 1.3.2** *For any $k$, $i \geq 0$, $k \neq i$ we have $\langle \nabla f(x_k), \nabla f(x_i) \rangle = 0$.*

*Proof* Let $k > i$. Consider the function

$$\phi(\lambda) = f\left( x_0 + \sum_{j=1}^{k} \lambda^{(j)} \nabla f(x_{j-1}) \right), \quad \lambda \in \mathbb{R}^k.$$

In view of Lemma 1.3.1, for some $\lambda_* \in \mathbb{R}^k$ we have $x_k = x_0 + \sum\limits_{j=1}^{k} \lambda_*^{(j)} \nabla f(x_{j-1})$. However, by definition, $x_k$ is the minimum point of $f(\cdot)$ on $x_0 + \mathscr{L}_k$. Therefore $\nabla \phi(\lambda_*) = 0$. It remains to compute the components of the gradient:

$$0 = \tfrac{\partial \phi(\lambda_*)}{\partial \lambda^{(j)}} = \langle \nabla f(x_k), \nabla f(x_{j-1}) \rangle, \quad j = 1, \ldots, k. \quad \square$$

This lemma has two evident consequences.

**Corollary 1.3.1** *The sequence generated by the Conjugate Gradient Method for problem (1.3.2) is finite.*

*Proof* Indeed, the number of nonzero orthogonal directions in $\mathbb{R}^n$ cannot exceed $n$. □

**Corollary 1.3.2** *For any $p \in \mathscr{L}_k$, $k \geq 1$, we have $\langle \nabla f(x_k), p \rangle = 0$.* □

The last auxiliary result explains the name of the method. Let $\delta_i = x_{i+1} - x_i$. It is clear that $\mathscr{L}_k = \mathrm{Lin}\{\delta_0, \ldots, \delta_{k-1}\}$.

**Lemma 1.3.3** *For any $k, i \geq 0$, $k \neq i$, we have $\langle A\delta_k, \delta_i \rangle = 0$.*

(Such directions are called *conjugate* with respect to $A$.)

*Proof* Without loss of generality, we can assume that $k > i$. Then

$$\langle A\delta_k, \delta_i \rangle = \langle A(x_{k+1} - x_k), \delta_i \rangle = \langle \nabla f(x_{k+1}) - \nabla f(x_k), \delta_i \rangle = 0$$

since $\delta_i = x_{i+1} - x_i \in \mathscr{L}_{i+1} \subseteq \mathscr{L}_k \subseteq \mathscr{L}_{k+1}$. □

Let us show how we can write down the Conjugate Gradient Method in a more algorithmic form. Since $\mathscr{L}_k = \mathrm{Lin}\{\delta_0, \ldots, \delta_{k-1}\}$, we can represent $x_{k+1}$ as follows:

$$x_{k+1} = x_k - h_k \nabla f(x_k) + \sum_{j=0}^{k-1} \lambda^{(j)} \delta_j.$$

In our notation, this is

$$\delta_k = -h_k \nabla f(x_k) + \sum_{j=0}^{k-1} \lambda^{(j)} \delta_j. \tag{1.3.4}$$

Let us compute the coefficients in this representation. Multiplying (1.3.4) by $A$ and $\delta_i$, $0 \le i \le k-1$, and using Lemma 1.3.3, we obtain

$$\begin{aligned} 0 = \langle A\delta_k, \delta_i \rangle &= -h_k \langle A\nabla f(x_k), \delta_i \rangle + \sum_{j=0}^{k-1} \lambda^{(j)} \langle A\delta_j, \delta_i \rangle \\ &= -h_k \langle A\nabla f(x_k), \delta_i \rangle + \lambda^{(i)} \langle A\delta_i, \delta_i \rangle \\ &= -h_k \langle \nabla f(x_k), A\delta_i \rangle + \lambda^{(i)} \langle A\delta_i, \delta_i \rangle \\ &= -h_k \langle \nabla f(x_k), \nabla f(x_{i+1}) - \nabla f(x_i) \rangle + \lambda^{(i)} \langle A\delta_i, \delta_i \rangle. \end{aligned}$$

Hence, in view of Lemma 1.3.2, $\lambda_i = 0$ for $i < k-1$. For $i = k-1$, we have

$$\lambda^{(k-1)} = \frac{h_k \|\nabla f(x_k)\|^2}{\langle A\delta_{k-1}, \delta_{k-1} \rangle} = \frac{h_k \|\nabla f(x_k)\|^2}{\langle \nabla f(x_k) - \nabla f(x_{k-1}), \delta_{k-1} \rangle}.$$

Thus, $x_{k+1} = x_k - h_k p_k$, where

$$p_k = \nabla f(x_k) - \frac{\|\nabla f(x_k)\|^2 \delta_{k-1}}{\langle \nabla f(x_k) - \nabla f(x_{k-1}), \delta_{k-1} \rangle} = \nabla f(x_k) - \frac{\|\nabla f(x_k)\|^2 p_{k-1}}{\langle \nabla f(x_k) - \nabla f(x_{k-1}), p_{k-1} \rangle}$$

since $\delta_{k-1} = -h_{k-1} p_{k-1}$ by definition of the directions $\{p_k\}$.

Note that we managed to write down the Conjugate Gradient Method in terms of the gradients of the objective function $f(\cdot)$. This provides us with the possibility of *formally* applying this scheme to minimize a general nonlinear function. Of course, such an extension destroys all properties of the process which are specific for quadratic functions. However, in a neighborhood of a strict local minimum, the objective function is close to quadratic. Therefore, asymptotically this method should be fast.

Let us present a general scheme of the Conjugate Gradient Method for minimizing a general nonlinear function.

```
Conjugate Gradient Method

0. Let $x_0 \in \mathbb{R}^n$. Compute $f(x_0), \nabla f(x_0)$. Set $p_0 = \nabla f(x_0)$.

1. kth iteration ($k \geq 0$).
        (a) Find $x_{k+1} = x_k - h_k p_k$(by "exact" line search).
        (b) Compute $f(x_{k+1})$ and $\nabla f(x_{k+1})$.
        (c) Compute the coefficient $\beta_k$.
        (d) Define $p_{k+1} = \nabla f(x_{k+1}) - \beta_k p_k$.
```

In this scheme, we have not yet specified the coefficient $\beta_k$. In fact, there exist many different formulas for this coefficient. All of them give the same results on quadratic functions. However, in the general nonlinear case, they generate different sequences. Let us present the three most popular expressions.

1. *Dai–Yuan*: $\beta_k = \frac{\|\nabla f(x_{k+1})\|^2}{\langle \nabla f(x_{k+1}) - \nabla f(x_k), p_k \rangle}$.
2. *Fletcher–Rieves*: $\beta_k = -\frac{\|\nabla f(x_{k+1})\|^2}{\|\nabla f(x_k)\|^2}$.
3. *Polak–Ribbiere*: $\beta_k = -\frac{\langle \nabla f(x_{k+1}), \nabla f(x_{k+1}) - \nabla f(x_k) \rangle}{\|\nabla f(x_k)\|^2}$.

Recall that in the quadratic case, the Conjugate Gradient Method terminates in $n$ iterations (or less). Algorithmically, this means that $p_n = 0$. In the general nonlinear case, this is not true. However, after $n$ iterations, this direction loses its interpretation. Therefore, in all practical schemes, there exists a *restarting* strategy, which at some moment sets $\beta_k = 0$ (usually after every $n$ iterations). This ensures the global convergence of the process (since we have the usual gradient step just after the restart, and all other iterations decrease the value of the objective function). In a neighborhood of a strict minimum, the conjugate gradient schemes demonstrate a local $n$-step quadratic convergence:

$$\| x_n - x^* \| \leq \text{const} \cdot \| x_0 - x^* \|^2 .$$

Note that this local convergence is slower than that of the variable metric methods. However, the conjugate gradient methods have the advantage of cheap iteration. As far as the global convergence is concerned, these schemes, in general, are not better than the simplest Gradient Method.

### *1.3.3 Constrained Minimization*

Let us discuss now the main ideas underlying the methods of optimization with functional constraints. The problem we consider here is as follows:

$$\begin{aligned} f_0(x) \;&\to\; \min_{x\in Q}, \\ f_j(x) &\leq 0,\;\; j = 1\dots m, \end{aligned} \tag{1.3.5}$$

where $Q$ is a simple closed set in $\mathbb{R}^n$, and the functional components $f_0(\cdot), \dots, f_m(\cdot)$ are continuous functions. Since these components are general nonlinear functions, we cannot expect this problem to be easier than an unconstrained minimization problem. Indeed, even the standard difficulties with stationary points, which we have in Unconstrained Minimization, appear in (1.3.5) in a much stronger form. Note that a stationary point of this problem (whatever its definition is) can be infeasible for the system of functional constraints. Hence, any minimization scheme attracted by such a point fails even to find a feasible solution of (1.3.5).

Therefore, the following reasoning looks quite convincing.

1. We have efficient methods for unconstrained minimization.[4]
2. Unconstrained minimization is simpler than constrained minimization.[5]
3. Therefore, let us try to approximate a solution of problem (1.3.5) by a sequence of solutions to some auxiliary *unconstrained* minimization problems.

This philosophy is implemented by the schemes of *Sequential Unconstrained Minimization*. There are three main groups of such methods.

- Lagrangian relaxation methods.
- Penalty function methods.
- Barrier methods.

Let us describe the main ideas of these approaches.

#### 1.3.3.1 Lagrangian Relaxation

This approach is based on the following fundamental *Minimax Principle*.

[4] In fact, this is not absolutely true. We will see that, in order to apply the unconstrained minimization methods to solve constrained problems, we need to be able to find a global minimum of some auxiliary problem, and we have already seen (Example 1.2.2) that this could be difficult.

[5] We are not going to discuss the correctness of this statement for general nonlinear problems. We just prevent the reader from extending it to other problem classes. In the following chapters, we will see that this statement is valid only up to a certain point.

**Theorem 1.3.1** *Let the function $F(x,\lambda)$ be defined for $x \in Q_1 \subseteq \mathbb{R}^n$ and $\lambda \in Q_2 \subseteq \mathbb{R}^m$, where both $Q_1$ and $Q_2$ are nonempty. Then,*

$$\sup_{\lambda \in Q_2} \inf_{x \in Q_1} F(x,\lambda) \le \inf_{x \in Q_1} \sup_{\lambda \in Q_2} F(x,\lambda). \tag{1.3.6}$$

*Proof* Indeed, for arbitrary $x \in Q_1$ and $\lambda \in Q_2$, we have

$$F(x,\lambda) \le \sup_{\xi \in Q_2} F(x,\xi).$$

Since this inequality is valid for all $x \in Q_1$, we conclude that

$$\inf_{x \in Q_1} F(x,\lambda) \le \inf_{x \in Q_1} \sup_{\xi \in Q_2} F(x,\xi).$$

It remains to note that this inequality is valid for all $\lambda \in Q_2$. □

Let us apply this principle to problem (1.3.5). Note that

$$f^* = \inf_{x \in Q} \{f_0(x) : \ f_j(x) \le 0, \ j = 1,\dots,m\}$$

$$= \inf_{x \in Q} \sup_{\lambda \in \mathbb{R}^m_+} \left\{ \mathscr{L}(x,\lambda) \overset{\text{def}}{=} f_0(x) + \langle \lambda, f(x) \rangle \right\},$$

where $f(x) = (f_1(x),\dots,f_m(x))^T$, $\mathbb{R}^m_+ = \{\lambda \in \mathbb{R}^m : \ \lambda^{(j)} \ge 0, \ j = 1,\dots,m\}$ is a *positive orthant*, and $\mathscr{L}(x,\lambda)$ is the *Lagrange function*, or *Lagrangian*, of problem (1.3.5). Let

$$\psi(\lambda) = \inf_{x \in Q} \mathscr{L}(x,\lambda),$$

$$\text{dom}\psi = \{\lambda \in \mathbb{R}^m : \ \psi(\lambda) > -\infty\}, \tag{1.3.7}$$

$$X^*(\lambda) = \text{Arg} \inf_{x \in Q} \mathscr{L}(x,\lambda),$$

where $X^*(\lambda)$ is the set of *global solutions* of the corresponding minimization problem. Note that at some $\lambda \in \mathbb{R}^m$ the value of function $\psi$ can be $-\infty$. For us, it is important to have $\text{dom}\psi \bigcap \mathbb{R}^m_+ \neq \emptyset$. For simplicity, we assume that, for all $\lambda$ from this set, $X^*(\lambda) \neq \emptyset$.

Thus, we come to the following *Lagrange dual problem*:

$$f_* \overset{\text{def}}{=} \sup_{\lambda} \left\{ \psi(\lambda) : \ \lambda \in \text{dom}\psi \bigcap \mathbb{R}^m_+ \right\} \overset{(1.3.6)}{\le} f^*. \tag{1.3.8}$$

Note that the objective function of the dual problem is very special. Indeed, for any two vectors $\lambda_1, \lambda_2$ from $\mathrm{dom}\psi$, and any $x_1 \in X^*(\lambda_1)$, $x_2 \in X^*(\lambda_2)$ we have

$$\begin{aligned} \psi(\lambda_2) &= f_0(x_2) + \sum_{j=1}^m \lambda_2^{(j)} f_j(x_2) \;\le\; f_0(x_1) + \sum_{j=1}^m \lambda_2^{(j)} f_j(x_1) \\ &= \psi(\lambda_1) + \langle f(x_1), \lambda_2 - \lambda_1 \rangle. \end{aligned} \tag{1.3.9}$$

This means that the function $\psi$ is *concave*, and (1.3.8) is a *convex* optimization problem. Such problems can be efficiently solved by numerical schemes (see Chap. 3), provided that for any $\lambda \in \mathrm{dom}\psi$ we are able to compute the vector $f(x(\lambda))$, where $x(\lambda)$ is one of the global solutions of problem (1.3.7).

Note that the dual problem (1.3.8) is not completely equivalent to the primal problem (1.3.5). Very often, we can observe the situation $f_* < f^*$ (the so-called *nonzero duality gap*). This is the reason why the problem (1.3.8) is often called the *Lagrangian relaxation* of problem (1.3.5).

Conditions for a zero duality gap, $f_* = f^*$, are usually quite restrictive and require convexity of all elements of problem (1.3.5). We will see many instances of such problems in Part II of this book. Here, we give a sufficient condition, which is sometimes useful.

**Theorem 1.3.2 (Certificate of Global Optimality)** *Let $\lambda_*$ be an optimal solution to problem (1.3.8). Assume that for some positive $\epsilon$ we have*

$$\Delta_\epsilon^+(\lambda^*) \overset{\mathrm{def}}{=} \{\lambda \in \mathbb{R}^m_+ : \; \|\lambda - \lambda_*\| \le \epsilon\} \subseteq \mathrm{dom}\psi.$$

*Let the vector $x(\lambda) \in X^*(\lambda)$, $\lambda \ne \lambda_*$, be uniquely defined and the following limit exist*

$$x^* = \lim_{\substack{\lambda \to \lambda_*, \\ \lambda \in \Delta_\epsilon^+(\lambda_*)}} x(\lambda).$$

*If $x^* \in X^*(\lambda_*)$, then it is an optimal global solution to problem (1.3.5).*

*Proof* Let $g(\lambda) = f(x(\lambda))$. Let $I^* = \{j : \lambda_*^{(j)} > 0\}$. Choosing $j \in I^*$ and $\epsilon > 0$ ensuring $\lambda_* \pm \epsilon e_j \in \mathrm{dom}\psi \bigcap \mathbb{R}^m_+$, we get

$$\psi(\lambda_*) \overset{(1.3.9)}{\le} \psi(\lambda_* + \epsilon e_j) + \langle g(\lambda_* + \epsilon e_j), -\epsilon e_j \rangle \;\le\; \psi(\lambda_*) + \langle g(\lambda_* + \epsilon e_j), -\epsilon e_j \rangle,$$

$$\psi(\lambda_*) \overset{(1.3.9)}{\le} \psi(\lambda_* - \epsilon e_j) + \langle g(\lambda_* - \epsilon e_j), \epsilon e_j \rangle \;\le\; \psi(\lambda_*) + \langle g(\lambda_* - \epsilon e_j), \epsilon e_j \rangle,$$

Thus, we have

$$\langle g(\lambda_* + \epsilon e_j), e_j \rangle \le 0 \;\le\; \langle g(\lambda_* - \epsilon e_j), e_j \rangle.$$

Taking the limit in both inequalities as $\epsilon \to 0$, we obtain $f_j(x^*) = 0$.

Similarly, if $j \notin I^*$, we can take $\epsilon$ small enough to have $\lambda_* + \epsilon e_j \in \mathrm{dom}\psi$. Then,

$$\begin{aligned} \psi(\lambda_*) &\overset{(1.3.9)}{\le} \psi(\lambda_* + \epsilon e_j) + \langle g(\lambda_* + \epsilon e_j), -\epsilon e_j \rangle \\ &\le \psi(\lambda_*) + \langle g(\lambda_* + \epsilon e_j), -\epsilon e_j \rangle. \end{aligned}$$

Hence, $\langle g(\lambda_* + \epsilon e_j), e_j \rangle \le 0$. Taking in this inequality the limit as $\epsilon \to 0$, we get $f_j(x^*) \le 0$.

Thus, the point $x^*$ is feasible for the problem (1.3.5), and

$$\lambda_*^{(j)} f_j(x^*) = 0, \quad j = 1, \ldots, m. \tag{1.3.10}$$

Therefore, we obtain

$$f_0(x^*) \overset{(1.3.10)}{=} f_0(x^*) + \sum_{j=1}^{m} \lambda_*^{(j)} f_j(x^*) \;=\; \psi(\lambda_*) \overset{(1.3.8)}{\le} f^*.$$

□

*Remark 1.3.1* The equality constraints in problem (1.3.5) can be treated in a similar way. The only difference is that in the dual problem (1.3.8), the corresponding Lagrange multipliers do not have sign restrictions. At the same time, the statement of Theorem 1.3.2 remains valid.

Let us show how this condition works in some simple situations.

*Example 1.3.2* Let us choose in the problem (1.3.5) $Q = \mathbb{R}^2$, and

$$f_0(x) = \tfrac{1}{2}\|x - \bar{e}_2\|^2, \quad f_1(x) = x^{(1)} - \tfrac{1}{2}(x^{(2)})^2,$$

where $\bar{e}_2 = (1, 1)^T$. Then, we can form the Lagrangian

$$\mathscr{L}(x, \lambda) = \tfrac{1}{2}\|x - \bar{e}_2\|^2 + \lambda \left[ x^{(1)} - \tfrac{1}{2}(x^{(2)})^2 \right],$$

and define $\psi(\lambda) = \inf\limits_{x \in \mathbb{R}^2} \mathscr{L}(x, \lambda)$. It is clear that $\mathrm{dom}\psi = (-\infty, 1)$, and for any feasible $\lambda$, the point $x(\lambda)$ can be found from the following equations:

$$x^{(1)}(\lambda) - 1 + \lambda = 0,$$

$$x^{(2)}(\lambda) - 1 - \lambda x^{(2)}(\lambda) = 0.$$

Thus, $x^{(1)}(\lambda) = 1 - \lambda$, and $x^{(2)}(\lambda) = \frac{1}{1-\lambda}$. Substituting this point into the Lagrangian, we obtain

$$\psi(\lambda) = \lambda - \tfrac{1}{2}\lambda^2 - \tfrac{1}{2(1-\lambda)} + \tfrac{1}{2}.$$

The maximum of $\psi$ is attained at $\lambda_* = 1 - \left(\frac{1}{2}\right)^{1/3}$. Since the trajectory $x(\lambda)$ is uniquely defined and continuous on the domain $\mathrm{dom}\psi$, by Theorem 1.3.2 we conclude that the point $x(\lambda_*) = \left(2^{-1/3}, 2^{1/3}\right)$ is the global optimal solution of our problem. □

We consider another example of application of Theorem 1.3.2 in Sect. 4.1.4.

#### 1.3.3.2 Penalty Functions

**Definition 1.3.1** A continuous function $\Phi(\cdot)$ is called a *penalty function* for a closed set $\mathscr{F} \subset \mathbb{R}^n$ if

- $\Phi(x) = 0$ for any $x \in \mathscr{F}$,
- $\Phi(x) > 0$ for any $x \notin \mathscr{F}$.

Sometimes, a penalty function is called just a *penalty* for the set $\mathscr{F}$. The main property of penalty functions is as follows.

> If $\Phi_1(\cdot)$ is a penalty for $\mathscr{F}_1$ and $\Phi_2(\cdot)$ is a penalty for $\mathscr{F}_2$, then $\Phi_1(\cdot) + \Phi_2(\cdot)$ is a penalty for the intersection $\mathscr{F}_1 \bigcap \mathscr{F}_2$.

Let us give several examples of such functions.

*Example 1.3.3* Define $(a)_+ = \max\{a, 0\}$, $a \in \mathbb{R}$. Let $f_1(\cdot), \ldots, f_m(\cdot)$ be continuous functions, and

$$\mathscr{F} = \{x \in \mathbb{R}^n \mid f_j(x) \le 0,\ j = 1 \ldots m\}.$$

Then, the following functions are penalties for $\mathscr{F}$:

1. *Quadratic penalty*: $\Phi(x) = \sum\limits_{j=1}^{m} (f_j(x))_+^2$.
2. *Nonsmooth penalty*: $\Phi(x) = \sum\limits_{j=1}^{m} (f_j(x))_+$.

The reader can easily continue the list. □

Let us present the general scheme of the Penalty Function Method as applied to problem (1.3.5).

**Penalty Function Method**

```
0. Choose $x_0 \in Q$. Choose a sequence of penalty coefficients:
       $0 < t_k < t_{k+1}$ and $t_k \to \infty$.
1. $k$th iteration ($k \geq 0$).
   Find $x_{k+1} = \arg\min_{x \in Q}\{f_0(x) + t_k \Phi(x)\}$ using $x_k$ as starting point.
```

It is easy to prove the convergence of this scheme assuming that $x_{k+1}$ is a global minimum of the auxiliary function.[6] Define

$$\Psi_k(x) = f_0(x) + t_k \Phi(x), \quad \Psi_k^* = \min_{x \in Q} \Psi_k(x) = \Psi_k(x_{k+1}).$$

($\Psi_k^*$ is the *global* optimal value of $\Psi_k(\cdot)$). Let $x^*$ be a global solution to (1.3.5).

**Theorem 1.3.3** *Let there exist a value* $\bar{t} > 0$ *such that the set*

$$S = \{x \in \mathbb{R}^n \mid f_0(x) + \bar{t}\Phi(x) \leq f_0(x^*)\}$$

*is bounded. Then*

$$\lim_{k\to\infty} f_0(x_k) = f_0(x^*), \quad \lim_{k\to\infty} \Phi(x_k) = 0.$$

*Proof* Note that $\Psi_k^* \leq \Psi_k(x^*) = f_0(x^*)$. At the same time, for any $x \in Q$ we have $\Psi_{k+1}(x) \geq \Psi_k(x)$. Therefore $\Psi_{k+1}^* \geq \Psi_k^*$. Thus, there exists a limit

$$\lim_{k\to\infty} \Psi_k^* \equiv \Psi^* \leq f_0(x^*).$$

If $t_k > \bar{t}$ then

$$f_0(x_{k+1}) + \bar{t}\Phi(x_{k+1}) \leq f_0(x_{k+1}) + t_k \Phi(x_{k+1}) = \Psi_k^* \leq f_0(x^*).$$

Therefore, $x_k \in S$ for $k$ large enough. Hence, the sequence $\{x_k\}$ has limit points. Since $\lim_{k\to\infty} t_k = +\infty$, for any such point $x_*$ we have $\Phi(x_*) = 0$. Thus, $x_* \in \mathscr{F}$ and $f_0(x_*) \leq f_0(x^*)$. Consequently, $f_0(x_*) = f_0(x^*)$. □

Note that this result is very general, but not too informative. There are still many questions which should be answered. For example, we do not know what

[6] If we assume that it is a strict local minimum, then the results are much weaker.

kind of penalty functions we should use. What should be the rules for choosing the penalty coefficients? What should be the accuracy for solving the auxiliary problems? In fact, all these questions are difficult to address in the framework of general Nonlinear Optimization. Traditionally, they are redirected to computational practice.

#### 1.3.3.3 Barrier Functions

Let us look at Barrier Methods.

**Definition 1.3.2** Let $\mathscr{F}$ be a closed set in $\mathbb{R}^n$ with nonempty interior. A continuous function $F(\cdot)$ is called a *barrier function* for $\mathscr{F}$ if $F(x) \to \infty$ as $x$ approaches the boundary of this set.

Sometimes a barrier function is called a *barrier* for short. Similarly to penalty functions, the barriers possess the following property.

> If $F_1(\cdot)$ is a barrier for $\mathscr{F}_1$ and $F_2(\cdot)$ is a barrier for $\mathscr{F}_2$, then $F_1(\cdot) + F_2(\cdot)$ is a barrier for the intersection $\mathscr{F}_1 \bigcap \mathscr{F}_2$ provided that its interior is nonexpty.

In order to apply the barrier approach, problem (1.3.5) must satisfy the *Slater condition:*

$$\exists \bar{x} \in \mathbb{R}^n : \quad f_j(\bar{x}) < 0, \quad j = 1 \ldots m. \tag{1.3.11}$$

Let us look at some examples of barrier functions.

*Example 1.3.4* Let $f_1(\cdot), \ldots, f_m(\cdot)$ be continuous functions and $\mathscr{F} = \{x \in \mathbb{R}^n \mid f_j(x) \le 0,\ j = 1 \ldots m\}$. Then all the functions below are barriers for $\mathscr{F}$:

1. *Power-function* barrier: $F(x) = \sum\limits_{j=1}^{m} \frac{1}{(-f_j(x))^p}$, $p \ge 1$.
2. *Logarithmic* barrier: $F(x) = -\sum\limits_{j=1}^{m} \ln(-f_j(x))$.
3. *Exponential* barrier: $F(x) = \sum\limits_{j=1}^{m} \exp\left(\frac{1}{-f_j(x)}\right)$.

The reader can easily extend this list. □

Let $\mathscr{F}_0 = Q \bigcap \mathrm{int}\mathscr{F}$ and let $F$ be a barrier for $\mathscr{F}$. The general scheme of the Barrier Method is as follows.

**Barrier Function Method**

**0.** Choose $x_0 \in \mathscr{F}_0$ and a sequence of penalty coefficients:

$$0 < t_k < t_{k+1} \text{ and } t_k \to \infty.$$

**1.** ***k*th iteration** ($k \geq 0$).
Find $x_{k+1} = \arg\min\limits_{x \in \mathscr{F}_0} \left\{ f_0(x) + \frac{1}{t_k} F(x) \right\}$ using $x_k$ as the starting point.

Let us prove the convergence of this method assuming that $x_{k+1}$ is a global minimum of the auxiliary function. Define

$$\Psi_k(x) = f_0(x) + \tfrac{1}{t_k} F(x), \quad \Psi_k^* = \min_{x \in \mathscr{F}_0} \Psi_k(x),$$

($\Psi_k^*$ is the global optimal value of $\Psi_k(\cdot)$) and let $f^*$ be the optimal value of the problem (1.3.5).

**Theorem 1.3.4** *Let the barrier $F(\cdot)$ be bounded below on $\mathscr{F}_0$. Then*

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

*Proof* Let $F(x) \geq F^*$ for all $x \in \mathscr{F}_0$. For arbitrary $\bar{x} \in \mathscr{F}_0$ we have

$$\limsup_{k \to \infty} \Psi_k^* \leq \lim_{k \to \infty} \left[ f_0(\bar{x}) + \tfrac{1}{t_k} F(\bar{x}) \right] = f_0(\bar{x}).$$

Therefore $\limsup\limits_{k \to \infty} \Psi_k^* \leq f^*$. On the other hand,

$$\Psi_k^* = \min_{x \in \mathscr{F}_0} \left\{ f_0(x) + \tfrac{1}{t_k} F(x) \right\} \geq \inf_{x \in \mathscr{F}_0} \left\{ f_0(x) + \tfrac{1}{t_k} F^* \right\} = f^* + \tfrac{1}{t_k} F^*.$$

Thus, $\lim\limits_{k \to \infty} \Psi_k^* = f^*$. □

As with Penalty Function Methods, many questions need to be answered. We do not know how to find the starting point $x_0$ and how to choose the best barrier function. We do not know theoretically justified rules for updating the penalty coefficients and the acceptable accuracy of the solution for the auxiliary problems. Finally, we have no ideas about the efficiency estimates of this process. And the reason is not in the lack of theory. Our problem (1.3.5) is still too complicated. We will see that all the questions above get *precise* answers in the framework of Convex Optimization (see Chap. 5).

We have finished our brief presentation of general Nonlinear Optimization. It was very short indeed, and there are many interesting theoretical topics that we did not mention. The reason is that the main goal of this book is to describe the areas of Optimization where we can obtain clear and comprehensive results on the performance of numerical methods. Unfortunately, general Nonlinear Optimization is just too complicated to fit the goal. However, it was impossible to skip this field since a lot of basic ideas underlying Convex Optimization have their origin in the general theory of Nonlinear Optimization. The Gradient Method and Newton's Method, Sequential Unconstrained Minimization and Barrier Functions were originally developed and used for general optimization problems. But only the framework of Convex Optimization allows these ideas to get their real power. In the following chapters of this book, we will see many examples of the second birth of these old ideas.

# Chapter 2
# Smooth Convex Optimization

In this chapter, we study the complexity of solving optimization problems formed by differentiable convex components. We start by establishing the main properties of such functions and deriving the lower complexity bounds, which are valid for all natural optimization methods. After that, we prove the worst-case performance guarantees for the Gradient Method. Since these bounds are quite far from the lower complexity bounds, we develop a special technique, based on the notion of estimating sequences, which allows us to justify the Fast Gradient Methods. These methods appear to be optimal for smooth convex problems. We also obtain performance guarantees for these methods targeting on generating points with small norm of the gradient. In order to treat problems with set constraints, we introduce the notion of a Gradient Mapping. This allows an automatic extension of methods for unconstrained minimization to the constrained case. In the last section, we consider methods for solving smooth optimization problems, defined by several functional components.

## 2.1 Minimization of Smooth Functions

(Smooth convex functions; Lower complexity bounds for $\mathscr{F}_L^{\infty,1}(\mathbb{R}^n)$; Strongly convex functions; Lower complexity bounds for $\mathscr{S}_{\mu,L}^{\infty,1}(\mathbb{R}^n)$; The Gradient Method.)

### *2.1.1 Smooth Convex Functions*

In this section, we consider the unconstrained minimization problem

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

Y. Nesterov, *Lectures on Convex Optimization*, Springer Optimization and Its Applications 137, https://doi.org/10.1007/978-3-319-91578-4_2

where the objective function $f(\cdot)$ is smooth enough. Recall that in the previous chapter we were trying to solve this problem under very weak assumptions on the function $f$. We have seen that in this general situation we cannot do too much: It is impossible to guarantee convergence even to a local minimum and it is impossible to get acceptable bounds on the global performance of minimization schemes, etc. Let us try to introduce some reasonable assumptions on the function $f$ in order to make our problem more tractable. For that, let us try to specify the desired properties of a hypothetical class of differentiable functions $\mathscr{F}$ we want to work with.

From the results of the previous chapter, we could come to the conclusion that the main reason for our troubles is the weakness of the first-order optimality condition (Theorem 1.2.1). Indeed, we have seen that, in general, the Gradient Method converges only to a stationary point of the function $f$ (see inequality (1.2.22) and Example 1.2.2). Therefore, the first additional property we definitely need is as follows.

**Assumption 2.1.1** *For any $f \in \mathscr{F}$, the first-order optimality condition is sufficient for a point to be a global solution to (2.1.1).*

Further, the main feature of any tractable functional class $\mathscr{F}$ is the possibility to verify the inclusion $f \in \mathscr{F}$ in a simple way. Usually, this is ensured by a set of *basic elements* of the class, endowed with a list of possible *operations* with elements of $\mathscr{F}$, which keep the result in the class (such operations are called *invariant*). An excellent example of such a construction is the class of differentiable functions. In order to check whether a function is differentiable or not, we just need to look at its analytical representation.

We do not want to restrict our class too much. Therefore, let us introduce only one invariant operation for the hypothetical class $\mathscr{F}$.

**Assumption 2.1.2** *If $f_1, f_2 \in \mathscr{F}$ and $\alpha, \beta \geq 0$, then $\alpha f_1 + \beta f_2 \in \mathscr{F}$.*

The reason for the restriction on the sign of coefficients in this assumption is evident: We would like to see $x^2$ in our class, but the function $-x^2$ is not suitable for our goals.

Finally, let us add to $\mathscr{F}$ some basic elements.

**Assumption 2.1.3** *Any linear function $\ell(x) = \alpha + \langle a, x \rangle$ belongs to $\mathscr{F}$.*[1]

Note that the linear function $\ell(\cdot)$ perfectly fits Assumption 2.1.1. Indeed, $\nabla \ell(x) = 0$ implies that this function is constant, and any point in $\mathbb{R}^n$ is its global minimum.

It turns out that we have already introduced enough assumptions to specify our functional class. Consider $f \in \mathscr{F}$. Let us fix some $x_0 \in \mathbb{R}^n$ and consider the function

$$\phi(y) = f(y) - \langle \nabla f(x_0), y \rangle.$$

[1] This is not a description of the whole set of basic elements. We just say that we want to have all linear functions in our class.

Then $\phi \in \mathscr{F}$ in view of Assumptions 2.1.2 and 2.1.3. Note that

$$\nabla\phi(y) \mid_{y=x_0} = \nabla f(x_0) - \nabla f(x_0) = 0.$$

Therefore, in view of Assumption 2.1.1, $x_0$ is the global minimum of function $\phi$, and for any $y \in \mathbb{R}^n$ we have

$$\phi(y) \geq \phi(x_0) = f(x_0) - \langle \nabla f(x_0), x_0 \rangle.$$

Hence, $f(y) \geq f(x_0) + \langle \nabla f(x_0), y - x_0 \rangle$.

This inequality is very well known in Optimization Theory. It defines the class of differentiable *convex* functions. Such functions may have a restricted domain. However, this domain must always be *convex*.

**Definition 2.1.1** A set $Q \subseteq \mathbb{R}^n$ is called *convex* if for any $x, y \in Q$ and $\alpha$ from $[0, 1]$ we have

$$\alpha x + (1 - \alpha) y \in Q.$$

Thus, a convex set contains the whole segment $[x, y]$ provided that the end points $x$ and $y$ belong to the set.

**Definition 2.1.2** A continuously differentiable function $f(\cdot)$ is called *convex* on a convex set $Q$ (notation $f \in \mathscr{F}^1(Q)$) if for any $x, y \in Q$ we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle. \tag{2.1.2}$$

If $-f(\cdot)$ is convex, we call $f(\cdot)$ *concave*.

In what follows we also consider the classes of convex functions $\mathscr{F}_L^{k,l}(Q)$ where the indices have the same meaning as for the classes $C_L^{k,l}(Q)$.

Let us check our assumptions, which now become the *properties* of the functional class.

**Theorem 2.1.1** *If $f \in \mathscr{F}^1(\mathbb{R}^n)$ and $\nabla f(x^*) = 0$ then $x^*$ is the global minimum of $f(\cdot)$ on $\mathbb{R}^n$.*

*Proof* In view of inequality (2.1.2), for any $x \in \mathbb{R}^n$ we have

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle = f(x^*). \quad \square$$

Thus, we get what we want in Assumption 2.1.1. Let us check Assumption 2.1.2.

**Lemma 2.1.1** *If $f_1$ and $f_2$ belong to $\mathscr{F}^1(Q)$ and $\alpha, \beta \geq 0$, then the function $f = \alpha f_1 + \beta f_2$ also belongs to $\mathscr{F}^1(Q)$.*

*Proof* For any $x, y \in Q$, we have

$$f_1(y) \geq f_1(x) + \langle \nabla f_1(x), y - x \rangle,$$

$$f_2(y) \geq f_2(x) + \langle \nabla f_2(x), y - x \rangle.$$

It remains to multiply the first equation by $\alpha$, the second one by $\beta$, and add the results. □

Thus, for differentiable functions our hypothetical class coincides with the class of convex functions. Let us present their main properties.

The next statement significantly increases our possibilities in *constructing* the convex functions.

**Lemma 2.1.2** *If $f \in \mathscr{F}^1(Q)$, $b \in \mathbb{R}^m$ and $A : \mathbb{R}^n \to \mathbb{R}^m$ then*

$$\phi(x) = f(Ax + b) \in \mathscr{F}^1(\hat{Q}), \quad \hat{Q} = \{x \in \mathbb{R}^n : \; Ax + b \in Q\}.$$

*Proof* Indeed, let $x, y \in Q$. Define $\bar{x} = Ax + b$, $\bar{y} = Ay + b$. Since

$$\nabla \phi(x) = A^T \nabla f(Ax + b),$$

we have

$$\begin{aligned} \phi(y) = f(\bar{y}) \geq & \; f(\bar{x}) + \langle \nabla f(\bar{x}), \bar{y} - \bar{x} \rangle \\ = & \; \phi(x) + \langle \nabla f(\bar{x}), A(y - x) \rangle \\ = & \; \phi(x) + \langle A^T \nabla f(\bar{x}), y - x \rangle \\ = & \; \phi(x) + \langle \nabla \phi(x), y - x \rangle. \qquad \square \end{aligned}$$

In order to make the verification of the inclusion $f \in \mathscr{F}^1(Q)$ easier, let us provide several equivalent definitions of this class.

**Theorem 2.1.2** *A continuously differentiable function $f$ belongs to the class $\mathscr{F}^1(Q)$ if and only if for any $x, y \in Q$ and $\alpha \in [0, 1]$ we have*[2]

$$f(\alpha x + (1 - \alpha) y) \leq \alpha f(x) + (1 - \alpha) f(y). \tag{2.1.3}$$

[2] Note that inequality (2.1.3) without the assumption of differentiability of $f$ serves as a definition of *general convex* functions. We will study these functions in detail in Chap. 3.

*Proof* Define $x_\alpha = \alpha x + (1-\alpha)y$. Let $f \in \mathscr{F}^1(Q)$. Then

$$f(x_\alpha) \leq f(y) - \langle \nabla f(x_\alpha), y - x_\alpha \rangle = f(y) - \alpha \langle \nabla f(x_\alpha), y - x \rangle,$$

$$f(x_\alpha) \leq f(x) - \langle \nabla f(x_\alpha), x - x_\alpha \rangle = f(x) + (1-\alpha) \langle \nabla f(x_\alpha), y - x \rangle.$$

Multiplying the first inequality by $(1-\alpha)$, the second one by $\alpha$, and adding the results, we get (2.1.3).

Let (2.1.3) be true for all $x, y \in Q$ and $\alpha \in [0, 1]$. Let us choose some $\alpha \in [0, 1)$. Then

$$\begin{aligned} f(y) &\geq \tfrac{1}{1-\alpha}[f(x_\alpha) - \alpha f(x)] = f(x) + \tfrac{1}{1-\alpha}[f(x_\alpha) - f(x)] \\ &= f(x) + \tfrac{1}{1-\alpha}[f(x + (1-\alpha)(y-x)) - f(x)]. \end{aligned}$$

Letting $\alpha$ tend to 1, we get (2.1.2). □

**Theorem 2.1.3** *A continuously differentiable function $f$ belongs to the class $\mathscr{F}^1(Q)$ if and only if for any $x, y \in Q$ we have*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0. \tag{2.1.4}$$

*Proof* Let $f$ be a convex continuously differentiable function. Then

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

Adding these inequalities, we get (2.1.4).

Let (2.1.4) hold for all $x, y \in Q$. Define $x_\tau = x + \tau(y-x) \in Q$. Then

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \langle \nabla f(x + \tau(y-x)), y - x \rangle d\tau \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x_\tau) - \nabla f(x), y - x \rangle d\tau \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \tfrac{1}{\tau} \langle \nabla f(x_\tau) - \nabla f(x), x_\tau - x \rangle d\tau \\ &\geq f(x) + \langle \nabla f(x), y - x \rangle. \end{aligned}$$ □

Sometimes it is more convenient to work with functions from a smaller class $\mathscr{F}^2(Q) \subset \mathscr{F}^1(Q)$.

**Theorem 2.1.4** *Let $Q$ be an open set. A twice continuously differentiable function $f$ belongs to the class $\mathscr{F}^2(Q)$ if and only if for any $x \in Q$ we have*

$$\nabla^2 f(x) \succeq 0. \tag{2.1.5}$$

*Proof* Let a function $f$ from $C^2(Q)$ be convex and $s \in \mathbb{R}^n$. Let $x_\tau = x + \tau s \in Q$ for $\tau > 0$ small enough. Then, in view of (2.1.4), we have

$$0 \leq \tfrac{1}{\tau^2}\langle \nabla f(x_\tau) - \nabla f(x), x_\tau - x\rangle = \tfrac{1}{\tau}\langle \nabla f(x_\tau) - \nabla f(x), s\rangle$$

$$= \tfrac{1}{\tau}\int\limits_0^\tau \langle \nabla^2 f(x + \lambda s)s, s\rangle d\lambda,$$

and we get (2.1.5) by letting $\tau$ tend to zero.

Let (2.1.5) hold for all $x \in Q$. Then for $y \in Q$ we have

$$f(y) = f(x) + \langle \nabla f(x), y - x\rangle + \int\limits_0^1 \int\limits_0^\tau \langle \nabla^2 f(x + \lambda(y - x))(y - x), y - x\rangle d\lambda d\tau$$

$$\geq f(x) + \langle \nabla f(x), y - x\rangle. \qquad \square$$

Let us look at some examples of differentiable convex functions on $\mathbb{R}^n$.

*Example 2.1.1*

1. Every linear function $f(x) = \alpha + \langle a, x\rangle$ is convex.
2. Let matrix $A$ be symmetric and positive semidefinite. Then the quadratic function

$$f(x) = \alpha + \langle a, x\rangle + \frac{1}{2}\langle Ax, x\rangle$$

is convex (since $\nabla^2 f(x) = A \succeq 0$).

3. The following functions of one variable belong to $\mathscr{F}^1(\mathbb{R})$:

$$f(x) = e^x,$$

$$f(x) = \mid x \mid^p, \quad p > 1,$$

$$f(x) = \tfrac{x^2}{1-|x|},$$

$$f(x) = \mid x \mid - \ln(1+ \mid x \mid).$$

We can check this using Theorem 2.1.4. Therefore, functions arising in *Geometric Optimization* (see Sect. 5.4.8), like

$$f(x) = \sum_{i=1}^{m} e^{\alpha_i + \langle a_i, x \rangle},$$

are convex (see Lemma 2.1.2). Similarly, functions arising in $\ell_p$*-norm* approximation problems, like

$$f(x) = \sum_{i=1}^{m} \mid \langle a_i, x \rangle - b_i \mid^p,$$

are convex too.

4. Consider the function $f(x) = \ln \left( \sum\limits_{i=1}^{n} e^{x^{(i)}} \right)$, $x \in \mathbb{R}^n$. Define $\varkappa(x) = \sum\limits_{i=1}^{n} e^{x^{(i)}}$. For an arbitrary $h \in \mathbb{R}^n$, we have

$$\langle \nabla f(x), h \rangle = \tfrac{1}{\varkappa(x)} \sum_{i=1}^{n} e^{x^{(i)}} h^{(i)},$$

$$\begin{aligned} \langle \nabla^2 f(x) h, h \rangle &= \tfrac{1}{\varkappa(x)} \sum_{i=1}^{n} e^{x^{(i)}} \left(h^{(i)}\right)^2 - \tfrac{1}{\varkappa^2(x)} \left( \sum_{i=1}^{n} e^{x^{(i)}} h^{(i)} \right)^2 \\ &= \tfrac{1}{\varkappa(x)} \langle \left( D(x) - \tfrac{1}{\varkappa(x)} d(x) d^T(x) \right) h, h \rangle, \end{aligned}$$

where $D(x)$ is a diagonal matrix with diagonal entries $e^{x^{(i)}}$, $i = 1, \dots, n$, and the vector $d(x) \in \mathbb{R}^n$ has the same entries. Since $\varkappa(x) = \langle d(x), \bar{e}_n \rangle$, it is easy to see that $D(x) \succeq \frac{1}{\varkappa(x)} d(x) d^T(x)$. Thus, by Theorem 2.1.4 the function $f$ is convex on $\mathbb{R}^n$. □

Note that for general convex functions, differentiability itself cannot ensure any favorable growth properties. Therefore, we need to consider the problem classes with some bounds on the derivatives. The most important functions of that type are convex functions whose gradient is Lipschitz continuous in the standard Euclidean norm. However, for future use in this book, let us explicitly state the necessary and sufficient conditions for Lipschitz continuity of the gradient with respect to an *arbitrary norm* $\| \cdot \|$ in $\mathbb{R}^n$. In this case, the size of linear functions on $\mathbb{R}^n$ (e.g. the gradients) must be measured in the dual norm

$$\|g\|_* = \max_{x \in \mathbb{R}^n} \{ \langle g, x \rangle : \ \|x\| \le 1 \}.$$

This definition is necessary and sufficient for the justification of the *Cauchy-Schwarz inequality*:

$$\langle g, x\rangle \le \|g\|_* \cdot \|x\|, \quad x, g \in \mathbb{R}^n. \tag{2.1.6}$$

Thus, for functions with Lipschitz continuous gradient with respect to the norm $\|\cdot\|$ we introduce a new notation: $f \in \mathscr{F}_L^{1,1}(Q, \|\cdot\|)$ means that $Q \subseteq \mathrm{dom} f$ and

$$\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\|, \quad \forall x, y \in Q. \tag{2.1.7}$$

If in this notation the norm is missing, then we are working with the standard Euclidean norm (e.g. $\mathscr{F}_L^{1,1}(\mathbb{R}^n)$). Let us prove that this norm is self-dual.

**Lemma 2.1.3** *For any $x$ and $s$ in $\mathbb{R}^n$ we have*

$$\max_{x\in\mathbb{R}^N}\left\{\langle s, x\rangle : \sum_{i=1}^n (x^{(i)})^2 \le 1\right\} = \left[\sum_{i=1}^n (s^{(i)})^2\right]^{1/2}.$$

*Proof* Let $\|\cdot\|$ be the standard Euclidean norm. By simple coordinate maximization, it is easy to check that

$$\max_{x\in\mathbb{R}^n}\{2\langle s, x\rangle - \|x\|^2\} = \max_{x\in\mathbb{R}^n}\left\{\sum_{i=1}^n \left[2s^{(i)}x^{(i)} - (x^{(i)})^2\right]\right\} = \|s\|^2.$$

On the other hand,

$$\begin{aligned}\max_{x\in\mathbb{R}^n}\{2\langle s, x\rangle - \|x\|^2\} &= \max_{x\in\mathbb{R}^n, \tau\in\mathbb{R}}\{2\tau\langle s, x\rangle - \tau^2\|x\|^2\} = \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\langle s,x\rangle^2}{\|x\|^2}\\ &= \max_{\|x\|\le 1}\langle s, x\rangle^2.\end{aligned}$$

□

Thus, the standard Euclidean norm can be used both for measuring sizes of points and gradients. Before we proceed, let us prove a simple property of general norms.

**Lemma 2.1.4** *For all $x, y \in \mathbb{R}^n$ and $\alpha \in [0, 1]$ we have*

$$\alpha\|x\|^2 + (1-\alpha)\|y\|^2 \ge \alpha(1-\alpha)(\|x\| + \|y\|)^2 \ge \alpha(1-\alpha)\|x - y\|^2. \tag{2.1.8}$$

*Proof* Using the inequality $a^2 + b^2 \ge 2ab$ with $a = \alpha\|x\|$ and $b = (1-\alpha)\|y\|$, we get the first inequality. The second one follows from the triangle inequality for norms. □

**Theorem 2.1.5** *All conditions below, holding for all* $x,\ y \in \mathbb{R}^n$ *and* $\alpha$ *from* $[0, 1]$, *are equivalent to the inclusion* $f \in \mathscr{F}_L^{1,1}(\mathbb{R}^n, \|\cdot\|)$*:*

$$0 \;\le\; f(y) - f(x) - \langle \nabla f(x), y - x\rangle \;\le\; \tfrac{L}{2} \parallel x - y \parallel^2, \tag{2.1.9}$$

$$f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{1}{2L} \parallel \nabla f(x) - \nabla f(y) \parallel_*^2 \le\; f(y), \tag{2.1.10}$$

$$\tfrac{1}{L} \parallel \nabla f(x) - \nabla f(y) \parallel_*^2 \le\; \langle \nabla f(x) - \nabla f(y), x - y\rangle, \tag{2.1.11}$$

$$0 \;\le\; \langle \nabla f(x) - \nabla f(y), x - y\rangle \;\le\; L \parallel x - y \parallel^2, \tag{2.1.12}$$

$$\begin{aligned}\alpha f(x) + (1-\alpha) f(y) \ge\; & f(\alpha x + (1-\alpha)y)\\ & + \tfrac{\alpha(1-\alpha)}{2L} \parallel \nabla f(x) - \nabla f(y) \parallel_*^2,\end{aligned} \tag{2.1.13}$$

$$\begin{aligned}0 \;\le\; & \alpha f(x) + (1-\alpha) f(y) - f(\alpha x + (1-\alpha)y)\\ & \le \alpha(1-\alpha)\tfrac{L}{2} \parallel x - y \parallel^2 .\end{aligned} \tag{2.1.14}$$

*Moreover, if* $f \in \mathscr{F}_L^{1,1}(Q)$, *then inequalities (2.1.9), (2.1.12), and (2.1.14) are valid for all* $x, y \in Q$.

*Proof* Indeed, the first inequality in (2.1.9) follows from the definition of convex functions. To prove the second one, note that

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x\rangle \quad &= \quad \int_0^1 \langle \nabla f(x + \tau(y-x)) \\ &\qquad -\nabla f(x), y - x\rangle d\tau \\ &\overset{(2.1.6),\,(2.1.7)}{\le} \int_0^1 L\tau \|y - x\|^2 d\tau \;=\; \tfrac{L}{2}\|y - x\|^2. \end{aligned}$$

Further, let us fix $x_0 \in \mathbb{R}^n$. Consider the function $\phi(y) = f(y) - \langle \nabla f(x_0), y\rangle$. Note that $\phi \in \mathscr{F}_L^{1,1}(\mathbb{R}^n, \|\cdot\|)$ and its optimal point is $y^* = x_0$. Therefore, in view of (2.1.9), we have

$$\begin{aligned}\phi(y^*) \;&=\; \min_{x \in \mathbb{R}^n} \phi(x) \overset{(2.1.9)}{\le} \min_{x \in \mathbb{R}^n} \left\{ \phi(y) + \langle \nabla\phi(y), x - y\rangle + \tfrac{L}{2}\|x - y\|^2 \right\} \\ &\overset{(2.1.6)}{=} \min_{r \ge 0} \left\{ \phi(y) - r\|\nabla\phi(y)\|_* + \tfrac{L}{2} r^2 \right\} \;=\; \phi(y) - \tfrac{1}{2L} \parallel \nabla\phi(y) \parallel_*^2,\end{aligned}$$

and we get (2.1.10) since $\nabla\phi(y) = \nabla f(y) - \nabla f(x_0)$.

We obtain (2.1.11) from inequality (2.1.10) by adding two copies of it with $x$ and $y$ interchanged. Applying the Cauchy–Schwarz inequality to (2.1.11), we get $\| \nabla f(x) - \nabla f(y) \|_* \leq L \| x - y \|$.

In the same way, we can obtain (2.1.12) from (2.1.9). In order to get (2.1.9) from (2.1.12), we apply integration:

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \int\limits_0^1 \langle \nabla f(x + \tau(y-x)) - \nabla f(x), y - x \rangle d\tau$$

$$\leq \tfrac{1}{2} L \|y - x\|^2.$$

Let us now prove two last inequalities. Define $x_\alpha = \alpha x + (1-\alpha)y$. Then, using (2.1.10), we get

$$f(x) \geq f(x_\alpha) + \langle \nabla f(x_\alpha), (1-\alpha)(x-y) \rangle + \tfrac{1}{2L} \| \nabla f(x) - \nabla f(x_\alpha) \|_*^2,$$

$$f(y) \geq f(x_\alpha) + \langle \nabla f(x_\alpha), \alpha(y-x) \rangle + \tfrac{1}{2L} \| \nabla f(y) - \nabla f(x_\alpha) \|_*^2 .$$

Adding these inequalities multiplied by $\alpha$ and $(1-\alpha)$ respectively, and using inequality (2.1.8), we get (2.1.13). It is easy to check that we get (2.1.10) from (2.1.13) by letting $\alpha \to 1$.

Similarly, from (2.1.9) we get

$$f(x) \leq f(x_\alpha) + \langle \nabla f(x_\alpha), (1-\alpha)(x-y) \rangle + \tfrac{L}{2} \| (1-\alpha)(x-y) \|^2,$$

$$f(y) \leq f(x_\alpha) + \langle \nabla f(x_\alpha), \alpha(y-x) \rangle + \tfrac{L}{2} \| \alpha(y-x) \|^2 .$$

Adding these inequalities multiplied by $\alpha$ and $(1-\alpha)$ respectively, we obtain (2.1.14), and we get back to (2.1.9) as $\alpha \to 1$. □

Finally, let us characterize the class $\mathscr{F}_L^{2,1}(\mathbb{R}^n, \|\cdot\|)$.

**Theorem 2.1.6** *A twice continuously differentiable function $f$ belongs to the class $\mathscr{F}_L^{2,1}(\mathbb{R}^n, \|\cdot\|)$ if and only if for any $x, h \in \mathbb{R}^n$ we have*

$$0 \leq \langle \nabla^2 f(x) h, h \rangle \ \leq \ L \|h\|^2. \tag{2.1.15}$$

*Proof* The first condition characterizes the convexity of the function $f(\cdot)$ and it was proved in Theorem 2.1.4. The second inequality is a limiting case of (2.1.12). □

Note that for the class $\mathscr{F}_L^{2,1}(\mathbb{R}^n)$, condition (2.1.15) can be written in the form of a matrix inequality:

$$0 \preceq \nabla^2 f(x) \ \preceq \ L I_n, \quad x \in \mathbb{R}^n. \tag{2.1.16}$$

### 2.1.2 *Lower Complexity Bounds for* $\mathscr{F}_L^{\infty,1}(\mathbb{R}^n)$

Let us check our potential ability to minimize smooth convex functions. In this section, we obtain the lower complexity bounds for optimization problems with objective functions from $\mathscr{F}_L^{\infty,1}(\mathbb{R}^n)$ (and, consequently, $\mathscr{F}_L^{1,1}(\mathbb{R}^n)$).

Recall that our problem class is as follows.

| | |
|---|---|
| **Model**: | $\min\limits_{x\in\mathbb{R}^n} f(x), \quad f \in \mathscr{F}_L^{\infty,1}(\mathbb{R}^n).$ |
| **Oracle**: | First-order local Black Box. |
| **Approximate solution**: | $\bar{x} \in \mathbb{R}^n,\ f(\bar{x}) - f^* \leq \epsilon.$ |

In order to make our considerations simpler, let us introduce the following assumption on iterative processes.

**Assumption 2.1.4** *An iterative method* $\mathscr{M}$ *generates a sequence of test points* $\{x_k\}$ *such that*

$$x_k \in x_0 + \mathrm{Lin}\{\nabla f(x_0), \ldots, \nabla f(x_{k-1})\}, \quad k \geq 1.$$

This assumption is not absolutely necessary and it can be avoided using more sophisticated reasoning. However, it holds for the majority of practical methods.

We can prove the lower complexity bounds for our problem class without developing a resisting oracle. Instead, we just point out the "worst function in the world" belonging to the class $\mathscr{F}_L^{\infty,1}(\mathbb{R}^n)$. This function appears to be difficult for *all* iterative schemes satisfying Assumption 2.1.4.

Let us fix some constant $L > 0$. Consider the following family of quadratic functions

$$f_k(x) = \tfrac{L}{4}\left\{\tfrac{1}{2}\left[(x^{(1)})^2 + \sum_{i=1}^{k-1}(x^{(i)} - x^{(i+1)})^2 + (x^{(k)})^2\right] - x^{(1)}\right\}$$

for $k = 1 \ldots n$. Note that for all $h \in \mathbb{R}^n$, we have

$$\langle \nabla^2 f_k(x)h, h\rangle = \tfrac{L}{4}\left[(h^{(1)})^2 + \sum_{i=1}^{k-1}(h^{(i)} - h^{(i+1)})^2 + (h^{(k)})^2\right] \geq 0,$$

and

$$\begin{aligned} \langle \nabla^2 f_k(x)h, h\rangle &\le \tfrac{L}{4}\left[ (h^{(1)})^2 + \sum_{i=1}^{k-1} 2((h^{(i)})^2 + (h^{(i+1)})^2) + (h^{(k)})^2 \right] \\ &\le L \sum_{i=1}^{n} (h^{(i)})^2. \end{aligned}$$

Thus, $0 \preceq \nabla^2 f_k(x) \preceq L I_n$. Therefore, $f_k(\cdot) \in \mathscr{F}_L^{\infty,1}(\mathbb{R}^n)$, $1 \le k \le n$.

Let us compute the minimal value of the function $f_k$. Note that $\nabla^2 f_k(x) = \frac{L}{4} A_k$ with

$$A_k = \left( \begin{array}{c|c} \begin{array}{c} k \\ \text{lines} \end{array} \left\{ \begin{array}{cccccc} 2 & -1 & 0 & & & \\ -1 & 2 & -1 & & 0 & \\ 0 & -1 & 2 & & & \\ & \dots & & & \dots & \\ & & & -1 & 2 & -1 \\ & 0 & & 0 & -1 & 2 \end{array} \right. & 0_{k,n-k} \\ \hline 0_{n-k,k} & 0_{n-k,n-k} \end{array} \right),$$

where $0_{k,p}$ is a $(k \times p)$ zero matrix. Therefore, the equation

$$\nabla f_k(x) = A_k x - e_1 = 0$$

has the following unique solution:

$$\bar{x}_k^{(i)} = \begin{cases} 1 - \frac{i}{k+1}, \; i = 1 \dots k, \\ \quad 0, \qquad k+1 \le i \le n. \end{cases}$$

Hence, the optimal value of the function $f_k$ is

$$f_k^* = \tfrac{L}{4}\left[ \tfrac{1}{2}\langle A_k \bar{x}_k, \bar{x}_k\rangle - \langle e_1, \bar{x}_k\rangle \right] = -\tfrac{L}{8}\langle e_1, \bar{x}_k\rangle \; = \; \tfrac{L}{8}\left( -1 + \tfrac{1}{k+1} \right). \tag{2.1.17}$$

Note also that

$$\sum_{i=1}^{k} i^2 = \tfrac{k(k+1)(2k+1)}{6} \le \tfrac{(k+1)^3}{3}. \tag{2.1.18}$$

Therefore,

$$\begin{aligned}\| \bar{x}_k \|^2 &= \sum_{i=1}^{n} \left(\bar{x}_k^{(i)}\right)^2 = \sum_{i=1}^{k} \left(1 - \tfrac{i}{k+1}\right)^2 \\ &= k - \tfrac{2}{k+1} \sum_{i=1}^{k} i + \tfrac{1}{(k+1)^2} \sum_{i=1}^{k} i^2 \\ &\leq k - \tfrac{2}{k+1} \cdot \tfrac{k(k+1)}{2} + \tfrac{1}{(k+1)^2} \cdot \tfrac{(k+1)^3}{3} = \tfrac{1}{3}(k+1).\end{aligned} \tag{2.1.19}$$

Let $\mathbb{R}^{k,n} = \{x \in \mathbb{R}^n \mid x^{(i)} = 0,\ k+1 \leq i \leq n\}$. This is the subspace of $\mathbb{R}^n$ in which only the first $k$ components of the point can differ from zero. From the analytical form of the functions $\{f_k\}$, it is easy to see that for all $x \in \mathbb{R}^{k,n}$ we have

$$f_p(x) \equiv f_k(x), \quad p = k, \dots, n.$$

Let us fix some $p$, $1 \leq p \leq n$.

**Lemma 2.1.5** *Let $x_0 = 0$. Then for any sequence $\{x_k\}_{k=0}^{p}$ satisfying the condition*

$$x_k \in \mathscr{L}_k \overset{\text{def}}{=} \mathrm{Lin}\{\nabla f_p(x_0), \dots, \nabla f_p(x_{k-1})\},$$

*we have $\mathscr{L}_k \subseteq \mathbb{R}^{k,n}$.*

*Proof* Since $x_0 = 0$, we have $\nabla f_p(x_0) = -\frac{L}{4} e_1 \in \mathbb{R}^{1,n}$. Thus $\mathscr{L}_1 \equiv \mathbb{R}^{1,n}$.

Let $\mathscr{L}_k \subseteq \mathbb{R}^{k,n}$ for some $k < p$. Since the matrix $A_p$ is tri-diagonal, for any $x \in \mathbb{R}^{k,n}$ we have $\nabla f_p(x) \in \mathbb{R}^{k+1,n}$. Therefore $\mathscr{L}_{k+1} \subseteq \mathbb{R}^{k+1,n}$, and we can complete the proof by induction. □

**Corollary 2.1.1** *For any sequence $\{x_k\}_{k=0}^{p}$ with $x_0 = 0$ and $x_k \in \mathscr{L}_k$, we have*

$$f_p(x_k) \geq f_k^*.$$

*Proof* Indeed, $x_k \in \mathscr{L}_k \subseteq \mathbb{R}^{k,n}$ and therefore $f_p(x_k) = f_k(x_k) \geq f_k^*$. □

Now we are ready to prove the main result of this section.

**Theorem 2.1.7** *For any $k$, $1 \leq k \leq \frac{1}{2}(n-1)$, and any $x_0 \in \mathbb{R}^n$ there exists a function $f \in \mathscr{F}_L^{\infty,1}(\mathbb{R}^n)$ such that for any first-order method $\mathscr{M}$ satisfying Assumption 2.1.4 we have*

$$f(x_k) - f^* \geq \tfrac{3L\|x_0 - x^*\|^2}{32(k+1)^2},$$

$$\| x_k - x^* \|^2 \geq \tfrac{1}{8} \| x_0 - x^* \|^2,$$

*where $x^*$ is the minimum of the function $f$ and $f^* = f(x^*)$.*

*Proof* It is clear that the methods of this type are invariant with respect to a simultaneous shift of all objects in the space of variables. Thus, the sequence of iterates, which is generated by such a method for the function $f(\cdot)$ starting from $x_0$, is just a shift of the sequence generated for $\bar{f}(x) = f(x + x_0)$ starting from the origin. Therefore, we can assume that $x_0 = 0$.

Let us prove the first inequality. For that, let us fix $k$ and apply $\mathscr{M}$ to minimize $f(x) = f_{2k+1}(x)$. Then $x^* = \bar{x}_{2k+1}$ and $f^* = f^*_{2k+1}$. Using Corollary 2.1.1, we conclude that

$$f(x_k) \equiv f_{2k+1}(x_k) = f_k(x_k) \geq f_k^*.$$

Hence, since $x_0 = 0$, in view of (2.1.17) and (2.1.19) we get the following estimate:

$$\frac{f(x_k)-f^*}{\|x_0-x^*\|^2} \geq \frac{\frac{L}{8}\left(-1+\frac{1}{k+1}+1-\frac{1}{2k+2}\right)}{\frac{1}{3}(2k+2)} = \frac{3}{8}L \cdot \frac{1}{4(k+1)^2}.$$

Let us prove the second inequality. Since $x_k \in \mathbb{R}^{k,n}$ and $x_0 = 0$, we have

$$\| x_k - x^* \|^2 \geq \sum_{i=k+1}^{2k+1} \left(\bar{x}^{(i)}_{2k+1}\right)^2 = \sum_{i=k+1}^{2k+1} \left(1 - \frac{i}{2k+2}\right)^2$$

$$= k + 1 - \frac{1}{k+1} \sum_{i=k+1}^{2k+1} i + \frac{1}{4(k+1)^2} \sum_{i=k+1}^{2k+1} i^2.$$

In view of (2.1.18), we have

$$\sum_{i=k+1}^{2k+1} i^2 = \frac{1}{6}\left[(2k+1)(2k+2)(4k+3) - k(k+1)(2k+1)\right]$$

$$= \frac{1}{6}(k+1)(2k+1)(7k+6).$$

Therefore, using (2.1.19) we finally obtain

$$\| x_k - x^* \|^2 \geq k + 1 - \frac{1}{k+1} \cdot \frac{(3k+2)(k+1)}{2} + \frac{(2k+1)(7k+6)}{24(k+1)}$$

$$= \frac{(2k+1)(7k+6)}{24(k+1)} - \frac{k}{2} = \frac{2k^2+7k+6}{24(k+1)}$$

$$\geq \frac{2k^2+7k+6}{16(k+1)^2} \| x_0 - \bar{x}_{2k+1} \|^2 \geq \frac{1}{8} \| x_0 - x^* \|^2 .$$

□

The above theorem is valid only under the assumption that the number of steps of the iterative scheme is not too large as compared with the dimension of the space of variables ($k \le \frac{1}{2}(n-1)$). Complexity bounds of this type are called *uniform* in the dimension. Clearly, they are valid for very large problems, in which we cannot even wait for $n$ iterates of the method. However, even for problems with a moderate dimension, these bounds also provide us with some information. Firstly, they describe the potential performance of numerical methods at the initial stage of the minimization process. Secondly, they warn us that without a direct use of finite-dimensional arguments we cannot justify a better complexity of the corresponding numerical scheme.

To conclude this section, let us note that the obtained lower bound for the value of the objective function is rather optimistic. Indeed, after one hundred iterations we could decrease the initial residual by $10^4$ times. However, the result on the behavior of the minimizing sequence is quite disappointing. The convergence to the optimal point can be *arbitrarily* slow. Since this is a lower bound, this conclusion is inevitable for our problem class. The only thing we can do is to try to find problem classes in which the situation could be better. This is the goal of the next section.

### 2.1.3 Strongly Convex Functions

Let us look at a possible restriction of the functional class $\mathscr{F}_L^{1,1}(\mathbb{R}^n, \| \cdot \|)$, for which we can guarantee a reasonable rate of convergence to a unique solution of the minimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \qquad f \in \mathscr{F}^1(\mathbb{R}^n, \| \cdot \|).$$

Recall that in Sect. 1.2.3 we have proved that in a small neighborhood of a nondegenerate local minimum the Gradient Method (1.2.15) converges linearly. Let us try to globalize this non-degeneracy assumption. Namely, let us assume that there exists some constant $\mu > 0$ such that for any $\bar{x}$ with $\nabla f(\bar{x}) = 0$ and any $x \in \mathbb{R}^n$ we have

$$f(x) \ge f(\bar{x}) + \frac{1}{2}\mu \parallel x - \bar{x} \parallel^2 .$$

Recall that the norm in this definition can be general.

Using the same reasoning as in the beginning of Sect. 2.1.1, we obtain the class of *strongly convex functions*.

**Definition 2.1.3** A continuously differentiable function $f(\cdot)$ is called *strongly convex* on $\mathbb{R}^n$ (notation $f \in \mathscr{S}^1_\mu(Q, \|\cdot\|)$) if there exists a constant $\mu > 0$ such that for any $x, y \in Q$ we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\mu \parallel y - x \parallel^2 . \qquad (2.1.20)$$

The constant $\mu$ is called the *convexity parameter* of function $f$.

We will also consider the classes $\mathscr{S}^{k,l}_{\mu,L}(Q, \|\cdot\|)$ where the indices $k$, $l$ and $L$ have the same meaning as for the class $C^{k,l}_L(Q)$.

Let us mention the most important properties of strongly convex functions.

**Theorem 2.1.8** *If $f \in \mathscr{S}^1_\mu(\mathbb{R}^n)$ and $\nabla f(x^*) = 0$, then*

$$f(x) \geq f(x^*) + \tfrac{1}{2}\mu \parallel x - x^* \parallel^2 \qquad (2.1.21)$$

*for all $x \in \mathbb{R}^n$.*

*Proof* Since $\nabla f(x^*) = 0$, for any $x \in \mathbb{R}^n$, we have

$$\begin{aligned} f(x) &\overset{(2.1.20)}{\geq} f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \tfrac{1}{2}\mu \parallel x - x^* \parallel^2 \\ &= f(x^*) + \tfrac{1}{2}\mu \parallel x - x^* \parallel^2 . \end{aligned}$$

□

Let us describe the result of addition of two strongly convex functions.

**Lemma 2.1.6** *If $f_1 \in \mathscr{S}^1_{\mu_1}(Q_1, \|\cdot\|)$, $f_2 \in \mathscr{S}^1_{\mu_2}(Q_2, \|\cdot\|)$ and $\alpha, \beta \geq 0$, then*

$$f = \alpha f_1 + \beta f_2 \in \mathscr{S}^1_{\alpha\mu_1 + \beta\mu_2}(Q_1 \bigcap Q_2, \|\cdot\|).$$

*Proof* For any $x, y \in Q_1 \bigcap Q_2$, we have

$$f_1(y) \geq f_1(x) + \langle \nabla f_1(x), y - x \rangle + \tfrac{1}{2}\mu_1 \parallel y - x \parallel^2,$$

$$f_2(y) \geq f_2(x) + \langle \nabla f_2(x), y - x \rangle + \tfrac{1}{2}\mu_2 \parallel y - x \parallel^2 .$$

It remains to add these equations multiplied by $\alpha$ and $\beta$ respectively. □

Note that the class $\mathscr{S}^1_0(Q, \|\cdot\|)$ coincides with $\mathscr{F}^1(Q, \|\cdot\|)$. Therefore, addition of a convex function and a strongly convex function gives a strongly convex function with the same value of convexity parameter.

Let us give several equivalent definitions of strongly convex functions.

**Theorem 2.1.9** *Let $f$ be continuously differentiable. Both conditions below, holding for all $x, y \in Q$ and $\alpha \in [0,1]$, are equivalent to inclusion $f \in \mathscr{S}^1_\mu(Q, \| \cdot \|)$:*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \parallel x - y \parallel^2, \tag{2.1.22}$$

$$\begin{aligned} \alpha f(x) + (1-\alpha) f(y) &\geq f(\alpha x + (1-\alpha) y) \\ &\quad + \alpha(1-\alpha)\tfrac{\mu}{2} \parallel x - y \parallel^2 . \end{aligned} \tag{2.1.23}$$

The proof of this theorem is very similar to the proof of Theorem 2.1.5 and we leave it as an exercise for the reader.

The next statement is sometimes useful.

**Theorem 2.1.10** *If $f \in \mathscr{S}^1_\mu(\mathbb{R}^n, \| \cdot \|)$, then for any $x$ and $y$ from $\mathbb{R}^n$ we have*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2\mu} \parallel \nabla f(x) - \nabla f(y) \parallel_*^2, \tag{2.1.24}$$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \tfrac{1}{\mu} \parallel \nabla f(x) - \nabla f(y) \parallel_*^2, \tag{2.1.25}$$

$$\mu \|x - y\| \leq \|\nabla f(x) - \nabla f(y)\|_*. \tag{2.1.26}$$

*Proof* Let us fix some $x \in \mathbb{R}^n$. Consider the function

$$\phi(y) = f(y) - \langle \nabla f(x), y \rangle \in \mathscr{S}^1_\mu(\mathbb{R}^n, \| \cdot \|).$$

Since $\nabla \phi(x) = 0$, for any $y \in \mathbb{R}^n$, we have

$$\begin{aligned} \phi(x) = \min_{v \in \mathbb{R}^n} \phi(v) &\overset{(2.1.20)}{\geq} \min_{v \in \mathbb{R}^n} [\phi(y) + \langle \nabla \phi(y), v - y \rangle + \tfrac{1}{2}\mu \|v - y\|^2] \\ &= \phi(y) - \tfrac{1}{2\mu} \|\nabla \phi(y)\|_*^2, \end{aligned}$$

and this is exactly (2.1.24). Adding two copies of (2.1.24) with $x$ and $y$ interchanged, we get (2.1.25). Finally, (2.1.26) follows from (2.1.25) and (2.1.22). □

Let us present a second-order characterization of the class $\mathscr{S}^1_\mu(Q, \| \cdot \|)$.

**Theorem 2.1.11** *Let a continuous function $f$ be twice continuously differentiable in* int$Q$. *It belongs to the class $\mathscr{S}^2_\mu(Q, \| \cdot \|)$ if and only if for all $x \in$* int$Q$ *and $h \in \mathbb{R}^n$ we have*

$$\langle \nabla^2 f(x) h, h \rangle \succeq \mu \|h\|^2. \tag{2.1.27}$$

*Proof* We get (2.1.27) from (2.1.22) by setting $y = x + \alpha h \in Q$ with $\alpha$ small enough and letting $\alpha \to 0$. □

In the case of the standard Euclidean norm, condition (2.1.27) can be written in the form of a matrix inequality:

$$\nabla^2 f(x) \succeq \mu I_n, \quad x \in \mathrm{int} Q. \tag{2.1.28}$$

Now we can look at some examples of strongly convex functions.

*Example 2.1.2*

1. Let a symmetric matrix $A$ satisfy the conditions $\mu I_n \preceq A \preceq L I_n$. Then, since $\nabla^2 f(x) = A$, we have

$$f(x) = \alpha + \langle a, x\rangle + \frac{1}{2}\langle Ax, x\rangle \in \mathscr{S}^{\infty,1}_{\mu,L}(\mathbb{R}^n) \subset \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n).$$

   Adding this function to a convex function, we get other examples of strongly convex functions.
2. Let $Q = \Delta_n^+ \overset{\mathrm{def}}{=} \{x \in \mathbb{R}^n_+ : \ \langle \bar{e}_n, x\rangle \leq 1\}$, where $\bar{e}_n \in \mathbb{R}^n$ is a vector of all ones. Consider the *entropy function*:

$$\eta(x) = \sum_{i=1}^n x^{(i)} \ln x^{(i)}, \quad x \in \Delta_n^+. \tag{2.1.29}$$

   For direction $h \in \mathbb{R}^n$, we have $\langle \nabla^2 \eta(x) h, h\rangle = \sum\limits_{i=1}^n \frac{(h^{(i)})^2}{x^{(i)}}$. We need to find the minimum of this expression in $x \in \mathrm{int}\Delta_n^+$. Since it is decreasing in $x$, we conclude that the inequality constraint is active and we need to compute $\min\limits_{\langle e_n x\rangle = 1} \sum\limits_{i=1}^n \frac{(h^{(i)})^2}{x^{(i)}}$. In view of Corollary 1.2.1, this minimum $x_*$ can be found from the system of equations

$$\frac{(h^{(i)})^2}{(x_*^{(i)})^2} = \lambda_*,$$

   where $\lambda^*$ is the optimal dual multiplier. It can be found from the equation

$$1 = \sum_{i=1}^n x_*^{(i)} = \frac{1}{\lambda_*^{1/2}} \sum_{i=1}^n |h^{(i)}|.$$

   Thus, $\langle \nabla^2 \eta(x) h, h\rangle \geq \sum\limits_{i=1}^n \frac{(h^{(i)})^2}{x_*^{(i)}} = \left(\sum\limits_{i=1}^n |h^{(i)}|\right)^2$, and by Theorem 2.1.11 we conclude that the entropy function is strongly convex on $\Delta_N^+$ in the $\ell_1$-norm with convexity parameter one. □

One of the most important functional classes is $\mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$ (recall that the corresponding norm is standard Euclidean). This class is described by the following inequalities:

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \parallel x - y \parallel^2, \tag{2.1.30}$$

$$\parallel \nabla f(x) - \nabla f(y) \parallel \leq L \parallel x - y \parallel . \tag{2.1.31}$$

The value $Q_f = L/\mu \geq 1$ is called the *condition number* of the function $f$.

It is important that inequality (2.1.30) can be strengthened by the additional information obtained from (2.1.31).

**Theorem 2.1.12** *If $f \in \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$, then for any $x, y \in \mathbb{R}^n$ we have*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \tfrac{\mu L}{\mu+L} \parallel x - y \parallel^2 + \tfrac{1}{\mu+L} \parallel \nabla f(x) - \nabla f(y) \parallel^2 . \tag{2.1.32}$$

*Proof* Define $\phi(x) = f(x) - \frac{1}{2}\mu\|x\|^2$. Then $\nabla\phi(x) = \nabla f(x) - \mu x$. Hence, by inequalities (2.1.30) and (2.1.12), $\phi \in \mathscr{F}^{1,1}_{L-\mu}(\mathbb{R}^n)$. If $\mu = L$, then (2.1.32) is proved. If $\mu < L$, then by (2.1.11) we have

$$\langle \nabla\phi(x) - \nabla\phi(y), y - x \rangle \geq \tfrac{1}{L-\mu}\|\nabla\phi(x) - \nabla\phi(y)\|^2,$$

and this is exactly (2.1.32). □

### 2.1.4 Lower Complexity Bounds for $\mathscr{S}^{\infty,1}_{\mu,L}(\mathbb{R}^n)$

Let us obtain the lower complexity bounds for unconstrained minimization of functions from the class $\mathscr{S}^{\infty,1}_{\mu,L}(\mathbb{R}^n) \subset \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$. Consider the following problem class.

| | |
|---|---|
| **Model**: | $\min\limits_{x\in\mathbb{R}^n} f(x), \quad f \in \mathscr{S}^{\infty,1}_{\mu,L}(\mathbb{R}^n),\ \mu > 0,\ n \geq 1.$ |
| **Oracle**: | First-order local Black Box. |
| **Approximate solution**: | $\bar{x}:\ f(\bar{x}) - f^* \leq \epsilon,\ \parallel \bar{x} - x^* \parallel^2 \leq \epsilon.$ |

As in the previous section, we consider methods satisfying Assumption 2.1.4. We are going to find the lower complexity bounds for our problems in terms of the *condition number* $Q_f = \frac{L}{\mu}$. Note that in the description of our problem class, we do not fix the dimension of the space of variables. Therefore, formally this class also includes an infinite-dimensional problem.

We are going to give an example of a bad function defined in an infinite-dimensional space. It is also possible to do this in finite dimensions, but the corresponding reasoning is more complicated.

Consider $\mathbb{R}^\infty \equiv \ell_2$, the space of all sequences $x = \{x^{(i)}\}_{i=1}^\infty$ with finite standard Euclidean norm

$$\| x \|^2 = \sum_{i=1}^{\infty} \left(x^{(i)}\right)^2 < \infty.$$

Let us choose two parameters, $\mu > 0$ and $Q_f > 1$, which define the following function

$$f_{\mu,Q_f}(x) = \frac{\mu(Q_f-1)}{8}\left\{(x^{(1)})^2 + \sum_{i=1}^{\infty}(x^{(i)} - x^{(i+1)})^2 - 2x^{(1)}\right\} + \frac{\mu}{2} \| x \|^2 .$$

Let $L = \mu Q_f$ and

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & \ddots \\ 0 & 0 & \ddots & \ddots \end{pmatrix}.$$

Then $\nabla^2 f_{\mu,Q_f}(x) = \frac{\mu(Q_f-1)}{4}A + \mu I$, where $I$ is the unit operator in $\mathbb{R}^\infty$. As in Sect. 2.1.2, we can see that $0 \preceq A \preceq 4I$. Therefore,

$$\mu I \ \preceq \ \nabla^2 f_{\mu,Q_f}(x) \ \preceq \ (\mu(Q_f-1)+\mu)I \ = \ \mu Q_f I \ = \ L\,I.$$

This means that $f_{\mu,Q_f} \in \mathscr{S}^{\infty,1}_{\mu,L}(\mathbb{R}^\infty)$. Note that the condition number of the function $f_{\mu,Q_f}$ is $Q_f$.

Let us find the minimum of the function $f_{\mu,Q_f}$. The first-order optimality condition

$$\nabla f_{\mu,Q_f}(x) \equiv \left(\frac{\mu(Q_f-1)}{4}A + \mu I\right)x - \frac{\mu(Q_f-1)}{4}e_1 = 0$$

can be written as

$$\left(A + \frac{4}{Q_f-1}I\right)x = e_1.$$

The coordinate form of this equation is as follows:

$$\begin{aligned} 2\tfrac{Q_f+1}{Q_f-1}x^{(1)} - x^{(2)} &= 1, \\ x^{(k+1)} - 2\tfrac{Q_f+1}{Q_f-1}x^{(k)} + x^{(k-1)} &= 0,\ k = 2, \dots. \end{aligned} \tag{2.1.33}$$

Let $q$ be the smallest root of the equation

$$q^2 - 2\tfrac{Q_f+1}{Q_f-1}q + 1 = 0,$$

that is $q = \frac{\sqrt{Q_f}-1}{\sqrt{Q_f}+1}$. Then the sequence $(x^*)^{(k)} = q^k$, $k = 1, 2, \dots$, satisfies the system (2.1.33). Thus, we come to the following result.

**Theorem 2.1.13** *For any $x_0 \in \mathbb{R}^\infty$ and any constants $\mu > 0$, $Q_f > 1$, there exists a function $f \in \mathscr{S}^{\infty,1}_{\mu,L}(\mathbb{R}^\infty)$ such that for any first-order method $\mathscr{M}$ satisfying Assumption 2.1.4, we have*

$$\| x_k - x^* \|^2 \ge \left(\frac{\sqrt{Q_f}-1}{\sqrt{Q_f}+1}\right)^{2k} \| x_0 - x^* \|^2, \tag{2.1.34}$$

$$f(x_k) - f(x^*) \ge \tfrac{\mu}{2}\left(\frac{\sqrt{Q_f}-1}{\sqrt{Q_f}+1}\right)^{2k} \| x_0 - x^* \|^2, \tag{2.1.35}$$

*where $x^*$ is the unique unconstrained minimum of function $f$.*

*Proof* Indeed, we can assume that $x_0 = 0$. Let us choose $f(x) = f_{\mu,Q_f}(x)$. Then

$$\| x_0 - x^* \|^2 = \sum_{i=1}^{\infty}[(x^*)^{(i)}]^2 = \sum_{i=1}^{\infty} q^{2i} = \tfrac{q^2}{1-q^2}.$$

Since $\nabla^2 f_{\mu,Q_f}(x)$ is a tri-diagonal operator and $\nabla f_{\mu,Q_f}(0) = -\frac{L-\mu}{4}e_1$, we conclude that $x_k \in \mathbb{R}^{k,\infty}$. Therefore

$$\| x_k - x^* \|^2 \ge \sum_{i=k+1}^{\infty}[(x^*)^{(i)}]^2 = \sum_{i=k+1}^{\infty} q^{2i} = \tfrac{q^{2(k+1)}}{1-q^2} = q^{2k} \| x_0 - x^* \|^2 .$$

The second bound of this theorem follows from (2.1.34) and Theorem 2.1.8. □

### *2.1.5 The Gradient Method*

Let us describe the performance of the Gradient Method as applied to the problem

$$\min_{x\in\mathbb{R}^n} f(x) \tag{2.1.36}$$

with $f \in \mathscr{F}_L^{1,1}(\mathbb{R}^n)$. Recall that the scheme of the Gradient Method is as follows.

```
Gradient Method

0. Choose $x_0 \in \mathbb{R}^n$.
1. $k$th iteration ($k \geq 0$).                                   (2.1.37)
    (a) Compute $f(x_k)$ and $\nabla f(x_k)$.
    (b) Find $x_{k+1} = x_k - h_k \nabla f(x_k)$ (see Sect. 1.2.3 for step-
        size rules).
```

In this section, we analyze the simplest variant of the gradient scheme with $h_k = h > 0$. It is possible to show that for all other reasonable step-size rules the rate of convergence of this method is similar. Denote by $x^*$ an arbitrary optimal point of our problem, and let $f^* = f(x^*)$.

**Theorem 2.1.14** *Let $f \in \mathscr{F}_L^{1,1}(\mathbb{R}^n)$ and $0 < h < \frac{2}{L}$. Then the Gradient Method generates a sequence of points $\{x_k\}$, with function values satisfying the inequality*

$$f(x_k) - f^* \leq \frac{2(f(x_0)-f^*)\,\|x_0-x^*\|^2}{2\|x_0-x^*\|^2+k\cdot h(2-Lh)\cdot(f(x_0)-f^*)}, \quad k \geq 0.$$

*Proof* Let $r_k = \| x_k - x^* \|$. Then

$$\begin{aligned} r_{k+1}^2 &= \| x_k - x^* - h\nabla f(x_k) \|^2 \\ &= r_k^2 - 2h\langle \nabla f(x_k), x_k - x^* \rangle + h^2 \| \nabla f(x_k) \|^2 \\ &\leq r_k^2 - h(\tfrac{2}{L} - h) \| \nabla f(x_k) \|^2 \end{aligned}$$

(we use (2.1.11) and $\nabla f(x^*) = 0$). Therefore, $r_k \leq r_0$. In view of (2.1.9), we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \tfrac{L}{2} \| x_{k+1} - x_k \|^2 \\ &= f(x_k) - \omega \| \nabla f(x_k) \|^2, \end{aligned}$$

where $\omega = h(1 - \frac{L}{2}h)$. Define $\Delta_k = f(x_k) - f^*$. Then

$$\Delta_k \overset{(2.1.2)}{\leq} \langle \nabla f(x_k), x_k - x^* \rangle \leq r_0 \parallel \nabla f(x_k) \parallel .$$

Therefore, $\Delta_{k+1} \leq \Delta_k - \frac{\omega}{r_0^2}\Delta_k^2$. Thus,

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{r_0^2} \cdot \frac{\Delta_k}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{r_0^2}.$$

Summing up these inequalities, we get

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_0} + \frac{\omega}{r_0^2}(k+1).$$

□

In order to choose the optimal step size, we need to maximize the function $\phi(h) = h(2 - Lh)$ with respect to $h$. The first-order optimality condition $\phi'(h) = 2 - 2Lh = 0$ provides us with the value $h^* = \frac{1}{L}$. In this case, we get the following rate of convergence for the Gradient Method:

$$f(x_k) - f^* \leq \frac{2L(f(x_0)-f^*)\|x_0-x^*\|^2}{2L\|x_0-x^*\|^2+k\cdot(f(x_0)-f^*)}. \tag{2.1.38}$$

Further, in view of (2.1.9) we have

$$f(x_0) \leq f^* + \langle \nabla f(x^*), x_0 - x^* \rangle + \frac{L}{2} \parallel x_0 - x^* \parallel^2 = f^* + \frac{L}{2} \parallel x_0 - x^* \parallel^2 .$$

Since the right-hand side of inequality (2.1.38) is increasing in $f(x_0) - f^*$, we obtain the following result.

**Corollary 2.1.2** *If $h = \frac{1}{L}$ and $f \in \mathscr{F}_L^{1,1}(\mathbb{R}^n)$, then*

$$f(x_k) - f^* \leq \frac{2L\|x_0-x^*\|^2}{k+4}. \tag{2.1.39}$$

Let us estimate the performance of the Gradient Method on the class of strongly convex functions.

**Theorem 2.1.15** *If $f \in \mathscr{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$ and $0 < h \leq \frac{2}{\mu+L}$, then the Gradient Method generates a sequence $\{x_k\}$ such that*

$$\parallel x_k - x^* \parallel^2 \leq \left(1 - \frac{2h\mu L}{\mu+L}\right)^k \parallel x_0 - x^* \parallel^2 .$$

*If* $h = \frac{2}{\mu+L}$, *then*

$$\| x_k - x^* \| \leq \left(\frac{Q_f-1}{Q_f+1}\right)^k \| x_0 - x^* \|,$$

$$f(x_k) - f^* \leq \frac{L}{2}\left(\frac{Q_f-1}{Q_f+1}\right)^{2k} \| x_0 - x^* \|^2,$$

*where* $Q_f = L/\mu$.

*Proof* Let $r_k = \| x_k - x^* \|$. Then

$$r_{k+1}^2 = \| x_k - x^* - h\nabla f(x_k) \|^2 = r_k^2 - 2h\langle \nabla f(x_k), x_k - x^* \rangle + h^2 \| \nabla f(x_k) \|^2$$

$$\leq \left(1 - \frac{2h\mu L}{\mu+L}\right) r_k^2 + h\left(h - \frac{2}{\mu+L}\right) \| \nabla f(x_k) \|^2$$

(we use (2.1.32) and $\nabla f(x^*) = 0$). The last inequality of the theorem follows from the previous one and (2.1.9). □

Note that the highest rate of convergence is achieved for $h = \frac{2}{\mu+L}$. In this case,

$$\|x_k - x^*\|^2 \leq \left(\frac{L-\mu}{L+\mu}\right)^{2k} \|x_0 - x^*\|^2. \tag{2.1.40}$$

We have already seen the step-size rule $h = \frac{2}{\mu+L}$ and the linear rate of convergence of the Gradient Method in Sect. 1.2.3, Theorem 1.2.4. However, this was only a local result.

Comparing the rate of convergence of the Gradient Method with the lower complexity bounds (Theorems 2.1.7 and 2.1.13), we can see that it is far from being optimal for the classes $\mathscr{F}_L^{1,1}(\mathbb{R}^n)$ and $\mathscr{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$. We should also note that on these problem classes the standard unconstrained minimization methods (Conjugate Gradients, Variable Metric) are not better. The optimal methods for minimizing smooth convex and strongly convex functions need the accumulation of some global information on the objective function. We will describe such schemes in the next section.

## 2.2 Optimal Methods

(Estimating sequences and Fast Gradient Methods; Decreasing the norm of the gradient; Convex sets; Constrained minimization problems; The gradient mapping; Minimization methods over simple sets.)

### 2.2.1 Estimating Sequences

Let us consider the following unconstrained minimization problem:

$$\min_{x \in \mathbb{R}^n} f(x), \tag{2.2.1}$$

where $f$ is strongly convex: $f \in \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$, $\mu \geq 0$. Since $\mathscr{S}^{1,1}_{0,L}(\mathbb{R}^n) \equiv \mathscr{F}^{1,1}_L(\mathbb{R}^n)$, this family of classes also contains the class of convex functions with Lipschitz continuous gradient. We assume that there exists a solution $x^*$ of problem (2.2.1) and define $f^* = f(x^*)$.

In Sect. 2.1, we proved the following convergence rates for the Gradient Method:

$$\begin{array}{ll} \mathscr{F}^{1,1}_L(\mathbb{R}^n): & f(x_k) - f^* \leq \frac{2L\|x_0-x^*\|^2}{k+4}, \\ \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n): & f(x_k) - f^* \leq \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k} \parallel x_0 - x^* \parallel^2 . \end{array}$$

These estimates differ from our lower complexity bounds (Theorem 2.1.7 and Theorem 2.1.13) by an order of magnitude. Of course, generally speaking, this does not mean that the Gradient Method is not optimal (it may be that the lower bounds are too optimistic). However, we will see that in our case the lower bounds are exact up to a constant factor. We prove this by constructing a method with rate of convergence proportional to these bounds.

Recall that the Gradient Method forms a relaxation sequence:

$$f(x_{k+1}) \leq f(x_k).$$

This fact is crucial for the justification of its convergence rate (Theorem 2.1.14). However, in Convex Optimization relaxation is not so important. Firstly, for some problem classes, this property is quite expensive. Secondly, the schemes and efficiency estimates of optimal methods are derived from some *global* topological properties of convex functions (see Theorem 2.1.5). From this point of view, the relaxation property is too microscopic to be useful.

The schemes and efficiency bounds of optimal methods are based on the notion of *estimating sequences*.

**Definition 2.2.1** A pair of sequences $\{\phi_k(x)\}_{k=0}^{\infty}$ and $\{\lambda_k\}_{k=0}^{\infty}$, $\lambda_k \geq 0$, are called the *estimating sequences* of the function $f(\cdot)$ if

$$\lambda_k \to 0,$$

and for any $x \in \mathbb{R}^n$ and all $k \geq 0$ we have

$$\phi_k(x) \leq (1 - \lambda_k) f(x) + \lambda_k \phi_0(x). \tag{2.2.2}$$

The next statement explains why these objects are useful.

**Lemma 2.2.1** *If for some sequence of points* $\{x_k\}$ *we have*

$$f(x_k) \leq \phi_k^* \overset{\text{def}}{=} \min_{x \in \mathbb{R}^n} \phi_k(x), \tag{2.2.3}$$

*then* $f(x_k) - f^* \leq \lambda_k[\phi_0(x^*) - f^*] \to 0$.

*Proof* Indeed,

$$\begin{aligned} f(x_k) \leq \phi_k^* &= \min_{x \in \mathbb{R}^n} \phi_k(x) \overset{(2.2.2)}{=} \min_{x \in \mathbb{R}^n} [(1 - \lambda_k) f(x) + \lambda_k \phi_0(x)] \\ &\leq (1 - \lambda_k) f(x^*) + \lambda_k \phi_0(x^*). \end{aligned}$$

□

Thus, for any sequence $\{x_k\}$, satisfying (2.2.3), we can derive its rate of convergence *directly* from the convergence rate of the sequence $\{\lambda_k\}$. However, at this moment we have two serious questions. Firstly, we do not know how to form the estimating sequences. Secondly, we do not know how to satisfy inequalities (2.2.3). The first question is simpler.

**Lemma 2.2.2** *Assume that:*

1. *a function* $f(\cdot)$ *belongs to the class* $\mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$,
2. $\phi_0(\cdot)$ *is an arbitrary convex function on* $\mathbb{R}^n$,
3. $\{y_k\}_{k=0}^{\infty}$ *is an arbitrary sequence of points in* $\mathbb{R}^n$,
4. *the coefficients* $\{\alpha_k\}_{k=0}^{\infty}$ *satisfy conditions* $\alpha_k \in (0, 1)$ *and* $\sum\limits_{k=0}^{\infty} \alpha_k = \infty$,
5. *we choose* $\lambda_0 = 1$.

*Then the pair of sequences* $\{\phi_k(\cdot)\}_{k=0}^{\infty}$ *and* $\{\lambda_k\}_{k=0}^{\infty}$, *defined recursively by the relations*

$$\begin{aligned} \lambda_{k+1} &= (1 - \alpha_k)\lambda_k, \\ \phi_{k+1}(x) &= (1 - \alpha_k)\phi_k(x) + \alpha_k \left[ f(y_k) + \langle \nabla f(y_k), x - y_k \rangle + \tfrac{\mu}{2} \parallel x - y_k \parallel^2 \right], \end{aligned} \tag{2.2.4}$$

*are estimating sequences.*

*Proof* Indeed, $\phi_0(x) \le (1-\lambda_0)f(x) + \lambda_0\phi_0(x) \equiv \phi_0(x)$. Further, let (2.2.2) hold for some $k \ge 0$. Then

$$\begin{aligned} \phi_{k+1}(x) &\overset{(2.1.20),(2.2.4)}{\le} (1-\alpha_k)\phi_k(x) + \alpha_k f(x) \\ &= (1-(1-\alpha_k)\lambda_k)f(x) + (1-\alpha_k)(\phi_k(x) - (1-\lambda_k)f(x)) \\ &\le (1-(1-\alpha_k)\lambda_k)f(x) + (1-\alpha_k)\lambda_k\phi_0(x) \\ &\overset{(2.2.4)}{\le} (1-\lambda_{k+1})f(x) + \lambda_{k+1}\phi_0(x). \end{aligned}$$

It remains to note that condition 4) ensures $\lambda_k \to 0$. □

Thus, the above statement provides us with some rules for updating the estimating sequences. Now we have two control sequences which can help us to maintain recursively the relation (2.2.3). At this moment, we are also free in our choice of initial function $\phi_0(x)$. Let us choose it as a simple quadratic function. Then, we can obtain a closed form recurrence for values $\phi_k^*$.

**Lemma 2.2.3** *Let $\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2} \parallel x - v_0 \parallel^2$. Then the process (2.2.4) preserves the canonical form of functions $\{\phi_k(x)\}$:*

$$\phi_k(x) \equiv \phi_k^* + \tfrac{\gamma_k}{2} \parallel x - v_k \parallel^2, \tag{2.2.5}$$

*where the sequences $\{\gamma_k\}$, $\{v_k\}$ and $\{\phi_k^*\}$ are defined as follows:*

$$\begin{aligned} \gamma_{k+1} &= (1-\alpha_k)\gamma_k + \alpha_k\mu, \\ v_{k+1} &= \tfrac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k v_k + \alpha_k\mu y_k - \alpha_k\nabla f(y_k)], \\ \phi_{k+1}^* &= (1-\alpha_k)\phi_k^* + \alpha_k f(y_k) - \tfrac{\alpha_k^2}{2\gamma_{k+1}} \parallel \nabla f(y_k) \parallel^2 \\ &\quad + \tfrac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}} \left( \tfrac{\mu}{2} \parallel y_k - v_k \parallel^2 + \langle \nabla f(y_k), v_k - y_k\rangle \right). \end{aligned}$$

*Proof* Note that $\nabla^2\phi_0(x) = \gamma_0 I_n$. Let us show that $\nabla^2\phi_k(x) = \gamma_k I_n$ for all $k \ge 0$. Indeed, if it is true for some $k$, then

$$\nabla^2\phi_{k+1}(x) = (1-\alpha_k)\nabla^2\phi_k(x) + \alpha_k\mu I_n = ((1-\alpha_k)\gamma_k + \alpha_k\mu)I_n \equiv \gamma_{k+1}I_n.$$

This justifies the canonical form (2.2.5) of the functions $\phi_k(\cdot)$. Further,

$$\phi_{k+1}(x) \overset{(2.2.4)}{=} (1-\alpha_k)\left(\phi_k^* + \tfrac{\gamma_k}{2} \parallel x - v_k \parallel^2\right)$$
$$+ \alpha_k[f(y_k) + \langle \nabla f(y_k), x - y_k\rangle + \tfrac{\mu}{2} \parallel x - y_k \parallel^2].$$

Therefore the equation $\nabla \phi_{k+1}(x) = 0$, which is the first-order optimality condition for the function $\phi_{k+1}(\cdot)$, is as follows:

$$(1-\alpha_k)\gamma_k(x - v_k) + \alpha_k \nabla f(y_k) + \alpha_k \mu(x - y_k) = 0.$$

From this equation, we get a closed form expression for the point $v_{k+1}$, the minimum of the function $\phi_{k+1}(\cdot)$.

Finally, let us compute $\phi_{k+1}^*$. In view of the recurrence (2.2.4) for the sequence $\{\phi_k(\cdot)\}$, we have

$$\begin{aligned} &\phi_{k+1}^* + \tfrac{\gamma_{k+1}}{2} \parallel y_k - v_{k+1} \parallel^2 \overset{(2.2.5)}{=} \phi_{k+1}(y_k) \\ &= (1-\alpha_k)\left(\phi_k^* + \tfrac{\gamma_k}{2} \parallel y_k - v_k \parallel^2\right) + \alpha_k f(y_k). \end{aligned} \tag{2.2.6}$$

By the recursive relation for $v_{k+1}$, we have

$$v_{k+1} - y_k = \tfrac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k(v_k - y_k) - \alpha_k \nabla f(y_k)].$$

Therefore,

$$\tfrac{\gamma_{k+1}}{2} \parallel v_{k+1} - y_k \parallel^2 = \tfrac{1}{2\gamma_{k+1}}[(1-\alpha_k)^2\gamma_k^2 \parallel v_k - y_k \parallel^2$$
$$-2\alpha_k(1-\alpha_k)\gamma_k\langle \nabla f(y_k), v_k - y_k\rangle + \alpha_k^2 \parallel \nabla f(y_k) \parallel^2].$$

It remains to substitute this relation into (2.2.6), taking into account that the multiplicative factor for the term $\parallel y_k - v_k \parallel^2$ in the resulting expression is as follows:

$$(1-\alpha_k)\tfrac{\gamma_k}{2} - \tfrac{1}{2\gamma_{k+1}}(1-\alpha_k)^2\gamma_k^2 = (1-\alpha_k)\tfrac{\gamma_k}{2}\left(1 - \tfrac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\right)$$
$$= (1-\alpha_k)\tfrac{\gamma_k}{2} \cdot \tfrac{\alpha_k \mu}{\gamma_{k+1}}. \qquad \square$$

The situation now is more transparent, and we are close to getting an algorithmic scheme. Indeed, assume that we already have $x_k$:

$$\phi_k^* \geq f(x_k).$$

Then, in view of Lemma 2.2.3,

$$\phi_{k+1}^* \geq (1-\alpha_k)f(x_k) + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}} \parallel \nabla f(y_k) \parallel^2$$

$$+\frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}} \langle \nabla f(y_k), v_k - y_k \rangle.$$

Since $f(x_k) \overset{(2.1.2)}{\geq} f(y_k) + \langle \nabla f(y_k), x_k - y_k \rangle$, we get the following estimate:

$$\phi_{k+1}^* \geq f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}} \parallel \nabla f(y_k) \parallel^2$$

$$+(1-\alpha_k)\langle \nabla f(y_k), \frac{\alpha_k \gamma_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k \rangle.$$

Let us look at this inequality. We want to have $\phi_{k+1}^* \geq f(x_{k+1})$. Recall that we can ensure the inequality

$$f(y_k) - \frac{1}{2L} \parallel \nabla f(y_k) \parallel^2 \geq f(x_{k+1})$$

in many different ways. The simplest one is just to take the gradient step

$$x_{k+1} = y_k - h_k \nabla f(y_k)$$

with $h_k = \frac{1}{L}$ (see (2.1.9)). Let us define $\alpha_k$ as a positive root of the quadratic equation

$$L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k \mu \quad (= \gamma_{k+1}).$$

Then $\frac{\alpha_k^2}{2\gamma_{k+1}} = \frac{1}{2L}$, and we can replace the previous inequality by the following one:

$$\phi_{k+1}^* \geq f(x_{k+1}) + (1-\alpha_k)\langle \nabla f(y_k), \frac{\alpha_k \gamma_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k \rangle.$$

Let us now use our freedom in the choice of $y_k$. It can be found from the equation:

$$\frac{\alpha_k \gamma_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k = 0.$$

This is $y_k = \frac{\alpha_k \gamma_k v_k + \gamma_{k+1} x_k}{\gamma_k + \alpha_k \mu}$, and we come to the following methods, which are often addressed as *Fast Gradient Methods*

**General Scheme of Optimal Method**

**0.** Choose the point $x_0 \in \mathbb{R}^n$, some $\gamma_0 > 0$, and set $v_0 = x_0$.

**1.** ***k*th iteration ($k \geq 0$).**

(a) Compute $\alpha_k \in (0, 1)$ from the equation

$$L\alpha_k^2 = (1 - \alpha_k)\gamma_k + \alpha_k \mu.$$

Set $\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k \mu$.

(b) Choose $y_k = \frac{1}{\gamma_k + \alpha_k \mu} [\alpha_k \gamma_k v_k + \gamma_{k+1} x_k]$. Compute $f(y_k)$ and $\nabla f(y_k)$.

(c) Find $x_{k+1}$ such that

$$f(x_{k+1}) \leq f(y_k) - \frac{1}{2L} \parallel \nabla f(y_k) \parallel^2$$

(see Sect. 1.2.3 for the step-size rules).

(d) Set $v_{k+1} = \frac{1}{\gamma_{k+1}} [(1 - \alpha_k)\gamma_k v_k + \alpha_k \mu y_k - \alpha_k \nabla f(y_k)]$.

(2.2.7)

Note that in Step 1(c) of this scheme we can choose an arbitrary $x_{k+1}$ satisfying the inequality $f(x_{k+1}) \leq f(y_k) - \frac{\omega}{2} \parallel \nabla f(y_k) \parallel^2$ with some $\omega > 0$. Then the constant $\frac{1}{\omega}$ replaces $L$ in the equation of Step 1(a).

**Theorem 2.2.1** *Scheme (2.2.7) generates a sequence of points $\{x_k\}_{k=0}^{\infty}$ such that*

$$f(x_k) - f^* \leq \lambda_k \left[ f(x_0) - f^* + \frac{\gamma_0}{2} \parallel x_0 - x^* \parallel^2 \right],$$

*where $\lambda_0 = 1$ and $\lambda_k = \Pi_{i=0}^{k-1}(1 - \alpha_i)$.*

*Proof* Indeed, let us choose $\phi_0(x) = f(x_0) + \frac{\gamma_0}{2} \parallel x - v_0 \parallel^2$. Then $f(x_0) = \phi_0^*$ and we get $f(x_k) \leq \phi_k^*$ by the rules of the scheme. It remains to use Lemma 2.2.1. □

Thus, in order to estimate the rate of convergence rate of method (2.2.7), we need to understand how quickly the sequence $\{\lambda_k\}$ approaches zero. Define

$$q_f = \frac{1}{Q_f} = \frac{\mu}{L}. \tag{2.2.8}$$

**Lemma 2.2.4** *If in the method (2.2.7) we choose $\gamma_0 \in (\mu, 3L+\mu]$, then for all $k \geq 0$ we have*

$$\lambda_k \leq \frac{4\mu}{(\gamma_0-\mu)\cdot\left[\exp\left(\frac{k+1}{2}q_f^{1/2}\right)-\exp\left(-\frac{k+1}{2}q_f^{1/2}\right)\right]^2} \leq \frac{4L}{(\gamma_0-\mu)(k+1)^2}. \tag{2.2.9}$$

*For $\gamma_0 = \mu$, we have $\lambda_k = \left(1-\sqrt{q_f}\right)^k$, $k \geq 0$.*

*Proof* Let us start from the case $\gamma_0 > \mu$. In accordance with Step 1(a) in (2.2.7),

$$\gamma_{k+1} - \mu = (1-\alpha_k)(\gamma_k - \mu) = \ldots = \lambda_{k+1}(\gamma_0 - \mu). \tag{2.2.10}$$

Since $\alpha_k = 1 - \frac{\lambda_{k+1}}{\lambda_k}$, from the quadratic equation of Step 1(a), we have

$$1 - \frac{\lambda_{k+1}}{\lambda_k} = \left[\frac{\gamma_{k+1}}{L}\right]^{1/2} \overset{(2.2.10)}{=} \left[\frac{\mu}{L} + \lambda_{k+1}\frac{\gamma_0-\mu}{L}\right]^{1/2}.$$

Therefore, $\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k} = \frac{1}{\lambda_{k+1}^{1/2}}\left[\frac{q_f}{\lambda_{k+1}} + \frac{\gamma_0-\mu}{L}\right]^{1/2}$. Thus,

$$\frac{1}{\lambda_{k+1}^{1/2}}\left[\frac{q_f}{\lambda_{k+1}} + \frac{\gamma_0-\mu}{L}\right]^{1/2} \leq \left(\frac{1}{\lambda_{k+1}^{1/2}} + \frac{1}{\lambda_k^{1/2}}\right)\cdot\left(\frac{1}{\lambda_{k+1}^{1/2}} - \frac{1}{\lambda_k^{1/2}}\right) \leq \frac{2}{\lambda_{k+1}^{1/2}}\left(\frac{1}{\lambda_{k+1}^{1/2}} - \frac{1}{\lambda_k^{1/2}}\right).$$

Defining $\xi_k = \left[\frac{L}{(\gamma_0-\mu)\lambda_k}\right]^{1/2}$, we get the following relation:

$$\xi_{k+1} - \xi_k \geq \frac{1}{2}\left[q_f\xi_{k+1}^2 + 1\right]^{1/2}. \tag{2.2.11}$$

Now, for $\delta = \frac{1}{2}\sqrt{q_f}$, we are going to prove by induction that

$$\xi_k \geq \frac{1}{4\delta}\left[e^{(k+1)\delta} - e^{-(k+1)\delta}\right], \quad k \geq 0. \tag{2.2.12}$$

For $k = 0$, in view of the upper bound on $\gamma_0$, we have

$$\xi_0 = \left[\frac{L}{\gamma_0-\mu}\right]^{1/2} \geq \frac{1}{3^{1/2}} > \frac{1}{2}\left[e^{1/2} - e^{-1/2}\right] \geq \frac{1}{4\delta}\left[e^{\delta} - e^{-\delta}\right]$$

since the right-hand side of the above inequality is increasing in $\delta$, and $\delta \leq \frac{1}{2}$.

Thus, for $k = 0$, inequality (2.2.12) is valid. Let us assume that it is valid for some $k \geq 0$. Consider the function $\psi(t) = \frac{1}{4\delta}\left[e^{(t+1)\delta} - e^{-(t+1)\delta}\right]$. Its derivative

$$\psi'(t) = \frac{1}{4}\left[e^{(t+1)\delta} + e^{-(t+1)\delta}\right]$$

is increasing in $t$. Thus, in view of Theorem 2.1.3 the function $\psi(\cdot)$ is convex. In view of our assumption,

$$\psi(t) \leq \xi_k \overset{(2.2.11)}{\leq} \xi_{k+1} - \tfrac{1}{2}\left[q_f \xi_{k+1}^2 + 1\right]^{1/2} \overset{\text{def}}{=} \gamma(\xi_{k+1}).$$

Note that $\gamma'(\xi) = 1 - \frac{1}{2}\frac{q_f \xi}{\left[q_f \xi_{k+1}^2 + 1\right]^{1/2}} > 0$. Suppose that $\xi_{k+1} < \psi(t+1)$. Then

$$\begin{aligned}\psi(t) &< \psi(t+1) - \tfrac{1}{2}\left[4\delta^2 \cdot \left(\tfrac{1}{4\delta}\left[e^{(t+2)\delta} - e^{-(t+2)\delta}\right]\right)^2 + 1\right]^{1/2}\\ &= \psi(t+1) - \tfrac{1}{4}\left[e^{(t+2)\delta} + e^{-(t+2)\delta}\right]\\ &= \psi(t+1) + \psi'(t+1)(t-(t+1)) \overset{(2.1.2)}{\leq} \psi(t).\end{aligned}$$

Thus, we get a contradiction with our second assumption, which proves the lower bound (2.2.12).

For the case $\gamma_0 = \mu$, we have $\gamma_k = \mu$ for all $k \geq 0$ (see (2.2.10)). By the quadratic equation of Step 1(a) in method (2.2.7), this means that $\alpha_k = \sqrt{q_f}$, $k \geq 0$. □

Let us present an exact statement on the optimality of (2.2.7).

**Theorem 2.2.2** *Let us take in (2.2.7)* $\gamma_0 = 3L + \mu$. *Then this scheme generates a sequence* $\{x_k\}_{k=0}^{\infty}$ *such that*

$$f(x_k) - f^* \leq \frac{2(4+q_f)\mu\|x_0 - x^*\|^2}{3\left[\exp\left(\frac{k+1}{2}q_f^{1/2}\right) - \exp\left(-\frac{k+1}{2}q_f^{1/2}\right)\right]^2} \leq \frac{2(4+q_f)L\|x_0-x^*\|^2}{3(k+1)^2}. \tag{2.2.13}$$

*This means that method (2.2.7) is* optimal *for solving the unconstrained minimization problem (2.2.1) with* $f \in \mathscr{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$ *and* $\mu \geq 0$, *when the accuracy* $\epsilon > 0$ *is small enough:*

$$\epsilon \leq \tfrac{\mu}{2}\|x_0 - x^*\|^2. \tag{2.2.14}$$

*If* $\mu = 0$, *then this method is optimal for*

$$\epsilon \leq \tfrac{3L}{32}\|x_0 - x^*\|^2. \tag{2.2.15}$$

*Proof* Indeed, since $f(x_0) - f^* \overset{(2.1.9)}{\leq} \frac{L}{2} \parallel x_0 - x^* \parallel^2$, by Theorem 2.2.1 we have

$$f(x_k) - f^* \leq \tfrac{\lambda_k}{2}(L + \gamma_0)\|x_0 - x^*\|^2.$$

Therefore, by Lemma 2.2.4, we obtain the following bounds:

$$f(x_k) - f^* \le \frac{2\mu(L+\gamma_0)\|x_0-x^*\|^2}{(\gamma_0-\mu)\cdot\left[\exp\left(\frac{k+1}{2}q_f^{1/2}\right)-\exp\left(-\frac{k+1}{2}q_f^{1/2}\right)\right]^2} \le \frac{2L(L+\gamma_0)\|x_0-x^*\|^2}{(\gamma_0-\mu)(k+1)^2}.$$

The upper bounds in the above relations are decreasing in $\gamma_0$. Hence, choosing it as the maximal allowed value, we get inequality (2.2.13).

Let $\mu > 0$. From the lower complexity bounds for the class (see Theorem 2.1.13), we have

$$f(x_k) - f^* \ge \frac{\mu}{2}\left(\frac{\sqrt{Q_f}-1}{\sqrt{Q_f}+1}\right)^{2k} R^2 \ge \frac{\mu}{2}\exp\left(-\frac{4k}{\sqrt{Q_f}-1}\right) R^2,$$

where $R =\| x_0 - x^* \|$. Therefore, the worst case lower bound for finding $x_k$ satisfying $f(x_k) - f^* \le \epsilon$ cannot be better than

$$k \ge \frac{\sqrt{Q_f}-1}{4} \ln \frac{\mu R^2}{2\epsilon} \tag{2.2.16}$$

calls of the oracle (in view of assumption (2.2.14), the right-hand side of this inequality is positive). For our scheme, we have

$$f(x_k) - f^* \overset{(2.2.13)}{\le} \frac{10\mu R^2}{3}\left[e^{(k+1)q_f^{1/2}} - 1\right]^{-1}.$$

Therefore, we guarantee that for $k > \sqrt{Q_f}\ln\left(1+\frac{10\mu R^2}{3\epsilon}\right)$ our problem will be solved. Since

$$\ln\left(1+\frac{10\mu R^2}{3\epsilon}\right) \overset{(2.2.14)}{\le} \ln\left(\frac{\mu R^2}{2\epsilon}+\frac{10\mu R^2}{3\epsilon}\right) = \ln\frac{\mu R^2}{2\epsilon} + \ln\frac{23}{3},$$

the upper bound for the number of iterations (= calls of the oracle) in method (2.2.7) is as follows:

$$\sqrt{Q_f}\cdot\left(\ln\frac{\mu R^2}{2\epsilon} + \ln\frac{23}{3}\right). \tag{2.2.17}$$

Clearly, this bound is proportional to the lower bound (2.2.16). Therefore, the method (2.2.7) is optimal.

The same reasoning can be used for the class $\mathscr{S}^{1,1}_{0,L}(\mathbb{R}^n)$. As above, we need to impose the upper bound (2.2.15) for accuracy in order to have a positive lower bound for the number of calls of the oracle (see Theorem 2.1.7). □

*Remark 2.2.1* Note that the scheme and the complexity analysis of method (2.2.7) is *continuous* in the convexity parameter $\mu$. Therefore, its version for convex functions with Lipschitz continuous gradient has the following rate of convergence:

$$f(x_k) - f^* \overset{(2.2.13)}{\leq} \frac{8L\|x_0-x^*\|^2}{3(k+1)^2}. \tag{2.2.18}$$

Let us analyze a variant of scheme (2.2.7), which uses a constant gradient step for finding the point $x_{k+1}$.

```
Constant Step Scheme I

0. Choose the point $x_0 \in \mathbb{R}^n$, some $\gamma_0 > 0$, and set $v_0 = x_0$.
1. kth iteration ($k \geq 0$).
    (a) Compute $\alpha_k \in (0, 1)$ from the equation

            $L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu.$        (2.2.19)

        Set $\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu$.
    (b) Choose $y_k = \frac{1}{\gamma_k+\alpha_k\mu}[\alpha_k\gamma_k v_k + \gamma_{k+1}x_k]$. Compute
        $f(y_k)$ and $\nabla f(y_k)$.
    (c) Set $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$ and

        $v_{k+1} = \frac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k v_k + \alpha_k\mu y_k - \alpha_k\nabla f(y_k)]$.
```

Let us show that this scheme can be rewritten in a simpler form. Note that

$$y_k = \frac{1}{\gamma_k+\alpha_k\mu}(\alpha_k\gamma_k v_k + \gamma_{k+1}x_k),$$

$$x_{k+1} = y_k - \tfrac{1}{L}\nabla f(y_k),$$

$$v_{k+1} = \frac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k v_k + \alpha_k\mu y_k - \alpha_k\nabla f(y_k)].$$

Therefore,

$$\begin{aligned} v_{k+1} &= \frac{1}{\gamma_{k+1}}\left\{\frac{(1-\alpha_k)}{\alpha_k}[(\gamma_k+\alpha_k\mu)y_k - \gamma_{k+1}x_k] + \alpha_k\mu y_k - \alpha_k\nabla f(y_k)\right\} \\ &= \frac{1}{\gamma_{k+1}}\left\{\frac{(1-\alpha_k)\gamma_k}{\alpha_k}y_k + \mu y_k\right\} - \frac{1-\alpha_k}{\alpha_k}x_k - \frac{\alpha_k}{\gamma_{k+1}}\nabla f(y_k) \\ &= x_k + \frac{1}{\alpha_k}(y_k - x_k) - \frac{1}{\alpha_k L}\nabla f(y_k) \;=\; x_k + \frac{1}{\alpha_k}(x_{k+1} - x_k). \end{aligned}$$

Hence,

$$\begin{aligned} y_{k+1} &= \tfrac{1}{\gamma_{k+1}+\alpha_{k+1}\mu}(\alpha_{k+1}\gamma_{k+1}v_{k+1} + \gamma_{k+2}x_{k+1}) \\ &= x_{k+1} + \tfrac{\alpha_{k+1}\gamma_{k+1}(v_{k+1}-x_{k+1})}{\gamma_{k+1}+\alpha_{k+1}\mu} = x_{k+1} + \beta_k(x_{k+1} - x_k), \end{aligned}$$

where $\beta_k = \frac{\alpha_{k+1}\gamma_{k+1}(1-\alpha_k)}{\alpha_k(\gamma_{k+1}+\alpha_{k+1}\mu)}$. Thus, we managed to eliminate the sequence $\{v_k\}$. Let us do the same with the coefficients $\{\gamma_k\}$. We have

$$\alpha_k^2 L = (1-\alpha_k)\gamma_k + \mu\alpha_k \equiv \gamma_{k+1}.$$

Therefore

$$\begin{aligned} \beta_k &= \tfrac{\alpha_{k+1}\gamma_{k+1}(1-\alpha_k)}{\alpha_k(\gamma_{k+1}+\alpha_{k+1}\mu)} = \tfrac{\alpha_{k+1}\gamma_{k+1}(1-\alpha_k)}{\alpha_k(\gamma_{k+1}+\alpha_{k+1}^2 L-(1-\alpha_{k+1})\gamma_{k+1})} \\ &= \tfrac{\gamma_{k+1}(1-\alpha_k)}{\alpha_k(\gamma_{k+1}+\alpha_{k+1}L)} = \tfrac{\alpha_k(1-\alpha_k)}{\alpha_k^2+\alpha_{k+1}}. \end{aligned}$$

Note also that $\alpha_{k+1}^2 = (1-\alpha_{k+1})\alpha_k^2 + q_f\alpha_{k+1}$, and

$$\alpha_0^2 L = (1-\alpha_0)\gamma_0 + \mu\alpha_0.$$

The latter relation means that $\gamma_0$ can be seen as a function of $\alpha_0$. Thus, we can completely eliminate the sequence $\{\gamma_k\}$. Let us write down the corresponding method.

**Constant Step Scheme II**

**0.** Choose the point $x_0 \in \mathbb{R}^n$, some $\alpha_0 \in (0, 1)$, and set $y_0 = x_0$.

**1.** ***k*th iteration ($k \ge 0$).**

(a) Compute $f(y_k)$ and $\nabla f(y_k)$. Set $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$.

(b) Compute $\alpha_{k+1} \in (0, 1)$ from the equation

$$\alpha_{k+1}^2 = (1-\alpha_{k+1})\alpha_k^2 + q_f\alpha_{k+1}.$$

Set $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2+\alpha_{k+1}}$ and $y_{k+1} = x_{k+1} + \beta_k(x_{k+1} - x_k)$.

(2.2.20)

The rate of convergence of this method can be derived from Theorem 2.2.1 and Lemma 2.2.4. Let us write down the corresponding statement in terms of $\alpha_0$.

**Theorem 2.2.3** *If in the method (2.2.20) we choose $\alpha_0$ in accordance with the conditions*

$$\sqrt{q_f} \leq \alpha_0 \ \leq \ \frac{2(3+q_f)}{3+\sqrt{21+4q_f}}, \tag{2.2.21}$$

*then*

$$\begin{aligned} f(x_k) - f^* &\leq \frac{4\mu\left[f(x_0)-f^*+\frac{\gamma_0}{2}\|x_0-x^*\|^2\right]}{(\gamma_0-\mu)\cdot\left[\exp\left(\frac{k+1}{2}q_f^{1/2}\right)-\exp\left(-\frac{k+1}{2}q_f^{1/2}\right)\right]^2} \\ &\leq \frac{4L}{(\gamma_0-\mu)(k+1)^2}\left[f(x_0) - f^* + \frac{\gamma_0}{2} \parallel x_0 - x^* \parallel^2\right], \end{aligned}$$

*where* $\gamma_0 = \frac{\alpha_0(\alpha_0 L-\mu)}{1-\alpha_0}$.

We do not need to prove this theorem since the initial scheme has not changed. We change only the notation. In Theorem 2.2.3, condition (2.2.21) is equivalent to the condition $\mu \leq \gamma_0 \leq 3L + \mu$ of Lemma 2.2.4.

Scheme (2.2.20) becomes very simple if we choose $\alpha_0 = \sqrt{q_f}$ (this corresponds to $\gamma_0 = \mu$). Then

$$\alpha_k = \sqrt{q_f}, \quad \beta_k = \frac{1-\sqrt{q_f}}{1+\sqrt{q_f}}$$

for all $k \geq 0$. Thus, we come to the following process.

**Constant Step scheme III**

**0.** Choose $y_0 = x_0 \in \mathbb{R}^n$.
**1.** ***k*th iteration ($k \geq 0$).** (2.2.22)

$$x_{k+1} = y_k - \tfrac{1}{L}\nabla f(y_k),$$

$$y_{k+1} = x_{k+1} + \frac{1-\sqrt{q_f}}{1+\sqrt{q_f}}(x_{k+1} - x_k).$$

In accordance with Theorem 2.2.1 and Lemma 2.2.4, it has the following rate of convergence:

$$f(x_k) - f^* \overset{(2.1.9)}{\leq} \frac{L+\mu}{2}\|x_0 - x^*\|^2 e^{-k\sqrt{q_f}}, \quad k \geq 0. \tag{2.2.23}$$

However, this method does not work for $\mu = 0$. The choice of a bigger value of the parameter $\gamma_0$ (which corresponds to another value of $\alpha_0$) is much safer.

Finally, let us prove the following statement.

**Theorem 2.2.4** *Let method (2.2.7) be applied to the function $f \in \mathscr{F}_L^{1,1}(\mathbb{R}^n)$ (this means that $\mu = 0$). Then for any $k \geq 0$ we have*

$$\|v_k - x^*\| \leq \left[1 + \tfrac{1}{\gamma_0} L\right]^{1/2} r_0, \tag{2.2.24}$$

$$\|x_k - x^*\| \leq \left[1 + \tfrac{1}{\gamma_0} L\right]^{1/2} r_0, \tag{2.2.25}$$

*where $r_0 \overset{\mathrm{def}}{=} \|x^* - x_0\|$. Moreover, for the vector $g_k = \frac{\lambda_k}{1-\lambda_k} \sum\limits_{i=0}^{k-1} \frac{\alpha_i}{\lambda_{i+1}} \nabla f(y_k)$, whose coefficients satisfy the equation $\sum\limits_{i=0}^{k-1} \frac{\alpha_i}{\lambda_{i+1}} = \frac{1-\lambda_k}{\lambda_k}$, $k \geq 1$, we have*

$$\|g_k\| \leq \tfrac{\lambda_k \gamma_0}{1-\lambda_k} \left(1 + \left[1 + \tfrac{1}{\gamma_0} L\right]^{1/2}\right) r_0. \tag{2.2.26}$$

*Choosing $\gamma_0 = 3L$, we get the following rate:*

$$\|g_k\| \overset{(2.2.9)}{\leq} \tfrac{4(3+2\sqrt{3})Lr_0}{3(k+1)^2-4}, \quad k \geq 1. \tag{2.2.27}$$

*Proof* As we have seen, method (2.2.7) recursively updates a sequence of estimating functions, which can be represented as follows:

$$\phi_k(x) = \ell_k(x) + \lambda_k(f(x_0) + \tfrac{1}{2}\gamma_0\|x - x_0\|^2), \quad k \geq 0,$$

where $\ell_k(\cdot)$ are linear functions updated by the rules $\ell_0(x) \equiv 0$,

$$\ell_{k+1}(x) = (1 - \alpha_k)\ell_k(x) + \alpha_k[f(y_k) + \langle \nabla f(y_k), x - y_k \rangle], \quad k \geq 0. \tag{2.2.28}$$

Let $\nabla \ell_k \equiv \nabla \ell_k(x)$, $x \in \mathbb{R}^n$.

Note that function $\phi_k$ is strongly convex with convexity parameter $\lambda_k \gamma_0$. Therefore, for all $x \in \mathbb{R}^n$ we have

$$\begin{aligned} f(x_k) + \tfrac{1}{2}\lambda_k\gamma_0\|x - v_k\|^2 &\leq \phi_k^* + \tfrac{1}{2}\lambda_k\gamma_0\|x - v_k\|^2 \overset{(2.1.21)}{\leq} \phi_k(x) \\ &\overset{(2.2.2)}{\leq} f(x) + \lambda_k(f(x_0) + \tfrac{1}{2}\gamma_0\|x - x_0\|^2 - f(x)). \end{aligned}$$

Taking in this inequality $x = x^*$, we get

$$\tfrac{1}{2}\lambda_k\gamma_0\|x^* - v_k\|^2 \le \lambda_k(f(x_0) - f(x^*) + \tfrac{1}{2}\gamma_0\|x^* - x_0\|^2) \overset{(2.1.9)}{\le} \tfrac{1}{2}\lambda_k(L+\gamma_0)r_0^2,$$

and this is the bound (2.2.24).

Let us prove by induction that the bound (2.2.25) holds for all $k \ge 0$. Since $x_0 = v_0$, it holds for $k = 0$. Assume it holds for some $k \ge 0$. Then, in view of Step (b) in (2.2.19), we have $\|y_k - x^*\| \le [1 + \frac{1}{\gamma_0}L]^{1/2}r_0$. It remains to note that the gradient step decreases the distance to the optimal point (see, for example, the proof of Theorem 2.1.14).

Let us look now at the evolution of the vectors $s_k \overset{\mathrm{def}}{=} \frac{1}{\lambda_k}\nabla\ell_k$. Note that $s_0 = 0$ and

$$\begin{aligned}\nabla\ell_{k+1} &\overset{(2.2.28)}{=} (1-\alpha_k)\nabla\ell_k + \alpha_k\nabla f(y_k)\\ &= \tfrac{\lambda_{k+1}}{\lambda_k}\nabla\ell_k + \alpha_k\nabla f(y_k), \quad k \ge 0.\end{aligned}$$

Thus, $s_k = \sum\limits_{i=0}^{k-1}\frac{\alpha_i}{\lambda_{i+1}}\nabla f(y_i)$, $k \ge 0$. On the other hand, for $\tau_i = \frac{\alpha_i}{\lambda_{i+1}}$ we have

$$\tau_i \overset{(2.2.4)}{=} \tfrac{\alpha_i}{(1-\alpha_i)\lambda_i} = \tfrac{1}{\lambda_{i+1}} - \tfrac{1}{\lambda_i}.$$

Thus, $\sum\limits_{i=0}^{k-1}\tau_i = \frac{1}{\lambda_k} - 1$, and $g_k = \frac{\lambda_k s_k}{1-\lambda_k} \equiv \frac{1}{1-\lambda_k}\nabla\ell_k(x)$, $x \in \mathbb{R}^n$. Note that

$$v_k = x_0 - \tfrac{1}{\lambda_k\gamma_0}\nabla\ell_k = x_0 - \tfrac{1-\lambda_k}{\lambda_k\gamma_0}g_k.$$

Hence,

$$\left[1 + \tfrac{1}{\gamma_0}L\right]^{1/2} r_0 \overset{(2.2.24)}{\ge} \|x_0 - \tfrac{1-\lambda_k}{\lambda_k\gamma_0}g_k - x^*\| \ge \tfrac{1-\lambda_k}{\lambda_k\gamma_0}\|g_k\| - r_0,$$

and we get inequality (2.2.26). □

Theorem 2.2.4 can be used to generate points with small gradient of the quadratic function $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$ with $A \succeq 0$. For that, we just compute the point

$$\hat{y}_k = \tfrac{\lambda_k}{1-\lambda_k}\sum_{i=0}^{k-1}\tfrac{\alpha_i}{\lambda_{i+1}}y_i, \quad k \ge 1. \tag{2.2.29}$$

Another example of employing the rule (2.2.29) is given in Sect. 2.2.3.

### 2.2.2 Decreasing the Norm of the Gradient

Sometimes, in solving the optimization problem (2.2.1) with $f \in \mathscr{F}^{1,1}_{\mu,L}(\mathbb{R}^n)$, we are interested in finding a point with small norm of the gradient:

$$\|\nabla f(x)\| \leq \epsilon. \tag{2.2.30}$$

(We will give an important example of this situation in Example 2.2.4 in Sect. 2.2.3.) What are the lower and upper complexity bounds for this goal? Since

$$f(x) - f^* \overset{(2.1.2)}{\leq} \|\nabla f(x)\| \cdot \|x - x^*\|,$$

the corresponding lower complexity bounds must be of the same order as for finding a point with small residual in function value: $f(x) - f^* \leq \epsilon$. Let us see which methods can be used to find points with small gradients.

First of all, let us look at the abilities of Gradient Method (2.1.37) with $h_k = \frac{1}{L}$. Denote $R_0 = \|x_0 - x^*\|$. Let us fix the total number of iterations $T \geq 3$. After the first $k$ iterations, $0 \leq k < T$, we have

$$f(x_k) - f^* \overset{(2.1.39)}{\leq} \frac{2LR_0^2}{k+4}.$$

If $i \geq k$, then $f(x_i) - f(x_{i+1}) \overset{(2.1.9)}{\geq} \frac{1}{2L}\|\nabla f(x_i))\|^2$. Define $g_{k,T} = \min\limits_{k \leq i \leq T} \|\nabla f(x_i)\|$. Then

$$(T-k+1)g_{k,T}^2 \leq \sum_{i=k}^{T} \|\nabla f(x_i)\|^2 \leq 2L \sum_{i=k}^{T}(f(x_i) - f(x_{i+1}))$$
$$= 2L(f(x_k) - f(x_{T+1})) \leq 2L(f(x_k) - f^*) \leq \frac{4L^2R_0^2}{k+4}.$$

Thus, $g_{0,T}^2 \leq \frac{4L^2R_0^2}{(k+4)(T-k+1)}$. We can choose $k$ by maximizing the quadratic function $q(k) = (k+4)(T-k+1)$ for integer $k$. Note that

$$q^* \overset{\text{def}}{=} \max_{k \in \mathbb{Z}} q(k) \geq q(\tau^* + \tfrac{1}{2}), \qquad \tau^* = \arg\max_{\tau \in \mathbb{R}} q(\tau).$$

Since $\tau^* = \frac{T-3}{2}$, we get $q^* \geq q(\frac{T-2}{2}) = \frac{1}{4}(T+4)(T+6)$.

Thus, we have proved the following theorem.

**Theorem 2.2.5** *Let $f \in \mathscr{F}_L^{1,1}(\mathbb{R}^n)$ and choose in method (2.1.37) $h_k = \frac{1}{L}$. Then for the total number of steps in this method $T \geq 3$ we have*

$$g_{0,T} \leq \frac{4LR_0}{[(T+4)(T+6)]^{1/2}}. \tag{2.2.31}$$

Thus, the Gradient Method ensures the goal (2.2.30) in $O(\frac{1}{\epsilon})$ iterations. Let us see what happens with a monotone version of the Optimal Method (2.2.19) in the case $\mu = 0$.

**Monotone Constant Step Scheme $\mathbf{I_A}$**

```
0. Choose the point $x_0 \in \mathbb{R}^n$. Set $\lambda_0 = 1$ and $v_0 = x_0$.
1. kth iteration ($k \geq 0$).
    (a) Compute $\alpha_k \in (0, 1)$ from equation $\alpha_k^2 = 3(1-\alpha_k)\lambda_k$.
    (b) Set $y_k = \alpha_k v_k + (1-\alpha_k)x_k$ and $\lambda_{k+1} = (1-\alpha_k)\lambda_k$.
    (c) Compute $\nabla f(y_k)$ and set $\hat{x}_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$.
    (d) Define $v_{k+1} = v_k - \frac{1}{L\alpha_k}\nabla f(y_k)$.
    (e) Set $\hat{y}_k = \arg\min\left\{f(y) :\ y \in \{x_k, \hat{x}_{k+1}\}\right\}$.
    (f) Compute $\nabla f(\hat{y}_k)$ and set $x_{k+1} = \hat{y}_k - \frac{1}{L}\nabla f(\hat{y}_k)$.
```

(2.2.32)

This scheme corresponds to the method (2.2.7) with $\gamma_0 = 3L$ and $\mu = 0$. Hence, $\gamma_k \equiv 3L\lambda_k$. Note that it ensures a monotone decrease of the objective function:

$$f(x_k) \overset{(2.2.32)_e}{\geq} f(\hat{y}_k) \overset{(2.2.32)_f}{\geq} f(x_{k+1}) + \frac{1}{2L}\|\nabla f(\hat{y}_k)\|^2. \tag{2.2.33}$$

As before, we divide the total number of iterations $T \geq 3$ into two parts. After the first $k$ iterations, $0 \leq k < T$, we have

$$f(x_k) - f^* \overset{(2.2.18)}{\leq} \frac{8LR_0^2}{3(k+1)^2}.$$

If $i \geq k$, then $f(x_i) - f(x_{i+1}) \overset{(2.2.33)}{\geq} \frac{1}{2L}\|\nabla f(\hat{y}_i))\|^2$. Define $g_{k,T} = \min\limits_{k \leq i \leq T} \|\nabla f(\hat{y}_i)\|$. Then

$$\begin{aligned}(T-k+1)g_{k,T}^2 &\leq \sum_{i=k}^{T} \|\nabla f(\hat{y}_i)\|^2 \ \leq\ 2L\sum_{i=k}^{T}(f(x_i) - f(x_{i+1}))\\ &= 2L(f(x_k) - f(x_{T+1})) \ \leq\ 2L(f(x_k) - f^*) \ \leq\ \frac{16L^2R_0^2}{3(k+1)^2}.\end{aligned}$$

Thus, $g_{0,T}^2 \le \frac{16L^2R_0^2}{3(k+1)^2(T-k+1)}$. We can choose $k$ by maximizing the cubic function $q(k) = (k+1)^2(T-k+1)$ for integer $k$. Note that $k^*$, the optimal solution of the problem $q^* \overset{\rm def}{=} \max\limits_{k \in \mathbb{Z}} q(k)$, belongs to the interval $[\tau^* - \frac{1}{2}, \tau^* + \frac{1}{2}]$, where $\tau^* = \arg \max\limits_{\tau \in \mathbb{R}_+} q(\tau)$. Moreover, since the function $q(\cdot)$ is concave in this interval, we have

$$\begin{aligned} q^* &\ge \min\{q(\tau^* - \tfrac{1}{2}), q(\tau^* + \tfrac{1}{2})\} \\ &= \min_{\delta = \pm\frac{1}{2}} \left\{ q(\tau^*) + \tfrac{1}{2} q''(\tau^*)(\tfrac{1}{2})^2 + \tfrac{1}{6} q'''(\tau^*)\delta^3 \right\} \\ &= q(\tau^*) + \tfrac{1}{8} q''(\tau^*) - \tfrac{1}{8}. \end{aligned}$$

Note that $q'(\tau) = (\tau+1)(2T+1-3\tau)$ and $q''(\tau) = 2T - 2 - 6k$. Therefore, $\tau^* = \frac{2T+1}{3}$, $q''(\tau^*) = -2T-4$, and $q(\tau^*) = \frac{4}{27}(T+2)^3$. Hence,

$$q^* \ge \tfrac{4}{27}(T+2)^3 - \tfrac{1}{4}(T+2) - \tfrac{1}{8}.$$

Thus, we have proved the following theorem.

**Theorem 2.2.6** *If $f \in \mathscr{F}_L^{1,1}(\mathbb{R}^n)$, then method (2.2.32) ensures the following rate of decrease for the norm of the gradient:*

$$g_{0,T} \le \frac{4LR_0}{[\frac{4}{3}(T+2)^3 - \frac{9}{4}(T+2) - \frac{9}{8}]^{1/2}}, \quad T \ge 1. \tag{2.2.34}$$

Thus, the Optimal Method (2.2.32) ensures the goal (2.2.30) in $O(\frac{1}{\epsilon^{2/3}})$ iterations. Let us show that we can be even faster if we apply a regularization technique.

Let us fix a regularization parameter $\delta > 0$ and consider the following function:

$$f_\delta(x) = f(x) + \tfrac{1}{2}\delta\|x - x_0\|^2.$$

In view of conditions (2.1.12) and (2.1.22), $f_\delta \in \mathscr{S}_{\delta, L+\delta}^{1,1}(\mathbb{R}^n)$. Denote by $x_\delta^*$ its unique optimal point, which satisfies the equation

$$\nabla f(x_\delta^*) + \delta(x_\delta^* - x_0) = 0. \tag{2.2.35}$$

Note that

$$f_\delta(x_\delta^*) + \tfrac{1}{2}\delta\|x_\delta^* - x^*\|^2 \overset{(2.1.21)}{\le} f_\delta(x^*) \;=\; f(x^*) + \tfrac{1}{2}\delta\|x^* - x^0\|^2.$$

Since $f(x^*) \leq f(x^*_\delta)$, we conclude that

$$\|x^*_\delta - x_0\|^2 + \|x^*_\delta - x^*\|^2 \leq \|x_0 - x^*\|^2. \tag{2.2.36}$$

Thus, by choosing an appropriate $\delta$, we can make the gradient $\nabla f(x^*_\delta)$ small:

$$\|\nabla f(x^*_\delta)\| \overset{(2.2.35)}{=} \delta\|x^*_\delta - x_0\| \overset{(2.2.36)}{\leq} \delta R_0.$$

Therefore, it is possible to find a point with small norm of the gradient by minimizing the function $f_\delta$. Let us estimate the complexity of this process.

Let us use for our goal the scheme (2.2.22) with parameters $L+\delta$ and $q_f = \frac{\delta}{\delta+L}$. Then after $T$ iterations of this method, we have

$$\begin{aligned} \|\nabla f(x_T)\| &\leq \|\nabla f(x^*_\delta)\| + \|\nabla f(x_T) - \nabla f(x^*_\delta)\| \overset{(1.2.8)}{\leq} \delta R_0 + L\|x_T - x^*_\delta\| \\ &\overset{(2.1.21)}{\leq} \delta R_0 + L\left[\tfrac{2}{\delta}(f_\delta(x_T) - f_\delta(x^*_\delta))\right]^{1/2} \\ &\overset{(2.2.23)}{\leq} \delta R_0 + L\left[\tfrac{L+2\delta}{\delta} R_0^2 e^{-T\sqrt{q_f}}\right]^{1/2}. \end{aligned}$$

Thus, choosing $\delta$ from condition $\delta R_0 = \frac{1}{2}\epsilon$, we get $\frac{1}{q_f} = 1 + \frac{2LR_0}{\epsilon}$. Therefore, the number of steps $T$ in our scheme is bounded by the solution of the following inequality:

$$LR_0\left[\tfrac{L+2\delta}{\delta}\right]^{1/2} \leq \tfrac{\epsilon}{2}\, e^{T\sqrt{q_f}/2}.$$

This is $T \geq \frac{2}{\sqrt{q_f}} \ln\left(\left(\frac{1}{q_f} - 1\right)\left(1 + \frac{1}{q_f}\right)^{1/2}\right)$. Thus, we have proved the following theorem

**Theorem 2.2.7** *Let $f \in \mathscr{F}^{1,1}_L(\mathbb{R}^n)$ and $\delta = \frac{\epsilon}{2R_0}$. Then the number of steps $T$ which is necessary for method (2.2.22) to generate a point $x_T$ with $\|\nabla f(x_T)\| \leq \epsilon$ by minimizing the function $f_\delta$ is bounded as follows:*

$$T \leq 3\sqrt{1 + \tfrac{2LR_0}{\epsilon}}\ \ln\left(1 + \tfrac{2LR_0}{\epsilon}\right). \tag{2.2.37}$$

Thus, up to a logarithmic factor, the complexity estimate of the regularization scheme is optimal. To the best of our knowledge, it is not known yet if this factor can be dropped.

### 2.2.3 Convex Sets

The next step in generalizing the unconstrained minimization problem (2.1.36) is a constrained minimization problem with no functional constraints:

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

where $Q$ is a *convex set* of $\mathbb{R}^n$. We have already introduced these sets in Definition 2.1.1, as natural domains of convex functions. Now we will need them as *simple constraints*.

Let us look at two important examples of convex sets.

**Lemma 2.2.5** *If $f(\cdot)$ is a convex function on $\mathbb{R}^n$, then for any $\beta \in \mathbb{R}$ its* level set

$$\mathscr{L}_f(\beta) = \{x \in \mathbb{R}^n \mid f(x) \leq \beta\}$$

*is either convex or empty.*

*Proof* Indeed, let $x$ and $y$ belong to $\mathscr{L}_f(\beta)$. Then $f(x) \leq \beta$ and $f(y) \leq \beta$. Therefore,

$$f(\alpha x + (1-\alpha)y) \overset{(2.1.3)}{\leq} \alpha f(x) + (1-\alpha) f(y) \leq \beta,$$

which means $\alpha x + (1-\alpha)y \in \mathscr{L}_f(\beta)$. □

**Lemma 2.2.6** *Let $f(\cdot)$ be a convex function on $\mathbb{R}^n$. Then its* epigraph

$$\mathscr{E}_f = \{(x,\tau) \in \mathbb{R}^{n+1} \mid f(x) \leq \tau\}$$

*is a convex set.*

*Proof* Indeed, let $z_1 = (x_1, \tau_1) \in \mathscr{E}_f$ and $z_2 = (x_2, \tau_2) \in \mathscr{E}_f$. Then for any $\alpha \in [0, 1]$ we have

$$z_\alpha \equiv \alpha z_1 + (1-\alpha) z_2 = (\alpha x_1 + (1-\alpha)x_2, \alpha\tau_1 + (1-\alpha)\tau_2),$$

$$f(\alpha x_1 + (1-\alpha)x_2) \overset{(2.1.3)}{\leq} \alpha f(x_1) + (1-\alpha) f(x_2) \leq \alpha\tau_1 + (1-\alpha)\tau_2.$$

Thus, $z_\alpha \in \mathscr{E}_f$. □

Let us consider now the most important *operations* with convex sets.

**Theorem 2.2.8** *Let $Q_1 \subseteq \mathbb{R}^n$ and $Q_2 \subseteq \mathbb{R}^m$ be closed convex sets, and $\mathscr{A}(\cdot)$ be a linear operator:*

$$\mathscr{A}(x) = Ax + b : \ \mathbb{R}^n \ \to \ \mathbb{R}^m.$$

1. *The intersection of two sets ($m = n$), $Q_1 \bigcap Q_2 = \{x \in \mathbb{R}^n \mid x \in Q_1, \ x \in Q_2\}$, is convex and closed.*
2. *The sum of two sets ($m = n$), $Q_1 + Q_2 = \{z = x + y \mid x \in Q_1, \ y \in Q_2\}$, is convex. It is closed provided that one of the sets is bounded.*
3. *The direct product of two sets, $Q_1 \times Q_2 = \{(x, y) \in \mathbb{R}^{n+m} \mid x \in Q_1, \ y \in Q_2\}$ is convex and closed.*
4. *The conic hull of a set, $\mathscr{K}(Q_1) = \{z \in \mathbb{R}^n \mid z = \beta x, \ x \in Q_1, \beta \geq 0\}$, is convex. It is closed if the set $Q_1$ is bounded and does not contain the origin.*
5. *The convex hull of two sets,*

$$\mathrm{Conv}(Q_1, Q_2) = \{z \in \mathbb{R}^n \mid z = \alpha x + (1 - \alpha)y, \ x \in Q_1, \ y \in Q_2, \ \alpha \in [0, 1]\},$$

   *is convex. It is closed if both sets are bounded.*
6. *The affine image of a set, $\mathscr{A}(Q_1) = \{y \in \mathbb{R}^m \mid y = \mathscr{A}(x), \ x \in Q_1\}$, is convex and closed.*
7. *The inverse affine image: $\mathscr{A}^{-1}(Q_2) = \{x \in \mathbb{R}^n \mid \mathscr{A}(x) \in Q_2\}$ is convex. It is closed if $Q_2$ is bounded.*

*Proof*

1. If $x_1 \in Q_1 \bigcap Q_2$ and $x_1 \in Q_1 \bigcap Q_2$, then $[x_1, x_2] \subset Q_1$ and $[x_1, x_2] \subset Q_2$. Therefore, $[x_1, x_2] \subset Q_1 \bigcap Q_2$. Closedness of intersection is evident.
2. If $z_1 = x_1 + y_1$ with $x_1 \in Q_1$, $y_1 \in Q_2$, and $z_2 = x_2 + y_2$ with $x_2 \in Q_1$, $y_2 \in Q_2$, then

$$\alpha z_1 + (1 - \alpha)z_2 = [\alpha x_1 + (1 - \alpha)x_2]_1 + [\alpha y_1 + (1 - \alpha)y_2]_2,$$

   where $[\cdot]_1 \in Q_1$ and $[\cdot]_2 \in Q_2$. Let us assume now that the set $Q_2$ is bounded. Consider a convergent sequence $z_k = x_k + y_k \to \bar{z}$ with $\{x_k\} \subset Q_1$ and $\{y_k\} \subset Q_2$. Since $Q_2$ is bounded, we can assume that the whole sequence $\{y_k\}$ converges (otherwise, select a converging subsequence). Then, the sequence $\{x_k\}$ also converges. This implies the inclusion $\bar{z} \in Q_1 + Q_2$.
3. If $z_1 = (x_1, y_1)$, $x_1 \in Q_1$, $y_1 \in Q_2$ and $z_2 = (x_2, y_2)$, $x_2 \in Q_1$, $y_2 \in Q_2$, then

$$\alpha z_1 + (1 - \alpha)z_2 = ([\alpha x_1 + (1 - \alpha)x_2]_1, [\alpha y_1 + (1 - \alpha)y_2]_2),$$

   where $[\cdot]_1 \in Q_1$ and $[\cdot]_2 \in Q_2$. Further, if a sequence $\{z_k = (x_k, y_k)\} \subset Q_1 \times Q_2$ converges to $\bar{z} = (\bar{x}, \bar{y})$, this means that $x_k \to \bar{x} \in Q_1$ and $y_k \to \bar{y} \in Q_2$. Hence, the point $\bar{z}$ belongs to $Q_1 \times Q_2$.

4. If $z_1 = \beta_1 x_1$ with $x_1 \in Q_1$ and $\beta_1 \geq 0$, and $z_2 = \beta_2 x_2$ with $x_2 \in Q_1$ and $\beta_2 \geq 0$, then for any $\alpha \in [0, 1]$ we have

$$\alpha z_1 + (1-\alpha) z_2 = \alpha\beta_1 x_1 + (1-\alpha)\beta_2 x_2 = \gamma(\bar{\alpha} x_1 + (1-\bar{\alpha}) x_2),$$

where $\gamma = \alpha\beta_1 + (1-\alpha)\beta_2$, and $\bar{\alpha} = \alpha\beta_1/\gamma \in [0, 1]$. Thus, the set $\mathscr{K}(Q_1)$ is convex.

Consider a convergent sequence $\{z_k = \beta_k x_k \to \bar{z}\}$ with $\{x_k\} \subset Q_1$. If $Q_1$ is bounded, then the sequence $\{x_k\}$ is bounded. If $0 \notin Q_1$, then the sequence $\{\beta_k\}$ is also bounded. Therefore, without loss of generality, we can assume that both sequences $\{\beta_k\}$ and $\{x_k\}$ are convergent. Hence, $\bar{z} \in \mathscr{K}(Q_1)$ and we conclude that this cone is closed.

5. If $z_1 = \beta_1 x_1 + (1-\beta_1) y_1$ with $x_1 \in Q_1$, $y_1 \in Q_2$, and $\beta_1 \in [0, 1]$, and $z_2 = \beta_2 x_2 + (1-\beta_2) y_2$ with $x_2 \in Q_1$, $y_2 \in Q_2$, and $\beta_2 \in [0, 1]$, then for any $\alpha \in [0, 1]$ we have

$$\begin{aligned}\alpha z_1 + (1-\alpha) z_2 &= \alpha(\beta_1 x_1 + (1-\beta_1) y_1) + (1-\alpha)(\beta_2 x_2 + (1-\beta_2) y_2)\\ &= \bar{\alpha}(\bar{\beta}_1 x_1 + (1-\bar{\beta}_1) x_2) + (1-\bar{\alpha})(\bar{\beta}_2 y_1 + (1-\bar{\beta}_2) y_2),\end{aligned}$$

where $\bar{\alpha} = \alpha\beta_1 + (1-\alpha)\beta_2$ and $\bar{\beta}_1 = \alpha\beta_1/\bar{\alpha}$, $\bar{\beta}_2 = \alpha(1-\beta_1)/(1-\bar{\alpha})$.

Let us assume that both sets are bounded. Considering now a convergent sequence $\{z_k = \beta_k x_k + (1-\beta_k) y_k \to \bar{z}\}$ with $\{\beta_k\} \subset [0, 1]$, $\{x_k\} \subset Q_1$, and $\{y_k\} \subset Q_2$, without loss of generality, we can assume that all these sequences are convergent. This implies that $\bar{z} \in \mathrm{Conv}\{Q_1, Q_2\}$.

6. If $y_1, y_2 \in \mathscr{A}(Q_1)$ then $y_1 = Ax_1 + b$ and $y_2 = Ax_2 + b$ for some $x_1, x_2 \in Q_1$. Therefore, for $y(\alpha) = \alpha y_1 + (1-\alpha) y_2$, $0 \leq \alpha \leq 1$, we have

$$y(\alpha) = \alpha(Ax_1 + b) + (1-\alpha)(Ax_2 + b) = A(\alpha x_1 + (1-\alpha) x_2) + b.$$

Thus, $y(\alpha) \in \mathscr{A}(Q_1)$. This set is closed in view of the continuity of linear operators.

7. If $x_1, x_2 \in \mathscr{A}^{-1}(Q_2)$ then $Ax_1 + b = y_1$ and $Ax_2 + b = y_2$ for some $y_1, y_2 \in Q_2$. Therefore, for $x(\alpha) = \alpha x_1 + (1-\alpha) x_2$, $0 \leq \alpha \leq 1$, we have

$$\begin{aligned}\mathscr{A}(x(\alpha)) &= A(\alpha x_1 + (1-\alpha) x_2) + b\\ &= \alpha(Ax_1 + b) + (1-\alpha)(Ax_2 + b) = \alpha y_1 + (1-\alpha) y_2 \in Q_2.\end{aligned}$$

Let $Q_2$ be bounded. Consider a convergent sequence $\{x_k \to \bar{x}\} \subset \mathscr{A}^{-1}(Q_2)$. Then, without loss of generality, we can assume that the sequence $\{y_k = \mathscr{A}(x_k)\} \subset Q_2$ is convergent to a point $\bar{y} \in Q_2$. Since $\bar{y} = A(\bar{x})$, we conclude that $\bar{x} \in \mathscr{A}^{-1}(Q_2)$. Thus, the inverse image of a bounded set is closed. □

Let us give examples justifying the additional assumptions of Theorem 2.2.8, which were introduced to ensure closedness of the results of some operations with convex sets.

*Example 2.2.1* In all examples below, we work with an unbounded convex set

$$Q = \left\{ x \in \mathbb{R}^2_+ : \ x^{(2)} \geq \tfrac{1}{x^{(1)}} \right\}.$$

- *Sum of two sets.* Consider the set $\mathbb{R}^{1,2}_+ \overset{\mathrm{def}}{=} \left\{ x \in \mathbb{R}^2 : \ x^{(1)} \geq 0, \ x^{(2)} = 0 \right\}$. Then

$$Q - \mathbb{R}^{1,2}_+ = \left\{ x \in \mathbb{R}^2 : \ x^{(2)} > 0 \right\}$$

  is an open set. At the same time, $Q + \mathbb{R}^{1,2}_+ \equiv Q$ is closed.
- *Conic hull.* Let $0_2 = (0,0)^T \in \mathbb{R}^2$. The set

$$\mathscr{K}(Q) \equiv \left\{ x \in \mathbb{R}^2 : \ x^{(1)} > 0, \ x^{(2)} > 0 \right\} \bigcup \{ 0_2 \}$$

  is not closed. Also, for $Q_1 = \left\{ x \in \mathbb{R}^2 : \ \| x - e_1 \| \leq 1 \right\}$, we have

$$\mathscr{K}(Q_1) = \left\{ x \in \mathbb{R}^2 : \ x^{(1)} > 0 \right\} \bigcup \{0_2\},$$

  which is not closed.
- *Convex hull.* Note that $\mathrm{Conv}\{0_2, Q\} = \mathscr{K}(Q)$, and the latter set is not closed.
- *Inverse affine image.* Note that

$$\{ x \in \mathbb{R} : \ \exists \tau > 0 \text{ such that } (\tau, x) \in Q \} = \{ x \in \mathbb{R} : \ x > 0 \},$$

  and this set is open. □

Using the statements above, we can justify the convexity of some important sets.

*Example 2.2.2*

1. *Half-space*. The set $\{ x \in \mathbb{R}^n \mid \ \langle a, x \rangle \leq \beta \}$ is convex since linear function is convex.
2. *Polytope*. The set $\{ x \in \mathbb{R}^n \mid \ \langle a_i, x \rangle \leq b_i, \ i = 1 \ldots m \}$ is convex as an intersection of convex sets.
3. *Ellipsoid*. Let $A = A^T \succeq 0$. Then the set $\{ x \in \mathbb{R}^n \mid \ \langle Ax, x \rangle \leq r^2 \}$ is convex since the function $\langle Ax, x \rangle$ is convex. □

Let us consider now a smooth optimization problem with the *set constraint*:

$$\min_{x \in Q} f(x), \quad f \in \mathscr{F}^1(Q, \| \cdot \|), \tag{2.2.38}$$

where $Q$ is a closed convex set. We assume that the optimal set of this problem $X^*$ is not empty. Our current goal consists in describing the *optimality conditions* for problem (2.2.38). It is clear that the old condition

$$\nabla f(x) = 0$$

does not work here.

*Example 2.2.3* Consider the following univariate minimization problem:

$$\min_{x \geq 0} \; x.$$

Here $Q = \{x \in \mathbb{R} : \; x \geq 0\}$, and $f(x) = x$. Note that $x^* = 0$, but $\nabla f(x^*) = 1 > 0$. □

**Theorem 2.2.9** *Let $f \in \mathscr{F}^1(Q)$ and the set $Q$ be closed and convex. A point $x^*$ is a solution of problem (2.2.38) if and only if*

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0 \tag{2.2.39}$$

*for all $x \in Q$.*

*Proof* Indeed, if (2.2.39) is true, then

$$f(x) \overset{(2.1.2)}{\geq} f(x^*) + \langle \nabla f(x^*), x - x^* \rangle \overset{(2.2.39)}{\geq} f(x^*)$$

for all $x \in Q$.

Let $x^*$ be a solution to (2.2.38). Assume that there exists some $x \in Q$ such that

$$\langle \nabla f(x^*), x - x^* \rangle < 0.$$

Consider the function $\phi(\alpha) = f(x^* + \alpha(x - x^*))$, $\alpha \in [0, 1]$. Note that

$$\phi(0) = f(x^*), \quad \phi'(0) = \langle \nabla f(x^*), x - x^* \rangle < 0.$$

Therefore, for $\alpha$ small enough we have

$$f(x^* + \alpha(x - x^*)) = \phi(\alpha) < \phi(0) = f(x^*).$$

This is a contradiction. □

The next statement is often addressed as the *growth property* of strongly convex functions.

**Corollary 2.2.1** *If $f \in \mathscr{S}^1_\mu(Q, \|\cdot\|)$, then for any $x \in Q$, we have*

$$f(x) \geq f(x^*) + \tfrac{\mu}{2}\|x - x^*\|^2. \tag{2.2.40}$$

*Proof* Indeed,

$$\begin{aligned} f(x) &\overset{(2.1.20)}{\geq} f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \tfrac{\mu}{2}\|x - x^*\|^2 \\ &\overset{(2.2.39)}{\geq} f(x^*) + \tfrac{\mu}{2}\|x - x^*\|^2. \qquad \square \end{aligned}$$

**Corollary 2.2.2** *Let $f \in C^{1,1}_L(\mathbb{R}^n, \|\cdot\|)$. Then, for any two points $x_1^*, x_2^* \in X^*$, we have*

$$\nabla f(x_1^*) = \nabla f(x_2^*), \quad \langle \nabla f(x_1^*), x_1^* \rangle = \langle \nabla f(x_2^*), x_2^* \rangle. \tag{2.2.41}$$

*Proof* Indeed, $\langle \nabla f(x_1^*), x_2^* - x_1^* \rangle \overset{(2.2.39)}{\geq} 0$ and $\langle \nabla f(x_2^*), x_1^* - x_2^* \rangle \overset{(2.2.39)}{\geq} 0$. Adding these two inequalities, we have

$$0 \geq \langle \nabla f(x_1^*) - \nabla f(x_2^*), x_1^* - x_2^* \rangle \overset{(2.1.11)}{\geq} \tfrac{1}{L}\|\nabla f(x_1^*) - \nabla f(x_2^*)\|_*^2.$$

For $x^* \in X^*$, let $g^* = \nabla f(x^*)$. Then,

$$\begin{aligned} 0 &\overset{(2.2.39)}{\geq} \langle \nabla f(x_2^*), x_2^* - x_1^* \rangle \overset{(2.2.41)}{=} \langle g^*, x_2^* - x_1^* \rangle \\ &\overset{(2.2.41)}{=} \langle \nabla f(x_1^*), x_2^* - x_1^* \rangle \overset{(2.2.39)}{\geq} 0. \qquad \square \end{aligned}$$

Let us now prove the existence theorem.

**Theorem 2.2.10** *Let $f \in \mathscr{S}^1_\mu(Q, \|\cdot\|)$ with $\mu > 0$ and the set $Q$ be closed and convex. Then there exists a unique solution $x^*$ of problem (2.2.38).*

*Proof* Let $x_0 \in Q$. Consider the set $\bar{Q} = \{x \in Q \mid f(x) \leq f(x_0)\}$. Note that the problem (2.2.38) is equivalent to the following

$$\min\{f(x) \mid x \in \bar{Q}\}. \tag{2.2.42}$$

However, the set $\bar{Q}$ is bounded: for all $x \in \bar{Q}$, we have

$$f(x_0) \geq f(x) \overset{(2.1.20)}{\geq} f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \tfrac{\mu}{2} \| x - x_0 \|^2 .$$

Hence, $\| x - x_0 \| \leq \frac{2}{\mu} \| \nabla f(x_0) \|_*$.

Thus, the solution $x^*$ of problem (2.2.42) ($\equiv$ (2.2.38)) exists. Let us prove that it is unique. Indeed, if $x_1^*$ is also an optimal solution to (2.2.38), then

$$f^* \;=\; f(x_1^*) \overset{(2.2.40)}{\geq} f^* + \tfrac{\mu}{2} \parallel x_1^* - x^* \parallel^2 .$$

Therefore $x_1^* = x^*$. □

*Example 2.2.4* Let $f \in \mathscr{F}_\mu^1(Q, \|\cdot\|_p)$. Consider the following *primal* minimization problem:

$$f^* \;=\; \min_{x \in Q}\{f(x) : \; Ax = b\}, \tag{2.2.43}$$

where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. In some applications the set $Q$ and function $f$ are very simple, and the complexity of this problem is related to the nontrivial intersection of the linear constraints with the set $Q$. In these cases, it is recommended to solve problem (2.2.43) by *dualizing* the linear constraints.

Let us introduce dual multipliers for equality constraints, and define the Lagrangian

$$\mathscr{L}(x, u) = f(x) + \langle u, b - Ax \rangle, \quad x \in Q, \; u \in \mathbb{R}^m.$$

Now we can define the dual function $\phi(u) = \min\limits_{x \in Q} \mathscr{L}(x, u)$. By Theorem 2.2.10, this function is well defined for all $u \in \mathbb{R}^m$. Let $x(u) = \arg\min\limits_{x \in Q} \mathscr{L}(x, u) \in Q$ and let $g(u) = b - Ax(u)$. Note that for arbitrary $u_1$ and $u_2 \in \mathbb{R}^m$ we have

$$\begin{aligned} \phi(u_1) &= f(x(u_1)) + \langle u_1, b - Ax(u_1) \rangle \;\leq\; f(x(u_2)) + \langle u_1, b - Ax(u_2) \rangle \\ &= \phi(u_2) + \langle u_1 - u_2, g(u_2) \rangle. \end{aligned}$$

Let us introduce in $\mathbb{R}^m$ the norm $\| \cdot \|_d$. Define

$$\|A\|_{p,d} = \max_{x,u}\{\langle Ax, u \rangle : \; \|x\|_p \leq 1, \|u\|_d \leq 1\} \overset{(2.1.6)}{=} \max_u\{\|A^T u\|_{p*} : \; \|u\|_d \leq 1\}.$$

Then, for any $u_1, u_2 \in \mathbb{R}^n$ we have

$$\langle \nabla f(x(u_2)), x(u_1) - x(u_2) \rangle \overset{(2.2.39)}{\geq} \langle A^T u_2, x(u_1) - x(u_2) \rangle. \tag{2.2.44}$$

Therefore,

$$\begin{aligned}
\phi(u_1) &= f(x(u_1)) + \langle u_1, b - Ax(u_1)\rangle \\
&\overset{(2.1.20)}{\geq} f(x(u_2)) + \langle \nabla f(x(u_2)), x(u_1) - x(u_2)\rangle + \tfrac{1}{2}\mu\|x(u_1) - x(u_2)\|_p^2 \\
&\quad + \langle u_1, b - Ax(u_1)\rangle \\
&\overset{(2.2.44)}{\geq} f(x(u_2)) + \langle u_2, A(x(u_1) - x(u_2))\rangle + \tfrac{1}{2}\mu\|x(u_1) - x(u_2)\|_p^2 \\
&\quad + \langle u_1, b - Ax(u_1)\rangle \\
&= \phi(u_2) + \langle g(u_2), u_1 - u_2\rangle - \langle u_1 - u_2, A(x(u_1) - x(u_2))\rangle \\
&\quad + \tfrac{1}{2}\mu\|x(u_1) - x(u_2)\|_p^2 \\
&\geq \phi(u_2) + \langle g(u_2), u_1 - u_2\rangle - \tfrac{1}{2\mu}(\|A^T(u_1 - u_2)\|_p^*)^2.
\end{aligned}$$

Since $\phi$ is concave, $g(u) = \nabla\phi(u)$ and $-\phi \overset{(2.1.9)}{\in} \mathscr{F}_L^{1,1}(\mathbb{R}^m, \|\cdot\|_d)$ with $L = \frac{1}{\mu}\|A\|_{p,d}^2$.

Now we can solve the Lagrangian dual problem

$$\min_{u\in\mathbb{R}^m}\{-\phi(u)\} \tag{2.2.45}$$

by any method for minimizing smooth convex functions. Assuming that the solution of this problem $u^*$ exists, we have

$$0 = \nabla\phi(u^*) \;=\; b - Ax(u^*).$$

Thus, $x(u^*)$ is feasible for problem (2.2.43). On the other hand,

$$f^* \overset{(1.3.6)}{\geq} f_* \overset{\text{def}}{=} \max_{u\in\mathbb{R}^m} \phi(u) \;=\; f(x(u^*)) + \langle u^*, \nabla\phi(u^*)\rangle \;=\; f(x(u^*)).$$

Hence, $f^* = f_*$ and $x(u^*)$ is the optimal solution of problem (2.2.43).

Now, assume that $\bar{u} \in \mathbb{R}^m$ is an approximate solution to the dual problem (2.2.45). Then it is clear that the norm of the gradient of the objective function at this point is very important. Indeed, it bounds the residual $b - A(x(\bar{u}))$. On the other hand,

$$f(x(\bar{u})) - f^* = \phi(\bar{u}) - \langle \bar{u}, \nabla\phi(\bar{u})\rangle - \phi(u^*) \;\leq\; \|\bar{u}\|_d \cdot \|\nabla\phi(\bar{u})\|_d^*.$$

Thus, the size of the gradient of the dual function bounds at the same time the level of infeasibility and the level of optimality.

We have already discussed in Sect. 2.2.2 how to compute a point with small norm of the gradient. However, for problem (2.2.45) the situation is even simpler. Indeed, Theorem 2.2.4 shows that the *average gradient* at points $\{y_k\}$ decreases as $O(\frac{1}{k^2})$. For problem (2.2.45), this means that the residual of the linear system $Ax = b$ at some average point of the sequence $\{x(v_k)\} \subset Q$ (with points $\{v_k\}$ corresponding to $\{y_k\}$ in method (2.2.7)) decreases as $O(\frac{1}{k^2})$. So, these average points can be taken as approximate solutions to the primal problem (2.2.43). □

To conclude this section, let us analyze the properties of *Euclidean projection* onto the convex set. Up to the end of this section the notation $\| \cdot \|$ is used for the standard Euclidean norm.

**Definition 2.2.2** Let $Q$ be a closed set and $x_0 \in \mathbb{R}^n$. Define

$$\pi_Q(x_0) = \arg\min_{x \in Q} \ \| x - x_0 \| . \tag{2.2.46}$$

We call $\pi_Q(x_0)$ the *Euclidean projection* of the point $x_0$ onto the set $Q$.

Let $f(x) = \frac{1}{2} \| x \|^2$. Since $\nabla^2 f(x) = I_n$, this function belongs to the class $\mathscr{S}_1^2(\mathbb{R}^n)$.

**Theorem 2.2.11** *If $Q$ is a convex set, then there exists a unique projection $\pi_Q(x_0)$.*

*Proof* Indeed, $\pi_Q(x_0) = \arg\min\limits_{x \in Q} f(x)$, where $f \in \mathscr{S}_{1,1}^{1,1}(\mathbb{R}^n)$. Therefore $\pi_Q(x_0)$ is unique and well defined in view of Theorem 2.2.10. □

Since $Q$ is closed, $\pi_Q(x_0) = x_0$ if and only if $x_0 \in Q$.

**Lemma 2.2.7** *Let $Q$ be a closed convex set and $x_0 \notin Q$. Then for any $x \in Q$, we have*

$$\langle \pi_Q(x_0) - x_0, x - \pi_Q(x_0) \rangle \geq 0. \tag{2.2.47}$$

*Proof* Note that $\pi_Q(x_0)$ is a solution of the minimization problem $\min\limits_{x \in Q} f(x)$ with $f(x) = \frac{1}{2} \| x - x_0 \|^2$. Therefore, in view of Theorem 2.2.9 we have

$$\langle \nabla f(\pi_Q(x_0)), x - \pi_Q(x_0) \rangle \geq 0$$

for all $x \in Q$. It remains to note that $\nabla f(x) = x - x_0$. □

**Corollary 2.2.3** *For any two points $x_1$ and $x_2 \in \mathbb{R}^n$, we have*

$$\|\pi_Q(x_1) - \pi_Q(x_2)\| \leq \|x_1 - x_2\|. \tag{2.2.48}$$

*Proof* Indeed, in view of inequality (2.2.47), we have

$$\langle \pi_Q(x_1) - x_1, \pi_Q(x_2) - \pi_Q(x_1) \geq 0,$$

$$\langle \pi_Q(x_2) - x_2, \pi_Q(x_1) - \pi_Q(x_2) \geq 0.$$

Adding these two inequalities, we get

$$\begin{aligned} \|\pi_Q(x_1) - \pi_Q(x_2)\|^2 \;&\leq\; \langle \pi_Q(x_1) - \pi_Q(x_2), x_1 - x_2 \rangle \\ &\leq\; \|\pi_Q(x_1) - \pi_Q(x_2)\| \cdot \|x_1 - x_2\|. \end{aligned}$$

□

Let us also mention a *triangle inequality* for projection (compare with (2.2.36)).

**Lemma 2.2.8** *For any two point $x \in Q$ and $y \in \mathbb{R}^n$, we have*

$$\| \, x - \pi_Q(y) \, \|^2 + \| \, \pi_Q(y) - y \, \|^2 \leq \| \, x - y \, \|^2 \, . \tag{2.2.49}$$

*Proof* Indeed, in view of (2.2.47), we have

$$\begin{aligned} \| \, x - \pi_Q(y) \, \|^2 - \| \, x - y \, \|^2 = \; &\langle y - \pi_Q(y), 2x - \pi_Q(y) - y \rangle \\ \leq \; &- \| \, y - \pi_Q(y) \, \|^2 \, . \end{aligned}$$

□

There exists a useful characterization of optimal solutions to problem (2.2.38) in terms of Euclidean projection.

**Theorem 2.2.12** *Let $x^*$ be an optimal solution to problem (2.2.38). Then, for any $\gamma > 0$ we have*

$$\pi_Q(x^* - \tfrac{1}{\gamma} \nabla f(x^*)) = x^*. \tag{2.2.50}$$

*Proof* Consider the minimization problem $\min\limits_{x \in Q} \frac{1}{2} \| x - x^* + \frac{1}{\gamma} \nabla f(x^*) \|^2$. Its objective function is strongly convex. Hence, in view of Theorem 2.2.10, its solution $x_*$ exists and is unique. Moreover, in view of Theorem 2.2.9, it is completely characterized by the following inequality:

$$\langle x_* - x^* + \tfrac{1}{\gamma} \nabla f(x^*), x - x_* \rangle \geq 0, \quad \forall x \in Q.$$

Hence, $x_* = x^*$. □

Finally, let us mention some properties of the *distance function* to a convex set:

$$\rho_Q(x) \overset{\text{def}}{=} \tfrac{1}{2}\|x - \pi_Q(x)\|^2, \quad x \in \mathbb{R}^n. \tag{2.2.51}$$

**Lemma 2.2.9** *A function $\rho_Q$ is convex and differentiable on $\mathbb{R}^n$ with gradient*

$$\nabla \rho_Q(x) = x - \pi_Q(x), \quad x \in \mathbb{R}^n, \tag{2.2.52}$$

*which is Lipschitz continuous in the standard Euclidean norm with constant one.*

*Proof* Let us fix two arbitrary points $x_1$ and $x_2$ in $\mathbb{R}^n$. Let $\pi_1 = \pi_Q(x_1) \in Q$, $\pi_2 = \pi_Q(x_2) \in Q$, $g_1 = x_1 - \pi_1$, and $g_2 = x_2 - \pi_2$. In view of the *Euclidean identity*

$$\tfrac{1}{2}\|g_2\|^2 = \tfrac{1}{2}\|g_1\|^2 + \langle g_1, g_2 - g_1\rangle + \tfrac{1}{2}\|g_2 - g_1\|^2, \tag{2.2.53}$$

we have

$$\begin{aligned} \rho_Q(x_2) \quad &\geq \quad \rho_Q(x_1) + \langle x_1 - \pi_Q(x_1), x_2 - x_1\rangle \\ &\qquad + \langle \pi_Q(x_1) - x_1, \pi_Q(x_2) - \pi_Q(x_1)\rangle \\ &\overset{(2.2.47)}{\geq} \quad \rho_Q(x_1) + \langle g_1, x_2 - x_1\rangle. \end{aligned}$$

On the other hand,

$$\begin{aligned} \rho_Q(x_2) - \rho_Q(x_1) &\overset{(2.2.53)}{=} \langle g_1, g_2 - g_1\rangle + \tfrac{1}{2}\|g_2 - g_1\|^2 \\ &= \langle g_1, x_2 - x_1\rangle + \langle g_1, \pi_1 - \pi_2 - g_2\rangle + \tfrac{1}{2}\|g_1\|^2 + \tfrac{1}{2}\|g_2\|^2 \\ &\overset{(2.2.46)}{\leq} \langle g_1, x_2 - x_1\rangle + \langle g_1, \pi_1 - x_2\rangle + \tfrac{1}{2}\|g_1\|^2 + \tfrac{1}{2}\|x_2 - \pi_1\|^2 \\ &= \langle g_1, x_2 - x_1\rangle + \tfrac{1}{2}\|x_2 - x_1\|^2. \end{aligned}$$

Thus, for arbitrary points $x_1$ and $x_2 \in \mathbb{R}^n$ we have proved the following relations:

$$\langle g_1, x_2 - x_1\rangle \leq \rho_Q(x_2) - \rho_Q(x_1) \;\leq\; \langle g_1, x_2 - x_1\rangle + \tfrac{1}{2}\|x_2 - x_1\|^2.$$

Hence the function $\rho_Q$ is differentiable at any point $x \in \mathbb{R}^n$ and $\nabla \rho_Q(x) = x - \pi_Q(x)$. Moreover, in view of condition (2.1.9), $f \in \mathscr{F}_1^{1,1}(\mathbb{R}^n)$. □

### 2.2.4 The Gradient Mapping

As compared with the unconstrained problem, in the constrained minimization problem (2.2.38), the gradient of the objective function should be treated differently. In the previous section, we have already seen that its role in optimality conditions is changing. Moreover, we can no longer use it for the gradient step since the result may be infeasible. If we look at the main properties of the gradient, which are useful for functions from the class $\mathscr{F}_L^{1,1}(\mathbb{R}^n)$, we can see that two of them are of the highest importance. The first is that the step along the direction of the anti-gradient decreases the function value by an amount comparable with the squared norm of the gradient:

$$f(x - \tfrac{1}{L}\nabla f(x)) \le f(x) - \tfrac{1}{2L} \parallel \nabla f(x) \parallel^2 .$$

The second is the inequality

$$\langle \nabla f(x), x - x^* \rangle \ge \tfrac{1}{L} \parallel \nabla f(x) \parallel^2 .$$

It turns out that for Constrained Minimization we can introduce an object which inherits both these important properties.

**Definition 2.2.3** Let us fix some $\gamma > 0$. Define

$$\begin{aligned} x_Q(\bar{x};\gamma) &= \arg\min_{x \in Q} \left[ f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \tfrac{\gamma}{2} \parallel x - \bar{x} \parallel^2 \right], \\ g_Q(\bar{x};\gamma) &= \gamma(\bar{x} - x_Q(\bar{x};\gamma)). \end{aligned} \tag{2.2.54}$$

We call $x_Q(\bar{x}, \gamma)$ the *gradient mapping*, and $g_Q(\bar{x}, \gamma)$ the *reduced gradient* of the function $f$ on $Q$.

Note that the objective function of the optimization problem in this definition can be written as

$$f(\bar{x}) + \tfrac{\gamma}{2}\|x - \bar{x} + \tfrac{1}{\gamma}\nabla f(\bar{x})\|^2 - \tfrac{1}{2\gamma}\|\nabla f(\bar{x})\|^2. \tag{2.2.55}$$

Thus, $x_Q(\bar{x};\gamma)$ is a projection of point $\bar{x} - \frac{1}{\gamma}\nabla f(\bar{x})$ onto the feasible set. For $Q \equiv \mathbb{R}^n$, we have

$$x_Q(\bar{x};\gamma) = \bar{x} - \tfrac{1}{\gamma}\nabla f(\bar{x}), \quad g_Q(\bar{x};\gamma) = \nabla f(\bar{x}).$$

The value $\frac{1}{\gamma}$ can be seen as a natural step size for the "gradient" step

$$\bar{x} \to x_Q(\bar{x};\gamma) \overset{(2.2.54)}{=} \bar{x} - \tfrac{1}{\gamma} g_Q(\bar{x};\gamma). \tag{2.2.56}$$

Note that the gradient mapping is well defined in view of Theorem 2.2.10. Moreover, it is defined for all $\bar{x} \in \mathbb{R}^n$, not necessarily from $Q$.

Let us write down the main property of the gradient mapping.

**Theorem 2.2.13** *Let $f \in \mathscr{S}^{1,1}_{\mu,L}(Q)$, $\gamma \geq L$, and $\bar{x} \in \mathbb{R}^n$. Then for any $x \in Q$, we have*

$$f(x) \geq f(x_Q(\bar{x};\gamma)) + \langle g_Q(\bar{x};\gamma), x - \bar{x}\rangle + \frac{1}{2\gamma} \parallel g_Q(\bar{x};\gamma) \parallel^2 + \frac{\mu}{2} \parallel x - \bar{x} \parallel^2 . \tag{2.2.57}$$

*Proof* Let $x_Q = x_Q(\gamma, \bar{x})$, $g_Q = g_Q(\gamma, \bar{x})$, and

$$\phi(x) = f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \frac{\gamma}{2} \parallel x - \bar{x} \parallel^2 .$$

Then $\nabla\phi(x) = \nabla f(\bar{x}) + \gamma(x - \bar{x})$, and for any $x \in Q$ we have

$$\langle \nabla f(\bar{x}) - g_Q, x - x_Q\rangle = \langle \nabla\phi(x_Q), x - x_Q\rangle \overset{(2.2.39)}{\geq} 0.$$

Hence,

$$\begin{aligned}
f(x) - \frac{\mu}{2} \parallel x - \bar{x} \parallel^2 &\overset{(2.1.20)}{\geq} f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle \\
&= f(\bar{x}) + \langle \nabla f(\bar{x}), x_Q - \bar{x}\rangle + \langle \nabla f(\bar{x}), x - x_Q\rangle \\
&\geq f(\bar{x}) + \langle \nabla f(\bar{x}), x_Q - \bar{x}\rangle + \langle g_Q, x - x_Q\rangle \\
&= \phi(x_Q) - \frac{\gamma}{2} \parallel x_Q - \bar{x} \parallel^2 + \langle g_Q, x - x_Q\rangle \\
&= \phi(x_Q) - \frac{1}{2\gamma} \parallel g_Q \parallel^2 + \langle g_Q, x - x_Q\rangle \\
&= \phi(x_Q) + \frac{1}{2\gamma} \parallel g_Q \parallel^2 + \langle g_Q, x - \bar{x}\rangle,
\end{aligned}$$

and $\phi(x_Q) \overset{(2.1.9)}{\geq} f(x_Q)$ since $\gamma \geq L$. □

**Corollary 2.2.4** *Let $f \in \mathscr{S}^{1,1}_{\mu,L}(Q)$, $\gamma \geq L$, and $\bar{x} \in Q$. Then*

$$f(x_Q(\bar{x};\gamma)) \leq f(\bar{x}) - \frac{1}{2\gamma} \parallel g_Q(\bar{x};\gamma) \parallel^2, \tag{2.2.58}$$

$$\begin{aligned}
\langle g_Q(\bar{x};\gamma), \bar{x} - x^*\rangle \geq \frac{1}{2\gamma} \parallel g_Q(\bar{x};\gamma) \parallel^2 + \frac{\mu}{2} \parallel \bar{x} - x^* \parallel^2 \\
+\frac{\mu}{2}\|x_Q(\bar{x};\gamma) - x^*\|^2.
\end{aligned} \tag{2.2.59}$$

*Proof* Indeed, using (2.2.57) with $x = \bar{x}$, we get (2.2.58). Using (2.2.57) with $x = x^*$, we get (2.2.59) since

$$f(x_Q(\bar{x};\gamma)) \overset{(2.2.40)}{\geq} f(x^*) + \tfrac{\mu}{2}\|x_Q(\bar{x};\gamma) - x^*\|^2. \qquad \square$$

### 2.2.5 Minimization over Simple Sets

Let us show that we can use the gradient mapping to solve the following problem:

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

where $f \in \mathscr{S}^{1,1}_{\mu,L}(Q)$ and $Q$ is a closed convex set. We assume that the set $Q$ is simple enough, so the gradient mapping can be computed by a closed form expression. This assumption is valid for some simple sets like positive orthants, $n$ dimensional boxes, simplexes, Euclidean balls, and some others.

Let us start with the Gradient Method.

**Gradient Method for Simple Set**

**0.** Choose a starting point $x_0 \in Q$ and a parameter $\gamma > 0$. (2.2.60)

**1. $k$th iteration ($k \geq 0$).**

$$x_{k+1} = x_k - \tfrac{1}{\gamma}\, g_Q(x_k;\gamma).$$

Note that in this scheme

$$x_{k+1} \overset{(2.2.56)}{=} x_Q(x_k;\gamma) = \pi_Q\left(x_k - \tfrac{1}{\gamma}\nabla f(x_k)\right). \tag{2.2.61}$$

The efficiency analysis of this scheme is very similar to the analysis of its unconstrained version.

**Theorem 2.2.14** *Let $f \in \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$. If in (2.2.60) $\gamma \geq \frac{L+\mu}{2}$, then*

$$\| x_k - x^* \| \leq \left(1 - \tfrac{\mu}{\gamma}\right)^k \| x_0 - x^* \|.$$

*Proof* Let $r_k = \| x_k - x^* \|$. Then, in view of Theorem 2.2.12, we have

$$\begin{aligned} r_{k+1}^2 &\overset{(2.2.61)}{=} \|\pi_Q(x_k - \tfrac{1}{\gamma}\nabla f(x_k)) - \pi_Q(x^* - \tfrac{1}{\gamma}\nabla f(x^*))\|^2 \\ &\overset{(2.2.48)}{\leq} \|x_k - x^* - \tfrac{1}{\gamma}(\nabla f(x_k) - \nabla f(x^*))\|^2 \\ &= r_k^2 - \tfrac{2}{\gamma}\langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle + \tfrac{1}{\gamma^2}\|\nabla f(x_k) - \nabla f(x^*)\|^2 \\ &\overset{(2.1.32)}{\leq} \left(1 - \tfrac{2}{\gamma} \cdot \tfrac{\mu L}{\mu + L}\right) r_k^2 + \left(\tfrac{1}{\gamma^2} - \tfrac{2}{\gamma} \cdot \tfrac{1}{\mu+L}\right) \|\nabla f(x_k) - \nabla f(x^*)\|^2 \\ &\overset{(2.1.26)}{\leq} \left(1 - \tfrac{2}{\gamma} \cdot \tfrac{\mu L}{\mu+L} + \mu^2 \left(\tfrac{1}{\gamma^2} - \tfrac{2}{\gamma} \cdot \tfrac{1}{\mu+L}\right)\right) r_k^2 \; = \; \left(1 - \tfrac{\mu}{\gamma}\right)^2 r_k^2. \qquad \square \end{aligned}$$

Thus, for the minimal value of the scaling parameter $\gamma = \frac{L+\mu}{2}$, method (2.2.60) has the same rate of convergence as for the unconstrained scheme (2.1.37):

$$\| x_k - x^* \| \leq \left(\tfrac{L-\mu}{L+\mu}\right)^k \| x_0 - x^* \| . \qquad (2.2.62)$$

Consider now the optimal schemes. We give only a sketch of their justification since it is very similar to the analysis of Sect. 2.2.1.

First of all, we define the estimating sequences. Assume that $x_0 \in Q$. Define

$$\begin{aligned} \phi_0(x) &= f(x_0) + \tfrac{\gamma_0}{2} \| x - x_0 \|^2, \\ \phi_{k+1}(x) &= (1-\alpha_k)\phi_k(x) + \alpha_k[f(x_Q(y_k; L)) + \tfrac{1}{2L} \| g_Q(y_k; L) \|^2 \\ &\quad + \langle g_Q(y_k; L), x - y_k \rangle + \tfrac{\mu}{2} \| x - y_k \|^2], \quad k \geq 0. \end{aligned}$$

Note that the recursive rule for updating the estimating functions $\phi_k(\cdot)$ has changed. The reason is that now we have to use inequality (2.2.57) instead of (2.1.20). However, this modification does not change the functional terms in the recursion, only the constant terms are affected. Therefore, it is possible to keep all complexity results of Sect. 2.2.1.

It is easy to see that the estimating sequence $\{\phi_k(x\cdot)\}$ can be represented in the canonical form

$$\phi_k(x) = \phi_k^* + \tfrac{\gamma_k}{2} \| x - v_k \|^2,$$

with the following recursive rules for $\gamma_k$, $v_k$ and $\phi_k^*$:

$$\begin{aligned}
\gamma_{k+1} &= (1-\alpha_k)\gamma_k + \alpha_k \mu,\\
v_{k+1} &= \tfrac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k v_k + \alpha_k \mu y_k - \alpha_k g_Q(y_k; L)],\\
\phi_{k+1}^* &= (1-\alpha_k)\phi_k^* + \alpha_k f(x_Q(y_k; L)) + \left(\tfrac{\alpha_k}{2L} - \tfrac{\alpha_k^2}{2\gamma_{k+1}}\right) \parallel g_Q(y_k; L) \parallel^2\\
&\quad + \tfrac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}} \left(\tfrac{\mu}{2} \parallel y_k - v_k \parallel^2 + \langle g_Q(y_k; L), v_k - y_k \rangle\right).
\end{aligned}$$

Further, assuming that $\phi_k^* \geq f(x_k)$ and using the inequality

$$\begin{aligned}
f(x_k) \overset{(2.2.57)}{\geq} & f(x_Q(y_k; L)) + \langle g_Q(y_k; L), x_k - y_k \rangle\\
& + \tfrac{1}{2L} \parallel g_Q(y_k; L) \parallel^2 + \tfrac{\mu}{2} \parallel x_k - y_k \parallel^2],
\end{aligned}$$

we come to the following lower bound:

$$\begin{aligned}
\phi_{k+1}^* \geq\ & (1-\alpha_k) f(x_k) + \alpha_k f(x_Q(y_k; L)) + \left(\tfrac{\alpha_k}{2L} - \tfrac{\alpha_k^2}{2\gamma_{k+1}}\right) \parallel g_Q(y_k; L) \parallel^2\\
& + \tfrac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}} \langle g_Q(y_k; L), v_k - y_k \rangle\\
\geq\ & f(x_Q(y_k; L)) + \left(\tfrac{1}{2L} - \tfrac{\alpha_k^2}{2\gamma_{k+1}}\right) \parallel g_Q(y_k; L) \parallel^2\\
& + (1-\alpha_k)\langle g_Q(y_k; L), \tfrac{\alpha_k \gamma_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k \rangle.
\end{aligned}$$

Thus, again we can choose

$$\begin{aligned}
x_{k+1} &= x_Q(y_k; L),\\
L\alpha_k^2 &= (1-\alpha_k)\gamma_k + \alpha_k \mu \equiv \gamma_{k+1},\\
y_k &= \tfrac{1}{\gamma_k + \alpha_k \mu}(\alpha_k \gamma_k v_k + \gamma_{k+1} x_k).
\end{aligned}$$

Let us write down the corresponding variant of scheme (2.2.20).

**Constant Step Scheme II for Simple Set** (2.2.63)

```
0. Choose $x_0 \in \mathbb{R}^n$ and $\alpha_0 \in \left[\sqrt{q_f}, \frac{2(3+q_f)}{3+\sqrt{21+4q}}\right]$. Set $y_0 = x_0$.
1. kth iteration ($k \geq 0$).
   (a) Compute $f(y_k)$ and $\nabla f(y_k)$. Set $x_{k+1} = x_Q(y_k; L)$.
   (b) Compute $\alpha_{k+1} \in (0, 1)$ from the equation
           $\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + q_f \alpha_{k+1}$.
       Set $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2+\alpha_{k+1}}$ and $y_{k+1} = x_{k+1} + \beta_k(x_{k+1} - x_k)$.
```

The rate of convergence of this method is given by Theorem 2.2.3. Note that only the points $\{x_k\}$ are feasible for $Q$. The sequence $\{y_k\}$ is used for computing the gradient mapping and it may be infeasible.

## 2.3 The Minimization Problem with Smooth Components

(Minimax problems: Gradient Mapping, Gradient Method, Optimal Methods; Problem with functional constraints; Methods for Constrained Minimization.)

### *2.3.1 The Minimax Problem*

Very often, the objective function in optimization problems is composed of several functional components. For example, the reliability of a complex system is usually defined as the minimal reliability of its parts. A constrained minimization problem with functional constraints also provides us with an example of the interaction of several nonlinear functions, etc.

The simplest problem of this type is called the *(discrete) minimax* problem. In this section, we consider the following *smooth* minimax problem:

$$\min_{x \in Q} \left[ f(x) = \max_{1 \leq i \leq m} f_i(x) \right], \tag{2.3.1}$$

where $f_i \in \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n, \|\cdot\|)$, $i = 1\dots m$, and $Q$ is a closed convex set. We call the function $f$ a *max-type* function composed of *components* $f_i(x)$. We write $f \in \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n, \|\cdot\|)$ if all components of the function $f$ belong to this class.

Note that in general, $f$ is not differentiable. However, provided that all $f_i$ are differentiable functions, we can introduce an object, which behaves exactly as a linear approximation of the differentiable function.

**Definition 2.3.1** Let $f$ be a max-type function:

$$f(x) = \max_{1\le i\le m} f_i(x).$$

The function

$$f(\bar{x}; x) = \max_{1\le i\le m} [f_i(\bar{x}) + \langle \nabla f_i(\bar{x}), x - \bar{x}\rangle],$$

is called the *linearization* of $f$ at the point $\bar{x}$.

Compare the following result with inequalities (2.1.20) and (2.1.9).

**Lemma 2.3.1** *For any two points $x$ and $\bar{x}$ in $\mathbb{R}^n$, we have*

$$f(x) \ge f(\bar{x}; x) + \tfrac{\mu}{2} \parallel x - \bar{x} \parallel^2, \tag{2.3.2}$$

$$f(x) \le f(\bar{x}; x) + \tfrac{L}{2} \parallel x - \bar{x} \parallel^2 . \tag{2.3.3}$$

*Proof* Indeed, for all $i = 1, \dots, m$, we have

$$f_i(x) \overset{(2.1.20)}{\ge} f_i(\bar{x}) + \langle \nabla f_i(\bar{x}), x - \bar{x}\rangle + \tfrac{\mu}{2} \parallel x - \bar{x} \parallel^2 .$$

Taking the maximum of these inequalities in $i$, we get (2.3.2).

To prove (2.3.3), we use inequalities

$$f_i(x) \overset{(2.1.9)}{\le} f_i(\bar{x}) + \langle \nabla f_i(\bar{x}), x - \bar{x}\rangle + \tfrac{L}{2} \parallel x - \bar{x} \parallel^2, \quad i = 1, \dots, m. \qquad \square$$

Let us write down the optimality conditions for problem (2.3.1) (compare with Theorem 2.2.9).

**Theorem 2.3.1** *The point $x^* \in Q$ is an optimal solution to problem (2.3.1) if and only if for any $x \in Q$ we have*

$$f(x^*; x) \ \ge \ f(x^*; x^*) \ = \ f(x^*). \tag{2.3.4}$$

*Proof* Indeed, if condition (2.3.4) holds, then

$$f(x) \overset{(2.3.2)}{\geq} f(x^*; x) \ \geq \ f(x^*; x^*) \ = \ f(x^*)$$

for all $x \in Q$.

Let $x^*$ be an optimal solution to (2.3.1). Assume that there exists an $x \in Q$ such that $f(x^*; x) < f(x^*)$. Consider the functions

$$\phi_i(\alpha) = f_i(x^* + \alpha(x - x^*)), \quad i = 1 \dots m.$$

Note that for all $i$, $1 \leq i \leq m$, we have

$$f_i(x^*) + \langle \nabla f_i(x^*), x - x^* \rangle < f(x^*) = \max_{1 \leq i \leq m} f_i(x^*).$$

Therefore, either $\phi_i(0) \equiv f_i(x^*) < f(x^*)$, or

$$\phi_i(0) = f(x^*), \quad \phi_i'(0) = \langle \nabla f_i(x^*), x - x^* \rangle < 0.$$

Thus, for $\alpha$ small enough, we have

$$f_i(x^* + \alpha(x - x^*)) = \phi_i(\alpha) < f(x^*)$$

for all $i$, $1 \leq i \leq m$. This is a contradiction. □

**Corollary 2.3.1** *Let $x^*$ be a minimum of the max-type function $f(\cdot)$ on the set $Q$. If $f$ belongs to $\mathscr{S}^1_\mu(\mathbb{R}^n, \| \cdot \|)$, then*

$$f(x) \geq f(x^*) + \tfrac{\mu}{2} \parallel x - x^* \parallel^2$$

*for all $x \in Q$.*

*Proof* Indeed, in view of (2.3.2) and Theorem 2.3.1, for any $x \in Q$, we have

$$\begin{aligned} f(x) &\geq f(x^*; x) + \tfrac{\mu}{2} \parallel x - x^* \parallel^2 \geq \ f(x^*; x^*) + \tfrac{\mu}{2} \parallel x - x^* \parallel^2 \\ &= f(x^*) + \tfrac{\mu}{2} \parallel x - x^* \parallel^2 . \end{aligned}$$

□

Finally, let us prove an existence theorem.

**Theorem 2.3.2** *Let the max-type function $f$ belong to the class $\mathscr{S}^1_\mu(\mathbb{R}^n, \| \cdot \|)$ with $\mu > 0$, and $Q$ be a closed convex set. Then there exists a unique optimal solution $x^*$ to problem (2.3.1).*

*Proof* Let $\bar{x} \in Q$. Consider the set $\bar{Q} = \{x \in Q \mid f(x) \leq f(\bar{x})\}$. Note that the problem (2.3.1) is equivalent to the following problem:

$$\min\{f(x) \mid x \in \bar{Q}\}. \tag{2.3.5}$$

However, the set $\bar{Q}$ is bounded: for any $x \in \bar{Q}$ we have

$$f(\bar{x}) \geq f_i(x) \overset{(2.1.20)}{\geq} f_i(\bar{x}) + \langle \nabla f_i(\bar{x}), x - \bar{x}\rangle + \tfrac{\mu}{2} \parallel x - \bar{x} \parallel^2, \quad i = 1, \dots, m.$$

Consequently,

$$\tfrac{\mu}{2} \parallel x - \bar{x} \parallel^2 \leq \parallel \nabla f_i(\bar{x}) \parallel_* \cdot \parallel x - \bar{x} \parallel + f(\bar{x}) - f_i(\bar{x}), \quad i = 1, \dots, m.$$

Thus, the solution $x^*$ of (2.3.5) (and of (2.3.1)) exists.

If $x_1^*$ is another solution to (2.3.1), then

$$f(x^*) = f(x_1^*) \overset{(2.3.2)}{\geq} f(x^*; x_1^*) + \tfrac{\mu}{2} \parallel x_1^* - x^* \parallel^2 \overset{(2.3.4)}{\geq} f(x^*) + \tfrac{\mu}{2} \parallel x_1^* - x^* \parallel^2 .$$

Therefore, $x_1^* = x^*$. □

### 2.3.2 Gradient Mapping

In Sect. 2.2.4, we introduced the reduced gradient, which replaces the usual gradient for a constrained minimization problem over a simple set. Since linearization of a max-type function behaves similarly to the linearization of a smooth function, we can adapt this notion to our particular situation. Up to the end of this chapter, we will be working with the standard Euclidean norm.

Let us fix some $\gamma > 0$ and point $\bar{x} \in \mathbb{R}^n$. For a max-type function $f$, define

$$f_\gamma(\bar{x}; x) = f(\bar{x}; x) + \tfrac{\gamma}{2} \parallel x - \bar{x} \parallel^2 .$$

The following definition is an extension of Definition 2.2.3.

**Definition 2.3.2** Define

$$f^*(\bar{x}; \gamma) = \min_{x \in Q} f_\gamma(\bar{x}; x),$$

$$x_f(\bar{x}; \gamma) = \arg\min_{x \in Q} f_\gamma(\bar{x}; x),$$

$$g_f(\bar{x}; \gamma) = \gamma(\bar{x} - x_f(\bar{x}; \gamma)).$$

We call $x_f(x;\gamma)$ the *Gradient Mapping* and $g_f(\bar{x};\gamma)$ the *Reduced Gradient* of a max-type function $f$ on $Q$.

For $m=1$, this definition is equivalent to Definition 2.2.3. Note that the point of *linearization* $\bar{x}$ does not necessarily belong to $Q$. At the same time, now the point $x_f(\bar{x};\gamma)$ cannot be interpreted as a projection (2.2.55).

It is clear that $f_\gamma(\bar{x};\cdot)$ is a max-type function composed by the components

$$f_i(\bar{x}) + \langle \nabla f_i(\bar{x}), x-\bar{x}\rangle + \tfrac{\gamma}{2} \parallel x-\bar{x} \parallel^2 \in \mathscr{S}^{1,1}_{\gamma,\gamma}(\mathbb{R}^n), \quad i = 1\ldots m.$$

Therefore, the gradient mapping is well defined (see Theorem 2.3.2).

Let us now prove the main result of this section, which highlights the similarity between the properties of the Gradient Mapping and the properties of the reduced gradient (compare with Theorem 2.2.13).

**Theorem 2.3.3** *For all $x\in Q$, $\gamma\geq L$, and $\bar{x}\in\mathbb{R}^n$, we have*

$$f(\bar{x};x) \geq f^*(\bar{x};\gamma) + \langle g_f(\bar{x};\gamma), x-\bar{x}\rangle + \tfrac{1}{2\gamma} \parallel g_f(\bar{x};\gamma) \parallel^2 . \tag{2.3.6}$$

*Proof* Let $x_f = x_f(\bar{x};\gamma)$, $g_f = g_f(\bar{x};\gamma)$. It is clear that $f_\gamma(\bar{x};\cdot)\in\mathscr{S}^{1,1}_{\gamma,\gamma}(\mathbb{R}^n)$ and it is a max-type function. Therefore, all results of the previous section can also be applied to the function $f_\gamma$.

Since $x_f = \arg\min\limits_{x\in Q} f_\gamma(\bar{x};x)$, in view of Corollary 2.3.1 and Theorem 2.3.1, we have

$$\begin{aligned}
f(\bar{x};x) &= f_\gamma(\bar{x};x) - \tfrac{\gamma}{2} \parallel x-\bar{x} \parallel^2\\
&\geq f_\gamma(\bar{x};x_f) + \tfrac{\gamma}{2}(\parallel x-x_f \parallel^2 - \parallel x-\bar{x} \parallel^2)\\
&\geq f^*(\bar{x};\gamma) + \tfrac{\gamma}{2}\langle \bar{x}-x_f, 2x-x_f-\bar{x}\rangle\\
&= f^*(\bar{x};\gamma) + \tfrac{\gamma}{2}\langle \bar{x}-x_f, 2(x-\bar{x})+\bar{x}-x_f\rangle\\
&= f^*(\bar{x};\gamma) + \langle g_f, x-\bar{x}\rangle + \tfrac{1}{2\gamma} \parallel g_f \parallel^2 . \qquad \square
\end{aligned}$$

In what follows, we often use the following corollary to Theorem 2.3.3.

**Corollary 2.3.2** *Let $f\in\mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$ and $\gamma\geq L$. Then:*

1. *For any $x\in Q$ and $\bar{x}\in\mathbb{R}^n$, we have*

$$f(x) \geq f(x_f(\bar{x};\gamma)) + \langle g_f(\bar{x};\gamma), x-\bar{x}\rangle + \tfrac{1}{2\gamma} \parallel g_f(\bar{x};\gamma) \parallel^2 + \tfrac{\mu}{2} \parallel x-\bar{x} \parallel^2 . \tag{2.3.7}$$

2. *If $\bar{x} \in Q$, then*

$$f(x_f(\bar{x};\gamma)) \leq f(\bar{x}) - \tfrac{1}{2\gamma} \parallel g_f(\bar{x};\gamma) \parallel^2 . \tag{2.3.8}$$

3. *For any $\bar{x} \in \mathbb{R}^n$, we have*

$$\langle g_f(\bar{x};\gamma), \bar{x} - x^* \rangle \geq \tfrac{1}{2\gamma} \parallel g_f(\bar{x};\gamma) \parallel^2 + \tfrac{\mu}{2} \parallel x^* - \bar{x} \parallel^2 . \tag{2.3.9}$$

*Proof* Assumption $\gamma \geq L$ implies that $f^*(\bar{x};\gamma) \geq f(x_f(\bar{x};\gamma))$. Therefore, (2.3.7) follows from (2.3.6) since

$$f(x) \geq f(\bar{x};x) + \tfrac{\mu}{2} \parallel x - \bar{x} \parallel^2$$

for all $x \in \mathbb{R}^n$ (see Lemma 2.3.1).

Using (2.3.7) with $x = \bar{x}$, we get (2.3.8), and using (2.3.7) with $x = x^*$, we get (2.3.9) since $f(x_f(\bar{x};\gamma)) - f(x^*) \geq 0$. □

Finally, let us estimate the variation of the optimal value $f^*(\bar{x};\gamma)$ as a function of $\gamma$.

**Lemma 2.3.2** *For any $\gamma_1, \gamma_2 > 0$, and $\bar{x} \in \mathbb{R}^n$, we have*

$$f^*(\bar{x};\gamma_2) \geq f^*(\bar{x};\gamma_1) + \tfrac{\gamma_2-\gamma_1}{2\gamma_1\gamma_2} \parallel g_f(\bar{x};\gamma_1) \parallel^2 .$$

*Proof* Let $x_i = x_f(\bar{x};\gamma_i)$, $g_i = g_f(\bar{x};\gamma_i)$, $i = 1, 2$. In view of (2.3.6), we have

$$\begin{aligned} f(\bar{x};x) + \tfrac{\gamma_2}{2} \parallel x - \bar{x} \parallel^2 \geq {} & f^*(\bar{x};\gamma_1) + \langle g_1, x - \bar{x} \rangle \\ & + \tfrac{1}{2\gamma_1} \parallel g_1 \parallel^2 + \tfrac{\gamma_2}{2} \parallel x - \bar{x} \parallel^2 \end{aligned} \tag{2.3.10}$$

for all $x \in Q$. In particular, for $x = x_2$ we obtain

$$\begin{aligned} f^*(\bar{x};\gamma_2) &= f(\bar{x};x_2) + \tfrac{\gamma_2}{2} \parallel x_2 - \bar{x} \parallel^2 \\ &\geq f^*(\bar{x};\gamma_1) + \langle g_1, x_2 - \bar{x} \rangle + \tfrac{1}{2\gamma_1} \parallel g_1 \parallel^2 + \tfrac{\gamma_2}{2} \parallel x_2 - \bar{x} \parallel^2 \\ &= f^*(\bar{x};\gamma_1) + \tfrac{1}{2\gamma_1} \parallel g_1 \parallel^2 - \tfrac{1}{\gamma_2}\langle g_1, g_2 \rangle + \tfrac{1}{2\gamma_2} \parallel g_2 \parallel^2 \\ &\geq f^*(\bar{x};\gamma_1) + \tfrac{1}{2\gamma_1} \parallel g_1 \parallel^2 - \tfrac{1}{2\gamma_2} \parallel g_1 \parallel^2 . \end{aligned}$$ □

### 2.3.3 Minimization Methods for the Minimax Problem

As usual, we start the presentation of numerical methods for problem (2.3.1) with a variant of the Gradient Method with constant step.

**Gradient Method for Minimax Problem**

**0.** Choose $x_0 \in Q$ and $h > 0$.

**1.** ***k*th iteration** ($k \geq 0$)**.**

$$x_{k+1} = x_k - hg_f(x_k; L). \tag{2.3.11}$$

**Theorem 2.3.4** *Let $f \in \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$. If in method (2.3.11) we choose $h \leq \frac{1}{L}$, then it forms a feasible sequence of points such that*

$$\| x_k - x^* \|^2 \leq (1 - \mu h)^k \| x_0 - x^* \|^2, \quad k \geq 0.$$

*Proof* Let $r_k = \| x_k - x^* \|$ and $g_k = g_f(x_k; L)$. Then, in view of (2.3.9), we have

$$r_{k+1}^2 = \| x_k - x^* - hg_k \|^2 = r_k^2 - 2h\langle g_k, x_k - x^* \rangle + h^2 \| g_k \|^2$$

$$\leq (1 - h\mu)r_k^2 + h\left(h - \tfrac{1}{L}\right) \| g_k \|^2 \leq (1 - h\mu)r_k^2.$$

Let $\alpha = hL \leq 1$. Then $x_{k+1} = (1-\alpha)x_k + \alpha x_f(x_k, L) \in Q$. □

With the maximal step size $h = \frac{1}{L}$, we have

$$x_{k+1} = x_k - \tfrac{1}{L} g_f(x_k; L) = x_f(x_k; L).$$

For this step size, the rate of convergence of method (2.3.11) is as follows:

$$\| x_k - x^* \|^2 \leq \left(1 - \tfrac{\mu}{L}\right)^k \| x_0 - x^* \|^2 .$$

As compared with Theorem 2.2.14, the Gradient Method for the minimax problem has a rate of convergence with a similar dependence on the condition number.

Let us check what we can say about the optimal methods. In order to develop an optimal scheme, we need to introduce estimating sequences with some recursive updating rules. Formally, the minimax problem differs from the unconstrained minimization problem only by the analytical form of the lower approximation of the objective function. In the case of unconstrained minimization, we use inequality (2.1.20) for updating the estimating sequence. Now we just replace it by the lower bound (2.3.7).

Let us introduce the estimating sequences for problem (2.3.1). We fix some point $x_0 \in Q$ and coefficient $\gamma_0 > 0$. Consider the sequences $\{y_k\} \subset \mathbb{R}^n$ and $\{\alpha_k\} \subset (0, 1)$. Define

$$\phi_0(x) = f(x_0) + \tfrac{\gamma_0}{2} \parallel x - x_0 \parallel^2,$$

$$\phi_{k+1}(x) = (1-\alpha_k)\phi_k(x) + \alpha_k[\boxed{f(x_f(y_k; L)) + \tfrac{1}{2L} \parallel g_f(y_k; L) \parallel^2}$$

$$+\langle g_f(y_k; L), x - y_k\rangle + \tfrac{\mu}{2} \parallel x - y_k \parallel^2].$$

Comparing these relations with (2.2.4), we can see the difference only in the constant term (shown in the frame). In (2.2.4), we used $f(y_k)$ in this position. This difference leads to a trivial modification of the results of Lemma 2.2.3: All appearances of $f(y_k)$ must be formally replaced by the expression in the frame, and $\nabla f(y_k)$ must be replaced by the reduced gradient $g_f(y_k; L)$. Thus, we come to the following lemma.

**Lemma 2.3.3** *For all $k \geq 0$ we have*

$$\phi_k(x) \equiv \phi_k^* + \tfrac{\gamma_k}{2} \parallel x - v_k \parallel^2,$$

*where the sequences $\{\gamma_k\}$, $\{v_k\}$ and $\{\phi_k^*\}$ are defined as $v_0 = x_0$, $\phi_0^* = f(x_0)$, and*

$$\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu,$$

$$v_{k+1} = \tfrac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k v_k + \alpha_k\mu y_k - \alpha_k g_f(y_k; L)],$$

$$\phi_{k+1}^* = (1-\alpha_k)\phi_k + \alpha_k(f(x_f(y_k; L)) + \tfrac{1}{2L} \parallel g_f(y_k; L) \parallel^2) + \tfrac{\alpha_k^2}{2\gamma_{k+1}} \parallel g_f(y_k; L) \parallel^2$$

$$+\tfrac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\tfrac{\mu}{2} \parallel y_k - v_k \parallel^2 + \langle g_f(y_k; L), v_k - y_k\rangle\right).$$

□

Now we can proceed exactly as in Sect. 2.2. Assume that $\phi_k^* \geq f(x_k)$. Inequality (2.3.7) with $x = x_k$ and $\bar{x} = y_k$ becomes as follows:

$$f(x_k) \geq f(x_f(y_k; L)) + \langle g_f(y_k; L), x_k - y_k\rangle + \tfrac{1}{2L} \parallel g_f(y_k; L) \parallel^2$$

$$+\tfrac{\mu}{2} \parallel x_k - y_k \parallel^2 .$$

Hence,

$$\begin{aligned}\phi^*_{k+1} &\geq (1-\alpha_k)f(x_k)+\alpha_k f(x_f(y_k;L))+\left(\tfrac{\alpha_k}{2L}-\tfrac{\alpha_k^2}{2\gamma_{k+1}}\right)\parallel g_f(y_k;L)\parallel^2\\ &\quad +\tfrac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\langle g_f(y_k;L), v_k-y_k\rangle\\ &\geq f(x_f(y_k;L))+\left(\tfrac{1}{2L}-\tfrac{\alpha_k^2}{2\gamma_{k+1}}\right)\parallel g_f(y_k;L)\parallel^2\\ &\quad +(1-\alpha_k)\langle g_f(y_k;L), \tfrac{\alpha_k\gamma_k}{\gamma_{k+1}}(v_k-y_k)+x_k-y_k\rangle.\end{aligned}$$

Thus, again we can choose

$$\begin{aligned}x_{k+1} &= x_f(y_k;L),\\ L\alpha_k^2 &= (1-\alpha_k)\gamma_k+\alpha_k\mu\equiv\gamma_{k+1},\\ y_k &= \tfrac{1}{\gamma_k+\alpha_k\mu}(\alpha_k\gamma_k v_k+\gamma_{k+1}x_k).\end{aligned}$$

Let us write down the resulting scheme in the form of (2.2.20), with eliminated sequences $\{v_k\}$ and $\{\gamma_k\}$.

**Constant Step Scheme II for Minimax Problem**

**0.** Choose $x_0\in\mathbb{R}^n$ and $\alpha_0\in\left[\sqrt{q_f}, \frac{2(3+q_f)}{3+\sqrt{21+4q_f}}\right]$. Set $y_0=x_0$.

**1.** ***k*th iteration ($k\geq 0$).**

(a) Compute $\{f_i(y_k)\}_{i=1}^m$ and $\{\nabla f_i(y_k)\}_{i=1}^m$. Set $x_{k+1}=x_f(y_k;L)$.

(b) Compute $\alpha_{k+1}\in(0,1)$ from the equation

$$\alpha_{k+1}^2=(1-\alpha_{k+1})\alpha_k^2+q_f\alpha_{k+1}.$$

Set $\beta_k=\frac{\alpha_k(1-\alpha_k)}{\alpha_k^2+\alpha_{k+1}}$ and $y_{k+1}=x_{k+1}+\beta_k(x_{k+1}-x_k)$.

(2.3.12)

The convergence analysis of this scheme is completely identical to the analysis used for scheme (2.2.20). Let us just give the final result.

**Theorem 2.3.5** *Let the max-type function $f$ belong to $\mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$. If in the method (2.3.12) we take $\alpha_0 \in \left[\sqrt{q_f}, \frac{2(3+q_f)}{3+\sqrt{21+4q_f}}\right]$, then*

$$f(x_k) - f^* \leq \frac{4\mu\left[f(x_0)-f^*+\frac{\gamma_0}{2}\|x_0-x^*\|^2\right]}{(\gamma_0-\mu)\cdot\left[\exp\left(\frac{k+1}{2}q_f^{1/2}\right)-\exp\left(-\frac{k+1}{2}q_f^{1/2}\right)\right]^2}$$

$$\leq \frac{4L}{(\gamma_0-\mu)(k+1)^2}\left[f(x_0) - f^* + \frac{\gamma_0}{2} \parallel x_0 - x^* \parallel^2\right],$$

*where $\gamma_0 = \frac{\alpha_0(\alpha_0 L-\mu)}{1-\alpha_0}$.* □

Note that the scheme (2.3.12) works for all $\mu \geq 0$. Let us write down the method for solving problem (2.3.1) with strictly convex components.

**Optimal Method for Minimax Problem with $f \in \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$**

**0.** Choose $x_0 \in Q$. Set $y_0 = x_0$, $\beta = \frac{1-\sqrt{q_f}}{1+\sqrt{q_f}}$.

**1. $k$th iteration ($k \geq 0$).**
Compute $\{f_i(y_k)\}$ and $\{\nabla f_i(y_k)\}$. Set $x_{k+1} = x_f(y_k; L)$ and

$$y_{k+1} = x_{k+1} + \beta(x_{k+1} - x_k).$$

(2.3.13)

**Theorem 2.3.6** *For scheme (2.3.13) we have*

$$f(x_k) - f^* \leq 2\left(1 - \sqrt{\frac{\mu}{L}}\right)^k (f(x_0) - f^*). \tag{2.3.14}$$

*Proof* Scheme (2.3.13) is a variant of (2.3.12) with $\alpha_0 = \sqrt{\frac{\mu}{L}}$. Under this choice, $\gamma_0 = \mu$ and we get (2.3.14) from Theorem 2.3.5 since, in view of Corollary 2.3.1, $\frac{\mu}{2} \parallel x_0 - x^* \parallel^2 \leq f(x_0) - f^*$. □

To conclude this section, let us look at the auxiliary problem, which we need to solve for computing the Gradient Mapping of the minimax problem. Recall that this problem is as follows:

$$\min_{x\in Q}\left\{\max_{1\leq i\leq m}[f_i(x_0) + \langle\nabla f_i(x_0), x - x_0\rangle] + \frac{\gamma}{2} \parallel x - x_0 \parallel^2\right\}.$$

Introducing an additional variable $t \in \mathbb{R}$, we can rewrite this problem in the following form:

$$\begin{aligned} &\min_{x,t} \left\{ t + \tfrac{\gamma}{2} \parallel x - x_0 \parallel^2 \right\} \\ \text{s. t. } &f_i(x_0) + \langle \nabla f_i(x_0), x - x_0 \rangle \leq t, \ i = 1 \dots m, \\ &x \in Q, \ t \in \mathbb{R}, \end{aligned} \tag{2.3.15}$$

If $Q$ is a polytope, then the problem (2.3.15) is a quadratic optimization problem. Such a problem can be solved by some special finite methods (simplex-type algorithms). It can also be solved by Interior Point Methods (see Chap. 5). In the latter case, we can treat much more complicated structures of the basic feasible set $Q$.

## 2.3.4 *Optimization with Functional Constraints*

Let us show that the methods of the previous section can be used to solve a constrained minimization problem with smooth functional constraints. Recall, that the analytical form of such a problem is as follows:

$$\begin{aligned} &\min_{x \in Q} \ f_0(x), \\ \text{s.t. } &f_i(x) \leq 0, \ i = 1 \dots m, \end{aligned} \tag{2.3.16}$$

where the functions $f_i$ are convex and smooth and $Q$ is a simple closed convex set. In this section, we assume that $f_i \in \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$, $i = 0 \dots m$, with some $\mu > 0$.

The relation between problem (2.3.16) and minimax problems is established by some special function of one variable. Consider the *parametric* max-type function

$$f(t; x) = \max\{f_0(x) - t; \ f_i(x), i = 1 \dots m\}, \quad t \in \mathbb{R}, \ x \in Q.$$

Let us introduce the auxiliary function

$$f^*(t) = \min_{x \in Q} \ f(t; x). \tag{2.3.17}$$

Note that the components of the max-type function $f(t; \cdot)$ are strongly convex in $x$. Therefore, for any $t \in \mathbb{R}$, the solution of problem (2.3.17), $x^*(t)$, exists and is unique in view of Theorem 2.3.2.

We will try to approach the solution of problem (2.3.16) by a process based on *approximate values* of the function $f^*(t)$. This approach can be seen as a variant of *Sequential Quadratic Optimization*. It can also be applied to nonconvex problems.

Let us establish some properties of function $f^*(\cdot)$. Clearly, this is a continuous function.

**Lemma 2.3.4** *Let $t^*$ be the optimal value of problem (2.3.16). Then*

$$f^*(t) \leq 0 \textit{ for all } t \geq t^*,$$

$$f^*(t) > 0 \textit{ for all } t < t^*.$$

*Proof* Let $x^*$ be the solution to problem (2.3.16). If $t \geq t^*$, then

$$f^*(t) \leq f(t; x^*) = \max\{f_0(x^*) - t;\, f_i(x^*)\} \;\leq\; \max\{t^* - t;\, f_i(x^*)\} \;\leq\; 0.$$

Suppose that $t < t^*$ and $f^*(t) \leq 0$. Then there exists a $y \in Q$ such that

$$f_0(y) \leq t < t^*, \quad f_i(y) \leq 0,\; i = 1 \ldots m.$$

Hence, $t^*$ cannot be the optimal value of problem (2.3.16). □

Thus, the smallest root of the function $f^*(\cdot)$ corresponds to the optimal value of problem (2.3.16). Note also that, using the methods of the previous section, we can only compute an approximation to the value $f^*(t)$. Hence, our goal now is to form a process for finding this root, based on this inexact information. To do so, we need to establish some properties of the function $f^*(\cdot)$.

**Lemma 2.3.5** *For any $\Delta \geq 0$, we have*

$$f^*(t) - \Delta \leq f^*(t + \Delta) \leq f^*(t).$$

*Proof* Indeed,

$$\begin{aligned} f^*(t+\Delta) &= \min_{x\in Q} \max_{1\leq i\leq m} \{f_0(x) - t - \Delta;\, f_i(x)\} \\ &\leq \min_{x\in Q} \max_{1\leq i\leq m} \{f_0(x) - t;\, f_i(x)\} \;=\; f^*(t), \\ f^*(t+\Delta) &= \min_{x\in Q} \max_{1\leq i\leq m} \{f_0(x) - t;\, f_i(x) + \Delta\} - \Delta \\ &\geq \min_{x\in Q} \max_{1\leq i\leq m} \{f_0(x) - t;\, f_i(x)\} - \Delta \;=\; f^*(t) - \Delta. \quad □ \end{aligned}$$

In other words the function $f^*(\cdot)$ is decreasing and Lipschitz continuous with constant one.

**Lemma 2.3.6** *For any $t_1 < t_2$ and $\Delta \geq 0$, we have*

$$f^*(t_1 - \Delta) \geq f^*(t_1) + \Delta \tfrac{f^*(t_1) - f^*(t_2)}{t_2 - t_1}. \qquad (2.3.18)$$

*Proof* Let $t_0 = t_1 - \Delta$, $\alpha = \frac{\Delta}{t_2 - t_0} \equiv \frac{\Delta}{t_2 - t_1 + \Delta} \in [0, 1]$. Then $t_1 = (1-\alpha)t_0 + \alpha t_2$, and inequality (2.3.18) can be written as follows:

$$f^*(t_1) \leq (1-\alpha) f^*(t_0) + \alpha f^*(t_2). \qquad (2.3.19)$$

Let $x_\alpha = (1-\alpha)x^*(t_0) + \alpha x^*(t_2)$. Then

$$\begin{aligned}
f^*(t_1) &\leq \max_{1 \leq i \leq m} \{ f_0(x_\alpha) - t_1;\ f_i(x_\alpha) \} \\
&\overset{(2.1.3)}{\leq} \max_{1 \leq i \leq m} \{ (1-\alpha)(f_0(x^*(t_0)) - t_0) + \alpha (f_0(x^*(t_2)) - t_2); \\
&\qquad (1-\alpha) f_i(x^*(t_0)) + \alpha f_i(x^*(t_2)) \} \\
&\leq (1-\alpha) \max_{1 \leq i \leq m} \{ f_0(x^*(t_0)) - t_0;\ f_i(x^*(t_0)) \} \\
&\qquad + \alpha \max_{1 \leq i \leq m} \{ f_0(x^*(t_2)) - t_2;\ f_i(x^*(t_2)) \} \\
&= (1-\alpha) f^*(t_0) + \alpha f^*(t_2),
\end{aligned}$$

and we get (2.3.18). □

Note that Lemmas 2.3.5 and 2.3.6 are valid for *any* parametric max-type functions, not necessarily formed by the functional components of problem (2.3.16).

Let us now study the properties of Gradient Mapping for the parametric max-type function. Define a *linearization* of parametric max-type function $f(t; \cdot)$:

$$f(t; \bar{x}; x) = \max_{1 \leq i \leq m} \{ f_0(\bar{x}) + \langle \nabla f_0(\bar{x}), x - \bar{x} \rangle - t;\ f_i(\bar{x}) + \langle \nabla f_i(\bar{x}), x - \bar{x} \rangle \}.$$

Now we can introduce a Gradient Mapping in the usual way. Let us fix some $\gamma > 0$. Define

$$f_\gamma(t;\bar{x};x) = f(t;\bar{x};x) + \tfrac{\gamma}{2} \parallel x - \bar{x} \parallel^2,$$

$$f^*(t;\bar{x};\gamma) = \min_{x \in Q} \; f_\gamma(t;\bar{x};x),$$

$$x_f(t;\bar{x};\gamma) = \arg\min_{x \in Q} \; f_\gamma(t;\bar{x};x),$$

$$g_f(t;\bar{x};\gamma) = \gamma(\bar{x} - x_f(t;\bar{x};\gamma)).$$

We call $x_f(t;\bar{x};\gamma)$ the *Constrained Gradient Mapping*, and $g_f(t;\bar{x},\gamma)$ the *Constrained Reduced Gradient* of problem (2.3.16). As usual, the point of linearization $\bar{x}$ is not necessarily feasible for $Q$.

Note that the function $f_\gamma(t;\bar{x};\cdot)$ itself is a max-type function composed of the components

$$f_0(\bar{x}) + \langle \nabla f_0(\bar{x}), x - \bar{x}\rangle - t + \tfrac{\gamma}{2} \parallel x - \bar{x} \parallel^2,$$

$$f_i(\bar{x}) + \langle \nabla f_i(\bar{x}), x - \bar{x}\rangle + \tfrac{\gamma}{2} \parallel x - \bar{x} \parallel^2, \; i = 1 \ldots m.$$

Moreover, $f_\gamma(t;\bar{x};\cdot) \in \mathscr{S}^{1,1}_{\gamma,\gamma}(\mathbb{R}^n)$. Therefore, in view of Theorem 2.3.2, the Constrained Gradient Mapping is well defined for any $t \in \mathbb{R}$.

Since $f(t;\cdot) \in \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$, we have

$$f_\mu(t;\bar{x};x) \overset{(2.3.2)}{\leq} f(t;x) \overset{(2.3.3)}{\leq} f_L(t;\bar{x};x)$$

for all $x \in \mathbb{R}^n$. Therefore

$$f^*(t;\bar{x};\mu) \leq f^*(t) \;\leq\; f^*(t;\bar{x};L).$$

Moreover, using Lemma 2.3.6, we obtain the following result.

*For any $\bar{x} \in \mathbb{R}^n$, $\gamma > 0$, $\Delta \geq 0$ and $t_1 < t_2$ and we have*

$$f^*(t_1 - \Delta;\bar{x};\gamma) \geq f^*(t_1;\bar{x};\gamma) + \tfrac{\Delta}{t_2 - t_1}(f^*(t_1;\bar{x};\gamma) - f^*(t_2;\bar{x};\gamma)). \tag{2.3.20}$$

There are two values, $\gamma = L$ and $\gamma = \mu$, which are important for us. Applying Lemma 2.3.2 to the max-type function $f_\gamma(t;\bar{x};x)$ with $\gamma_1 = L$ and $\gamma_2 = \mu$, we get the following inequality:

$$f^*(t;\bar{x};\mu) \geq f^*(t;\bar{x};L) - \tfrac{L-\mu}{2\mu L} \parallel g_f(t;\bar{x};L) \parallel^2 . \tag{2.3.21}$$

Since we are interested in finding a root of the function $f^*(\cdot)$, let us look first at the roots of the function $f^*(\cdot; \bar{x}; \gamma)$, which can be seen as an approximation to $f^*(\cdot)$.

Define

$$t^*(\bar{x}, t) = \mathrm{root}_t(f^*(t; \bar{x}; \mu))$$

(the notation $\mathrm{root}_t(\cdot)$ corresponds to the root in $t$ of the function $(\cdot)$).

**Lemma 2.3.7** *Let $\bar{x} \in \mathbb{R}^n$ and $\bar{t} < t^*$ be such that*

$$f^*(\bar{t}; \bar{x}; \mu) \geq (1 - \varkappa) f^*(\bar{t}; \bar{x}; L)$$

*for some $\varkappa \in (0, 1)$. Then $\bar{t} < t^*(\bar{x}, \bar{t}) \leq t^*$. Moreover, for any $t < \bar{t}$ and $x \in \mathbb{R}^n$ we have*

$$f^*(t; x; L) \geq 2(1 - \varkappa) f^*(\bar{t}; \bar{x}; L) \sqrt{\tfrac{\bar{t} - t}{t^*(\bar{x}, \bar{t}) - \bar{t}}}.$$

*Proof* Since $\bar{t} < t^*$, we have

$$0 < f^*(\bar{t}) \leq f^*(\bar{t}; \bar{x}; L) \leq \tfrac{1}{1-\varkappa} f^*(\bar{t}; \bar{x}; \mu).$$

Thus, $f^*(\bar{t}; \bar{x}; \mu) > 0$ and, since $f^*(\cdot; \bar{x}; \mu)$ is decreasing, we get

$$t^*(\bar{x}, \bar{t}) > \bar{t}.$$

Let $\Delta = \bar{t} - t$. Then, in view of inequality (2.3.20), we have

$$\begin{aligned} f^*(t; x; L) &\geq f^*(t) \geq f^*(\bar{t}; \bar{x}; \mu) \geq f^*(\bar{t}; \bar{x}; \mu) + \tfrac{\Delta}{t^*(\bar{x}, \bar{t}) - \bar{t}} f^*(\bar{t}; \bar{x}; \mu) \\ &\geq (1 - \varkappa) \left(1 + \tfrac{\Delta}{t^*(\bar{x}, \bar{t}) - \bar{t}}\right) f^*(\bar{t}; \bar{x}; L) \\ &\geq 2(1 - \varkappa) f^*(\bar{t}; \bar{x}; L) \sqrt{\tfrac{\Delta}{t^*(\bar{x}, \bar{t}) - \bar{t}}}. \end{aligned}$$

In the last inequality, we use the relation $1 + \tau \geq 2\sqrt{\tau}$, $\tau \geq 0$. □

### 2.3.5 The Method for Constrained Minimization

Now we are ready to analyze the following process.

**Constrained Minimization Scheme**

**0.** Choose $x_0 \in Q$, $\varkappa \in (0, \frac{1}{2})$, $t_0 < t^*$, and accuracy $\epsilon > 0$.

**1.** ***k*th iteration ($k \geq 0$).**

(a) Generate the sequence $\{x_{k,j}\}$ by method (2.3.13) as applied to $f(t_k; \cdot)$ with starting point $x_{k,0} = x_k$. If

$$f^*(t_k; x_{k,j}; \mu) \;\geq\; (1 - \varkappa) f^*(t_k; x_{k,j}; L),$$

then stop the internal process and set $j(k) = j$,

$$j^*(k) = \arg \min_{0 \leq j \leq j(k)} f^*(t_k; x_{k,j}; L),$$

$$x_{k+1} = x_f(t_k; x_{k,j^*(k)}; L).$$

**Global Stop:** $f^*(t_k; x_{k,j}; L) \leq \epsilon$ at some iteration of the internal scheme.

(b) Set $t_{k+1} = t^*(x_{k,j(k)}, t_k)$.

(2.3.22)

This is the first time in this book we have met a two-level process. Clearly, its analysis is more complicated. Firstly, we need to estimate the rate of convergence of the upper-level process in (2.3.22) (called the *Master Process*). Secondly, we need to estimate the total complexity of the internal processes in Step 1(a). Since we are interested in the analytical complexity of this method, the arithmetical cost of computation of the root $t^*(x, t)$ and optimal value $f^*(t; x, \gamma)$ is not important for us now.

Let us describe the convergence of the Master Process.

**Lemma 2.3.8**

$$f^*(t_k; x_{k+1}; L) \leq \tfrac{t^* - t_0}{1 - \varkappa} \left[ \tfrac{1}{2(1-\varkappa)} \right]^k .$$

*Proof* Let $\beta = \frac{1}{2(1-\varkappa)}$ $(< 1)$ and

$$\delta_k = \tfrac{f^*(t_k; x_{k,j(k)}; L)}{\sqrt{t_{k+1} - t_k}}.$$

Since $t_{k+1} = t^*(x_{k,j(k)}, t_k)$, in view of Lemma 2.3.7, for $k \geq 1$ we have

$$2(1 - \varkappa) \tfrac{f^*(t_k; x_{k,j(k)}; L)}{\sqrt{t_{k+1} - t_k}} \leq \tfrac{f^*(t_{k-1}; x_{k-1,j(k-1)}; L))}{\sqrt{t_k - t_{k-1}}}.$$

Thus, $\delta_k \le \beta\delta_{k-1}$ and we obtain

$$\begin{aligned} f^*(t_k; x_{k,j(k)}; L) = \delta_k\sqrt{t_{k+1}-t_k} \ &\le \ \beta^k\delta_0\sqrt{t_{k+1}-t_k} \\ &= \beta^k f^*(t_0; x_{0,j(0)}; L)\sqrt{\tfrac{t_{k+1}-t_k}{t_1-t_0}}. \end{aligned}$$

Further, in view of Lemma 2.3.5, we have $t_1 - t_0 \ge f^*(t_0; x_{0,j(0)}; \mu)$. Hence,

$$\begin{aligned} f^*(t_k; x_{k,j(k)}; L) &\le \beta^k f^*(t_0; x_{0,j(0)}; L)\sqrt{\tfrac{t_{k+1}-t_k}{f^*(t_0;x_{0,j(0)};\mu)}} \\ &\le \tfrac{\beta^k}{1-\varkappa}\sqrt{f^*(t_0; x_{0,j(0)}; \mu)(t_{k+1}-t_k)} \\ &\le \tfrac{\beta^k}{1-\varkappa}\sqrt{f^*(t_0)(t^*-t_0)}. \end{aligned}$$

It remains to note that $f^*(t_0) \le t^* - t_0$ (see Lemma 2.3.5), and

$$f^*(t_k; x_{k+1}; L) \equiv f^*(t_k; x_{k,j^*(k)}; L) \le f^*(t_k; x_{k,j(k)}; L). \qquad \square$$

The above result provides us with an estimate for the number of upper-level iterations, which we need for finding an $\epsilon$-solution to problem (2.3.16). Indeed, let $f^*(t_k; x_{k,j}; L) \le \epsilon$. Then for $x_* = x_f(t_k; x_{k,j}; L)$, we have

$$f(t_k; x_*) = \max_{1\le i\le m}\{f_0(x_*) - t_k;\ f_i(x_*)\} \le f^*(t_k; x_{k,j}; L) \le \epsilon.$$

Since $t_k \le t^*$, we conclude that

$$\begin{aligned} f_0(x_*) &\le t^* + \epsilon, \\ f_i(x_*) &\le \epsilon,\ i = 1\dots m. \end{aligned} \tag{2.3.23}$$

In view of Lemma 2.3.8, we can get (2.3.23) at most in

$$N(\epsilon) = \tfrac{1}{\ln[2(1-\varkappa)]}\ln\tfrac{t^*-t_0}{(1-\varkappa)\epsilon} \tag{2.3.24}$$

*full* iterations of the master process (the last iteration of the process, in general, is not full since it is terminated by the Global Stop rule). Note that in estimate (2.3.24), $\varkappa$ is an absolute constant (for example, $\varkappa = \frac{1}{4}$).

Let us analyze the complexity of the internal process. Assume that the sequence $\{x_{k,j}\}$ is generated by (2.3.13) starting from the point $x_{k,0} = x_k$. In view of Theorem 2.3.6, we have

$$
\begin{aligned}
f(t_k; x_{k,j}) - f^*(t_k) &\le 2\left(1 - \sqrt{q_f}\right)^j (f(t_k; x_k) - f^*(t_k)) \\
&\le 2e^{-\sigma \cdot j}(f(t_k; x_k) - f^*(t_k)) \;\le\; 2e^{-\sigma \cdot j} f(t_k; x_k),
\end{aligned}
$$

where $\sigma \overset{\text{def}}{=} \sqrt{q_f}$. Recall that $Q_f = \frac{1}{q_f} = \frac{L}{\mu}$.

Let $N$ be the number of full iterations of process (2.3.22) ($N \le N(\epsilon)$). Thus, $j(k)$ is well defined for all $k$, $0 \le k \le N$. Note that $t_k = t^*(x_{k-1,j(k-1)}, t_{k-1}) > t_{k-1}$. Therefore

$$
f(t_k; x_k) \le f(t_{k-1}; x_k) \le f^*(t_{k-1}; x_{k-1,j^*(k-1)}, L).
$$

Define

$$
\Delta_k = f^*(t_{k-1}; x_{k-1,j^*(k-1)}, L), \quad k \ge 1, \quad \Delta_0 = f(t_0; x_0).
$$

Then, for all $k \ge 0$ we have

$$
f(t_k; x_k) - f^*(t_k) \le \Delta_k.
$$

**Lemma 2.3.9** *For all $k$, $0 \le k \le N$, the internal process no longer works if the following condition is satisfied:*

$$
f(t_k; x_{k,j}) - f^*(t_k) \le \frac{\varkappa}{Q_f - 1} \cdot f^*(t_k; x_{k,j}; L). \tag{2.3.25}
$$

*Proof* Assume that (2.3.25) is satisfied. Then, in view of (2.3.8), we have

$$
\begin{aligned}
\frac{1}{2L} \parallel g_f(t_k; x_{k,j}; L \parallel^2 &\le f(t_k; x_{k,j}) - f(t_k; x_f(t_k; x_{k,j}; L)) \\
&\le f(t_k; x_{k,j}) - f^*(t_k).
\end{aligned}
$$

Therefore, using (2.3.21), we obtain

$$
\begin{aligned}
f^*(t_k; x_{k,j}; \mu) &\ge f^*(t_k; x_{k,j}; L) - \frac{L-\mu}{2\mu L} \parallel g_f(t_k; x_{k,j}; L \parallel^2 \\
&\ge f^*(t_k; x_{k,j}; L) - (Q_f - 1) \cdot (f(t_k; x_{k,j}) - f^*(t_k)) \\
&\overset{(2.3.25)}{\ge} (1 - \varkappa) f^*(t_k; x_{k,j}; L),
\end{aligned}
$$

which is the termination criterion of Step 1(a) in (2.3.22). □

The above result, combined with the estimate of the rate of convergence for the internal process, provide us with the total complexity estimate for the constrained minimization scheme.

**Lemma 2.3.10** *For all $k$, $0 \leq k \leq N$, we have*

$$j(k) \leq 1 + \sqrt{Q_f} \cdot \ln \tfrac{2(Q_f - 1)\Delta_k}{\varkappa \Delta_{k+1}}.$$

*Proof* Assume that

$$j(k) - 1 > \tfrac{1}{\sigma} \ln \tfrac{2(Q_f - 1)\Delta_k}{\varkappa \Delta_{k+1}}, \tag{2.3.26}$$

where $\sigma = \sqrt{q_f}$. Recall that $\Delta_{k+1} = \min\limits_{0 \leq j \leq j(k)} f^*(t_k; x_{k,j}; L)$. Note that the stopping criterion of the internal process was not satisfied for $j = j(k) - 1$. Therefore, in view of Lemma 2.3.9, we have

$$f^*(t_k; x_{k,j}; L) \leq \tfrac{Q_f - 1}{\varkappa}(f(t_k; x_{k,j}) - f^*(t_k)) \leq 2\tfrac{Q_f - 1}{\varkappa} e^{-\sigma \cdot j} \Delta_k \overset{(2.3.26)}{<} \Delta_{k+1}.$$

This is a contradiction with the definition of $\Delta_{k+1}$. □

**Corollary 2.3.3**

$$\sum_{k=0}^{N} j(k) \leq (N+1)\left[1 + \sqrt{Q_f} \cdot \ln \tfrac{2(L-\mu)}{\varkappa \mu}\right] + \sqrt{Q_f} \cdot \ln \tfrac{\Delta_0}{\Delta_{N+1}}. \qquad \square$$

It remains to estimate the number of internal iterations in the last step of the Master Process. Denote this number by $j^*$.

**Lemma 2.3.11**

$$j^* \leq 1 + \sqrt{Q_f} \cdot \ln \tfrac{2(Q_f - 1)\Delta_{N+1}}{\varkappa \epsilon}.$$

*Proof* The proof is very similar to the proof of Lemma 2.3.10. Suppose that

$$j^* - 1 > \sqrt{Q_f} \cdot \ln \tfrac{2(Q_f - 1)\Delta_{N+1}}{\varkappa \epsilon}.$$

Note that for $j = j^* - 1$ we have

$$\begin{aligned} \epsilon \leq f^*(t_{N+1}; x_{N+1,j}; L) &\leq \tfrac{Q_f - 1}{\varkappa}(f(t_{N+1}; x_{N+1,j}) - f^*(t_{N+1})) \\ &\leq 2\tfrac{Q_f - 1}{\varkappa} e^{-\sigma \cdot j} \Delta_{N+1} < \epsilon. \end{aligned}$$

This is a contradiction. □

**Corollary 2.3.4**

$$j^* + \sum_{k=0}^{N} j(k) \leq (N+2)\left[1 + \sqrt{Q_f} \cdot \ln \frac{2(Q_f - 1)}{\varkappa}\right] + \sqrt{Q_f} \cdot \ln \frac{\Delta_0}{\epsilon}.$$

Let us put everything together. Substituting the estimate (2.3.24) for the number of full iterations $N$ into the estimate of Corollary 2.3.4, we come to the following bound for the total number of internal iterations of process (2.3.22):

$$\begin{gathered}\left[\frac{1}{\ln[2(1-\varkappa)]} \ln \frac{t_0 - t^*}{(1-\varkappa)\epsilon} + 2\right] \cdot \left[1 + \sqrt{Q_f} \cdot \ln \frac{2(Q_f - 1)}{\varkappa}\right] \\ + \sqrt{Q_f} \cdot \ln\left(\frac{1}{\epsilon} \cdot \max_{1 \leq i \leq m} \{f_0(x_0) - t_0;\ f_i(x_0)\}\right).\end{gathered} \tag{2.3.27}$$

Note that method (2.3.13), which is used in the internal process, calls the oracle of problem (2.3.16) only once at each iteration. Therefore, the estimate (2.3.27) is an upper bound for the analytical complexity of problem (2.3.16) which $\epsilon$-solution is defined by relations (2.3.23).

Let us check how far this estimate is from the lower bound. The principal term in the estimate (2.3.27) is of the order

$$\ln \frac{t_0 - t^*}{\epsilon} \cdot \sqrt{Q_f} \cdot \ln Q_f.$$

This value differs from the *lower bound* for an unconstrained minimization problem by a factor of $\ln \frac{L}{\mu}$. This means that the scheme (2.3.22) is at least *suboptimal* for constrained optimization problems.

To conclude this section, let us address two technical questions. Firstly, in scheme (2.3.22) it is assumed that we know some estimate $t_0 < t^*$. This assumption is not binding since it is possible to choose $t_0$ as the optimal value of the minimization problem

$$\min_{x \in Q} [f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{\mu}{2} \parallel x - x_0 \parallel^2].$$

Clearly, this value is less than or equal to $t^*$.

Secondly, we assume that we are able to compute $t^*(\bar{x}, t)$. Recall that $t^*(\bar{x}, t)$ is a root of the function

$$f^*(t; \bar{x}; \mu) = \min_{x \in Q} \ f_\mu(t; \bar{x}; x),$$

where $f_\mu(t; \bar{x}; x)$ is a max-type function composed of the components

$$f_0(\bar{x}) + \langle \nabla f_0(\bar{x}), x - \bar{x} \rangle + \tfrac{\mu}{2} \parallel x - \bar{x} \parallel^2 -t,$$

$$f_i(\bar{x}) + \langle \nabla f_i(\bar{x}), x - \bar{x} \rangle + \tfrac{\mu}{2} \parallel x - \bar{x} \parallel^2, \ i = 1 \dots m.$$

In view of Lemma 2.3.4, it is the optimal value of the following minimization problem:

$$\min_{x \in Q} [f_0(\bar{x}) + \langle \nabla f_0(\bar{x}), x - \bar{x} \rangle + \tfrac{\mu}{2} \parallel x - \bar{x} \parallel^2],$$

$$\text{s.t. } f_i(\bar{x}) + \langle \nabla f_i(\bar{x}), x - \bar{x} \rangle + \tfrac{\mu}{2} \parallel x - \bar{x} \parallel^2 \leq 0, \ i = 1 \dots m.$$

This problem is not a pure problem of Quadratic Optimization since the constraints are not linear. However, it still can be solved in finite time by a simplex-type procedure, since the objective function and the constraints have the same Hessian. This problem can also be solved by Interior-Point Methods (see Chap. 5).

# Chapter 3
# Nonsmooth Convex Optimization

In this chapter, we consider the most general convex optimization problems, which are formed by non-differentiable convex functions. We start by studying the main properties of these functions and the definition of subgradients, which are the main directions used in the corresponding optimization schemes. We also prove the necessary facts from Convex Analysis, including different variants of Minimax Theorems. After that, we establish the lower complexity bounds and prove the convergence rate of the Subgradient Method for constrained and unconstrained optimization problems. This method appears to be optimal uniformly in the dimension of the space of variables. In the next section, we consider other optimization methods, which can work in spaces of moderate dimension (the Method of Centers of Gravity, the Ellipsoid Algorithm). The chapter concludes with a presentation of methods based on a complete piece-wise linear model of the objective function (Kelley's method, the Level Method).

## 3.1 General Convex Functions

(Equivalent definitions; Closed functions; The discrete minimax theorem; Continuity of convex functions; Separation theorems; Subgradients; Computation rules; Optimality conditions; the Karush–Kuhn–Tucker Theorem; The exact penalty function; Minimax theorems; Basic elements of primal-dual methods.)

Y. Nesterov, *Lectures on Convex Optimization*, Springer Optimization
and Its Applications 137, https://doi.org/10.1007/978-3-319-91578-4_3

### 3.1.1 Motivation and Definitions

In this chapter, we consider methods for solving the most general *convex* minimization problem

$$\begin{aligned} &\min_{x \in Q} \; f_0(x), \\ &\text{s.t. } f_i(x) \leq 0, \; i = 1 \ldots m, \end{aligned} \tag{3.1.1}$$

where $Q \subseteq \mathbb{R}^n$ is a closed convex set and $f_i(\cdot)$, $i = 0 \ldots m$, are *general convex* functions. The term *general* means that these functions can be nondifferentiable. Clearly, such a problem is more difficult than a problem with differentiable components.

Note that nonsmooth minimization problems arise frequently in different applications. Quite often, some components of a model are composed of max-type functions:

$$f(x) = \max_{1 \leq j \leq p} f_j(x),$$

where $f_j(\cdot)$ are convex and differentiable. In Sect. 2.3, we have seen that such a function can be minimized by methods based on Gradient Mapping. However, if the number of smooth components $p$ is *very big*, the computation of the Gradient Mapping becomes too expensive. Then, it is reasonable to treat this max-type function as a general convex function. Another source of nondifferentiable functions is the situation when some components of the problem (3.1.1) are given *implicitly*, as solutions of some auxiliary problems. Such functions are called the functions with *implicit structure*. Very often, these functions are nondifferentiable.

Let us start our considerations with the definition of a general convex function. In the sequel, the term "general" is often omitted.

Denote by

$$\text{dom}\, f = \{x \in \mathbb{R}^n : |\, f(x) \,| < \infty\}$$

the *domain* of function $f$. We always assume that $\text{dom}\, f \neq \emptyset$.

**Definition 3.1.1** A function $f(\cdot)$ is called *convex* if its domain is convex and for all $x, y \in \text{dom}\, f$ and $\alpha \in [0, 1]$ the following inequality holds:

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y). \tag{3.1.2}$$

If this inequality is strict, the function is called *strictly convex*. We call $f$ *concave* if $-f$ is convex.

At this point, we are not yet ready to speak about any methods for solving problem (3.1.1). In Chap. 2, our optimization schemes were based on *gradients* of smooth functions. For nonsmooth functions, such objects do not exist and we have to find something to replace them. However, in order to do that, we should first study the properties of general convex functions and justify a possible definition of a computable generalized gradient. This route is quite long, but we have to follow it up to the end.

A straightforward consequence of Definition 3.1.1 is the following.

**Lemma 3.1.1 (Jensen's Inequality)** *For any* $x_1, \dots, x_m \in \operatorname{dom} f$ *and positive coefficients* $\alpha_1, \dots, \alpha_m$ *such that*

$$\sum_{i=1}^{m} \alpha_i = 1, \tag{3.1.3}$$

*we have*

$$f\left(\sum_{i=1}^{m} \alpha_i x_i\right) \le \sum_{i=1}^{m} \alpha_i f(x_i). \tag{3.1.4}$$

*Proof* Let us prove this statement by induction over $m$. Definition 3.1.1 justifies inequality (3.1.4) for $m = 2$. Assume it is true for some $m \ge 2$. For a set of $m+1$ points we have

$$\sum_{i=1}^{m+1} \alpha_i x_i = \alpha_1 x_1 + (1-\alpha_1) \sum_{i=1}^{m} \beta_i x_i,$$

where $\beta_i = \frac{\alpha_{i+1}}{1-\alpha_1}$, $i = 1, \dots, m$. Clearly,

$$\sum_{i=1}^{m} \beta_i = 1, \quad \beta_i > 0, \; i = 1 \dots m.$$

Therefore, using Definition 3.1.1 and our inductive assumption, we have

$$\begin{aligned} f\left(\sum_{i=1}^{m+1} \alpha_i x_i\right) &= f\left(\alpha_1 x_1 + (1-\alpha_1) \sum_{i=1}^{m} \beta_i x_i\right) \\ &\le \alpha_1 f(x_1) + (1-\alpha_1) f\left(\sum_{i=1}^{m} \beta_i x_i\right) \le \sum_{i=1}^{m+1} \alpha_i f(x_i). \qquad \square \end{aligned}$$

A point $x = \sum_{i=1}^{m} \alpha_i x_i$ with positive coefficients $\alpha_i$ satisfying the normalizing condition (3.1.3) is called a *convex combination* of points $\{x_i\}_{i=1}^{m}$.

Let us mention two important consequences of Jensen's inequality.

**Corollary 3.1.1** *Let $x$ be a convex combination of points $x_1, \dots, x_m$. Then*

$$f(x) \le \max_{1 \le i \le m} f(x_i).$$

*Proof* Indeed, by Jensen's inequality and condition (3.1.3), we have

$$f(x) = f\left(\sum_{i=1}^{m} \alpha_i x_i\right) \le \sum_{i=1}^{m} \alpha_i f(x_i) \le \max_{1 \le i \le m} f(x_i). \qquad \square$$

**Corollary 3.1.2** *Let*

$$\Delta = \mathrm{Conv}\,\{x_1, \dots, x_m\} \equiv \left\{ x = \sum_{i=1}^{m} \alpha_i x_i \mid \alpha_i \ge 0, \ \sum_{i=1}^{m} \alpha_i = 1 \right\}.$$

*Then* $\max\limits_{x \in \Delta} f(x) = \max\limits_{1 \le i \le n} f(x_i)$. □

There exist two other equivalent definitions of convex functions.

**Theorem 3.1.1** *A function $f$ is convex if and only if for all $x, y \in \mathrm{dom}\, f$ and $\beta \ge 0$ such that $y + \beta(y - x) \in \mathrm{dom}\, f$, we have*

$$f(y + \beta(y - x)) \ge f(y) + \beta(f(y) - f(x)). \tag{3.1.5}$$

*Proof* Let $f$ be convex. Define $\alpha = \frac{\beta}{1+\beta}$ and $u = y + \beta(y - x)$. Then

$$y = \tfrac{1}{1+\beta}(u + \beta x) = (1 - \alpha)u + \alpha x.$$

Therefore,

$$f(y) \le (1 - \alpha) f(u) + \alpha f(x) = \tfrac{1}{1+\beta} f(u) + \tfrac{\beta}{1+\beta} f(x).$$

Assume now that (3.1.5) holds. Let us fix $x, y \in \mathrm{dom}\, f$ and $\alpha \in (0, 1]$. Define $\beta = \frac{1-\alpha}{\alpha}$ and $u = \alpha x + (1 - \alpha) y$. Then

$$x = \tfrac{1}{\alpha}(u - (1 - \alpha) y) = u + \beta(u - y).$$

Therefore, $f(x) \ge f(u) + \beta(f(u) - f(y)) = \frac{1}{\alpha} f(u) - \frac{1-\alpha}{\alpha} f(y)$. □

**Theorem 3.1.2** *A function $f$ is convex if and only if its* epigraph

$$\mathrm{epi}\,(f) = \{(x, t) \in \mathrm{dom}\, f \times \mathbb{R} \mid t \ge f(x)\}$$

*is a convex set.*

*Proof* Indeed, if $(x_1, t_1) \in \text{epi}\,(f)$ and $(x_2, t_2) \in \text{epi}\,(f)$, then for any $\alpha \in [0, 1]$ we have

$$\alpha t_1 + (1-\alpha)t_2 \geq \alpha f(x_1) + (1-\alpha) f(x_2) \geq f(\alpha x_1 + (1-\alpha)x_2).$$

Thus, $(\alpha x_1 + (1-\alpha)x_2, \alpha t_1 + (1-\alpha)t_2) \in \text{epi}\,(f)$.

Let $\text{epi}\,(f)$ be convex. Note that for $x_1, x_2 \in \text{dom}\, f$, the corresponding points of the graph of the function belong to the epigraph:

$$(x_1, f(x_1)) \in \text{epi}\,(f), \quad (x_1, f(x_2)) \in \text{epi}\,(f).$$

Therefore $(\alpha x_1 + (1-\alpha)x_2, \alpha f(x_1) + (1-\alpha) f(x_2)) \in \text{epi}\,(f)$. This means that

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2). \qquad \square$$

We also need the following property of the level sets of convex functions.

**Theorem 3.1.3** *If a function $f$ is convex, then all level sets*

$$\mathscr{L}_f(\beta) = \{x \in \text{dom}\, f \mid f(x) \leq \beta\}, \quad \beta \in \mathbb{R},$$

*are either convex or empty.*

*Proof* Indeed, if $x_1 \in \mathscr{L}_f(\beta)$ and $x_2 \in \mathscr{L}_f(\beta)$, then for any $\alpha \in [0, 1]$ we have

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2) \leq \alpha\beta + (1-\alpha)\beta = \beta. \qquad \square$$

In Example 3.1.1(6) we will see that behavior of a general convex function on the boundary of its domain is sometimes out of any control. Therefore, we need to introduce one convenient notion, which will be very useful in our analysis.

**Definition 3.1.2** A function $f$ is called *closed* and convex on a convex set $Q \subseteq \text{dom}\, f$ if its *constrained epigraph*

$$\text{epi}_Q(f) = \{(x, t) \in Q \times \mathbb{R} : \ t \geq f(x)\}$$

is a closed convex set. If $Q = \text{dom}\, f$, we call $f$ a *closed convex function*.

Note that in this definition the set $Q$ is not necessarily closed. Let us prove the following natural statement.

**Lemma 3.1.2** *Let a function $f$ be closed and convex on $Q$. Then for any closed convex set $Q_1 \subseteq Q$, this function is closed and convex on $Q_1$.*

*Proof* Indeed, the set $\{(x, t) : \ x \in Q_1\}$ is closed. Hence, the statement follows from Item 1 of Theorem 2.2.8. $\square$

Let us mention the most important topological properties of closed convex functions.

**Theorem 3.1.4** *Let a function $f$ be closed and convex.*

1. *For any sequence $\{x_k\} \subset \operatorname{dom} f$ convergent to a point $\bar{x} \in \operatorname{dom} f$ we have*

$$\liminf_{k\to\infty} f(x_k) \geq f(\bar{x}). \tag{3.1.6}$$

   *(This means that $f$ is* lower semi-continuous.*)*
2. *For any sequence $\{x_k\} \subset \operatorname{dom} f$ convergent to some point $\bar{x} \notin \operatorname{dom} f$ we have*

$$\lim_{k\to\infty} f(x_k) = +\infty. \tag{3.1.7}$$

3. *All level sets of the function $f$ are either empty or closed and convex.*
4. *Let $f$ be closed and convex on a set $Q$ and its constrained level sets be bounded. Then problem*

$$\min_{x\in Q} f(x)$$

   *is solvable.*
5. *Let $f$ be closed and convex on $Q$. If the optimal set $X^* = \operatorname{Arg}\min_{x\in Q} f(x)$ is nonempty and bounded, then all level sets of the function $f$ on $Q$ are either empty or bounded.*

*Proof*

1. Note that the sequence $\{(x_k, f(x_k))\}$ belongs to the closed set $\operatorname{epi}(f)$. If it has a subsequence convergent to $(\bar{x}, \bar{f}) \in \operatorname{epi}(f)$, then $\bar{x} \in \operatorname{dom} f$ and $\bar{f} \geq f(\bar{x})$. This is the inequality (3.1.6).

   If there is no convergent subsequence in $\{f(x_k)\}$, we need to consider two cases. Assume that $\liminf\limits_{k\to\infty} f(x_k) = -\infty$. Since $\bar{x} \in \operatorname{dom} f$, the sequence $\{(x_k, f(\bar{x}) - 1)\}$ belongs to $\operatorname{epi}(f)$ for $k$ large enough, but it converges to the point $(\bar{x}, f(\bar{x}) - 1) \notin \operatorname{epi}(f)$. This contradicts our assumption. Thus, the only possibility is $\lim\limits_{k\to\infty} f(x_k) = +\infty$. Hence, (3.1.6) is also satisfied.
2. Let $\bar{x} \notin \operatorname{dom} f$. If the sequence $\{f(x_k)\}$ contains a bounded subsequence, then the corresponding points $(x_k, \tau)$ with $\tau$ big enough belong to the epigraph. However, their limit is not in this set. This contradiction proves (3.1.7).
3. By its definition, $(\mathscr{L}_f(\beta), \beta) = \operatorname{epi}(f) \bigcap \{(x, t) \mid t = \beta\}$. Therefore, the level set $\mathscr{L}_f(\beta)$ is closed and convex as an intersection of two closed convex sets.
4. Consider a sequence $\{x_k\} \subset Q$ such that $\lim\limits_{k\to\infty} f(x_k) = f_* \overset{\text{def}}{=} \inf\limits_{x\in Q} f(x)$. Since the level sets of the function $f$ on $Q$ are bounded, we can assume that it is a convergent sequence: $\lim\limits_{k\to\infty} x_k = x^*$. Assume that $f_* = -\infty$. Consider the points

$y_k = (1-\alpha_k)x_0 + \alpha_k x_k \in Q$, $k \geq 0$, with slowly decreasing coefficients $\alpha_k \downarrow 0$. Note that we can always ensure

$$f(y_k) \overset{(3.1.2)}{\leq} f(x_0) + \alpha_k(f(x_k) - f(x_0)) \;\to\; -\infty,$$

and this contradicts the closedness of the set $\mathrm{epi}_Q(f)$.

Thus, $f_* > -\infty$, and we can assume that the whole sequence $\{(x_k, f(x_k))\}$ converges to a certain point $(x^*, f_*)$ from $\mathrm{epi}_Q(f)$. However, by definition of this set, $x^* \in Q$ and $f(x^*) \leq f_*$.

5. Assume that some set $\mathscr{L}_f(\beta)$ with $\beta > f^* = \min\limits_{x \in Q} f(x)$ is unbounded. Let us fix a point $x^* \in X^*$ and choose $R > \max\limits_{y \in X^*} \|y - x^*\|$. Consider a sequence $\{x_k\} \subset \mathscr{L}_f(\beta)$ with $\rho_k \overset{\mathrm{def}}{=} \|x_k - x^*\| \to \infty$. Without loss of generality, we can assume that all $\rho_k \geq R$. Define $y_k = x^* + \frac{1}{\rho_k}R(x_k - x^*)$. Clearly, $y_k \in Q$ and $\|y_k - x^*\| = R$. However,

$$f(y_k) \overset{(3.1.2)}{\leq} f^* + \tfrac{1}{\rho_k}R(f(x_k) - f^*) \;\to\; f^*, \quad k \to \infty.$$

Since the sequence $\{y_k\}_{k \geq 0}$ is compact and the level set $\mathscr{L}_f(\beta)$ is closed (see Item 3), we can assume that the limit $\lim\limits_{k\to\infty} y_k \overset{\mathrm{def}}{=} \bar{y} \in \mathscr{L}_f(\beta)$ exists. However, by (3.1.6) we have $f(\bar{y}) = f^*$, and this contradicts the choice of $R$. □

Note that, if $f$ is convex and continuous and its domain $\mathrm{dom}\, f$ is closed, then $f$ is a closed function. However, in general, a closed convex function is not necessarily continuous.

Let us look at some examples of closed convex functions.

*Example 3.1.1*

1. A linear function is closed and convex.
2. $f(x) = \mid x \mid$, $x \in \mathbb{R}$, is closed and convex since its epigraph is

$$\{(x,t) \mid \; t \geq x, \; t \geq -x\},$$

which is the intersection of two closed convex sets (see Theorem 3.1.2).

3. All *continuous* and convex functions on $\mathbb{R}^n$ belong to the class of general closed convex functions.
4. The function $f(x) = \frac{1}{x}$, $x > 0$, is convex and closed. However, its domain $\mathrm{dom}\, f = \mathrm{int}\, \mathbb{R}_+$ is open.
5. The function $f(x) = \parallel x \parallel$, where $\parallel \cdot \parallel$ is any *norm*, is closed and convex:

$$f(\alpha x_1 + (1-\alpha)x_2) = \parallel \alpha x_1 + (1-\alpha)x_2 \parallel \leq \parallel \alpha x_1 \parallel + \parallel (1-\alpha)x_2 \parallel$$

$$= \alpha \parallel x_1 \parallel + (1-\alpha) \parallel x_2 \parallel$$

for any $x_1, x_2 \in \mathbb{R}^n$ and $\alpha \in [0,1]$. The most popular norms in Numerical Analysis are so-called *$\ell_p$-norms*:

$$\| x \|_{(p)} = \left[ \sum_{i=1}^{n} | x^{(i)} |^p \right]^{1/p}, \quad p \geq 1.$$

Among them, there are three norms, which are commonly used:

- *Euclidean norm* $\| x \|_{(2)} = [\sum\limits_{i=1}^{n} (x^{(i)})^2]^{1/2}$, $p = 2$. Since it is used very often, usually we drop the subscript if no ambiguity arises.
- *$\ell_1$-norm* $\| x \|_{(1)} = \sum\limits_{i=1}^{n} | x^{(i)} |$, $p = 1$.
- *$\ell_\infty$-norm* (*Chebyshev* norm, *uniform* norm, *infinity* norm)

$$\| x \|_{(\infty)} = \max_{1 \leq i \leq n} | x^{(i)} | .$$

Any norm defines a system of *balls*,

$$B_{\|\cdot\|}(x_0, r) = \{ x \in \mathbb{R}^n \mid \| x - x_0 \| \leq r \}, \quad r \geq 0,$$

where $r$ is the *radius* of the ball and $x_0 \in \mathbb{R}^n$ is its *center*. We call the ball $B_{\|\cdot\|}(0, 1)$ the *unit* ball of the norm $\| \cdot \|$. Clearly, these balls are convex sets (see Theorem 3.1.3). For $\ell_p$-balls of radius $r$ we also use the notation

$$B_p(x_0, r) = \{ x \in \mathbb{R}^n \mid \| x - x_0 \|_{(p)} \leq r \}.$$

For $\ell_1$-balls, we often use the following representation:

$$B_1(x_0, r) = \{ x \in \mathbb{R}^n : \ \|x - x_0\|_{(1)} \leq r \} \ = \ \mathrm{Conv}\, \{ x_0 \pm r e_i, \ i = 1, \ldots, n \}, \tag{3.1.8}$$

where $e_i$ are coordinate vectors in $\mathbb{R}^n$.

6. Up to now, none of our examples have demonstrated any pathological behavior. However, let us look at the following function of two variables:

$$f(x, y) = \begin{cases} \quad 0, & \text{if } x^2 + y^2 < 1, \\ \phi(x, y), & \text{if } x^2 + y^2 = 1, \end{cases}$$

where $\phi(x, y)$ is an *arbitrary* nonnegative function defined on the boundary of a unit circle. The domain of this function is the unit Euclidean disk, which is closed and convex. Moreover, it is easy to see that $f$ is convex. However, it has no reasonable properties at the boundary of its domain. Definitely, we want to exclude such functions from our considerations. This was the main reason for introducing the notion of the closed function. It is clear that $f(\cdot,\cdot)$ is not closed unless $\phi(x,y) \equiv 0$.

Another possibility would be to consider a smaller class of continuous convex functions. However, we will see that for closedness of a convex function there exist very natural sufficient conditions, and this is not the case for continuity. □

### 3.1.2 Operations with Convex Functions

In the previous section, we have seen several examples of convex functions. Let us describe a set of invariant operations which allow us to create more complicated objects.

**Theorem 3.1.5** *Let functions $f_1$ and $f_2$ be closed and convex on convex sets $Q_1$ and $Q_2$, and $\beta \geq 0$. Then all functions below are closed and convex on the corresponding sets $Q$:*

1. $f(x) = \beta f_1(x)$ , $Q = Q_1$.
2. $f(x) = f_1(x) + f_2(x)$, $Q = Q_1 \bigcap Q_2$.[1]
3. $f(x) = \max\{f_1(x), f_2(x)\}$, $Q = Q_1 \bigcap Q_2$.

*Proof*

1. The first item is evident:

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \beta(\alpha f_1(x_1) + (1-\alpha) f_1(x_2)), \quad x_1, x_2 \in Q_1.$$

2. For all $x_1, x_2 \in Q = Q_1 \bigcap Q_2$ and $\alpha \in [0,1]$ we have

$$\begin{aligned} & f_1(\alpha x_1 + (1-\alpha)x_2) + f_2(\alpha x_1 + (1-\alpha)x_2) \\ & \leq \alpha f_1(x_1) + (1-\alpha) f_1(x_2) + \alpha f_2(x_1) + (1-\alpha) f_2(x_2) \\ & = \alpha(f_1(x_1) + f_2(x_1)) + (1-\alpha)(f_1(x_2) + f_2(x_2)). \end{aligned}$$

Thus, $f$ is convex on the set $Q$. Let us prove that it is also closed on $Q$. Consider a convergent sequence $\{(x_k, t_k)\} \subset \text{epi}_Q(f)$:

$$t_k \geq f_1(x_k) + f_2(x_k), \quad x_k \in Q, \quad \lim_{k\to\infty} x_k = \bar{x}, \quad \lim_{k\to\infty} t_k = \bar{t}.$$

[1] Recall that without additional assumptions, we cannot guarantee closedness of the sum of two closed convex sets (see Item 2 in Theorem 2.2.8 and Example 2.2.1). For that, we need boundedness of one of them. However, the epighraphs are never bounded.

Since the functions $f_1$ and $f_2$ are closed on $Q_1$ and $Q_2$ respectively, we have

$$\liminf_{k\to\infty} f_1(x_k) \overset{(3.1.6)}{\geq} f_1(\bar{x}), \quad \bar{x} \in Q_1, \quad \liminf_{k\to\infty} f_2(x_k) \overset{(3.1.6)}{\geq} f_2(\bar{x}), \quad \bar{x} \in Q_2.$$

Therefore, $\bar{x} \in Q_1 \bigcap Q_2$, and

$$\bar{t} = \lim_{k\to\infty} t_k \;\geq\; \liminf_{k\to\infty} f_1(x_k) + \lim_{k\to\infty}\inf f_2(x_k) \;\geq\; f(\bar{x}).$$

Thus, $(\bar{x}, \bar{t}) \in \text{epi}_Q(f)$.

3. The constrained epigraph of the function $f$ can be represented as follows:

$$\text{epi}_Q(f) = \{(x,t) \mid\; t \geq f_1(x),\; t \geq f_2(x),\; x \in Q_1 \bigcap Q_2\}$$

$$\equiv \text{epi}_{Q_1}(f_1) \bigcap \text{epi}_{Q_2}(f_2).$$

Thus, $\text{epi}_Q(f)$ is closed and convex as an intersection of two closed convex sets. □

Let us prove that convexity is an *affine-invariant* property.

**Theorem 3.1.6** *Let a function $\phi$ be closed and convex on a bounded set $S \subseteq \mathbb{R}^m$. Consider a linear operator*

$$\mathscr{A}(x) = Ax + b : \quad \mathbb{R}^n \to \mathbb{R}^m.$$

*Then the function $f(x) = \phi(\mathscr{A}(x))$ is closed and convex on the* inverse image *of the set $S$ defined as follows:*

$$Q = \{x \in \mathbb{R}^n \mid\; \mathscr{A}(x) \in S\}.$$

*Proof* For $x_1$ and $x_2$ in $Q$, define $y_1 = \mathscr{A}(x_1)$, $y_2 = \mathscr{A}(x_2)$. Then for $\alpha \in [0, 1]$ we have

$$f(\alpha x_1 + (1-\alpha)x_2) = \phi(\mathscr{A}(\alpha x_1 + (1-\alpha)x_2)) \;=\; \phi(\alpha y_1 + (1-\alpha)y_2)$$

$$\leq \alpha\phi(y_1) + (1-\alpha)\phi(y_2) \;=\; \alpha f(x_1) + (1-\alpha)f(x_2).$$

Thus, the function $f$ is convex. The closedness of its constrained epigraph follows from the continuity of the linear operator $\mathscr{A}(\cdot)$. □

The next two theorems are the main providers of closed convex functions with implicit structure.

**Theorem 3.1.7** *Let $Q$ be a convex set, and let the function $\phi$ be convex with* $\mathrm{dom}\,\phi \supseteq Q$. *Then the function*

$$f(x) = \inf_y\{\phi(x, y) : \ (x, y) \in Q\} \tag{3.1.9}$$

*is convex on* $\hat{Q} = \{x : \ \exists y \text{such that } (x, y) \in Q\}$.

*Proof* Let us take arbitrary points $x_1, x_2 \in \hat{Q}$. Consider two sequences $\{y_{1,k}\}$ and $\{y_{2,k}\}$ such that $\{(x_1, y_{1,k})\} \subset Q$, $\{(x_2, y_{2,k})\} \subset Q$, and

$$\lim_{k\to\infty} \phi(x_1, y_{1,k}) = f(x_1), \quad \lim_{k\to\infty} \phi(x_2, y_{2,k}) = f(x_2).$$

Since $\phi$ is jointly convex in $(x, y)$, for any $\alpha \in [0, 1]$ we have

$$\begin{aligned} f(\alpha x_1 + (1-\alpha)x_2) &\overset{(3.1.9)}{\leq} \phi(\alpha x_1 + (1-\alpha)x_2, \alpha y_{1,k} + (1-\alpha)y_{2,k}) \\ &\leq \alpha\phi(x_1, y_{1,k}) + (1-\alpha)\phi(x_2, y_{2,k}). \end{aligned}$$

Taking the limit of the right-hand side of this inequality, we get the convexity condition (3.1.2) for the function $f$. □

Conditions for closedness of the function (3.1.9) will be presented later in Theorem 3.1.25 and Theorem 3.1.28.

**Theorem 3.1.8** *Let $\Delta$ be an arbitrary set and*

$$f(x) = \sup_y\{\phi(x, y) \mid \ y \in \Delta\}.$$

*Suppose that for any $y \in \Delta$ functions $\phi(\cdot, y)$ are closed and convex on some set $Q$. Then $f(\cdot)$ is a closed convex function on the set*

$$\hat{Q} = \left\{ x \in Q \mid \ \sup_{y\in\Delta} \phi(x, y) \ < \ +\infty \right\}. \tag{3.1.10}$$

*Proof* Indeed, if $x \in \hat{Q}$, then $f(x) < +\infty$ and we conclude that $Q \subseteq \mathrm{dom}\, f$. Further, it is clear that $(x, t) \in \mathrm{epi}_Q(f)$ if and only if for all $y \in \Delta$ we have

$$x \in Q, \quad t \geq \phi(x, y).$$

This means that

$$\mathrm{epi}_Q(f) = \bigcap_{y\in\Delta} \mathrm{epi}_Q(\phi(\cdot, y)).$$

Thus, $\mathrm{epi}_Q(f)$ is closed and convex since each set $\mathrm{epi}_Q(\phi(\cdot, y))$ is closed and convex. □

**Theorem 3.1.9** *Let a function* $\psi(\cdot)$ *be convex and* $\varphi$ *be a univariate convex function which is non-decreasing on the set*

$$\mathrm{Im}\,\psi = \{\tau = \psi(x),\ x \in \mathrm{dom}\,\psi\}.$$

*Then the function* $f(x) = \varphi(\psi(x))$, $x \in \mathrm{dom}\,\psi$, *is convex.*

*Proof* Indeed, for any points $x$ and $y$ from $\mathrm{dom}\, f$, and $\alpha \in [0, 1]$, we have

$$\begin{aligned} f(\alpha x + (1 - \alpha y)) &= \varphi(\psi(\alpha x + (1 - \alpha)y)) \\ &\leq \varphi(\alpha\psi(x) + (1 - \alpha)\psi(y)) \\ &\leq \alpha\varphi(\psi(x)) + (1 - \alpha)\varphi(\psi(y)) \\ &= \alpha f(x) + (1 - \alpha) f(y). \qquad \Box \end{aligned}$$

Now we are ready to look at more sophisticated examples of convex functions.

*Example 3.1.2*

1. The function $f(x) = \max\limits_{1 \leq i \leq n} \{x^{(i)}\}$ is closed and convex. Another example of a closed convex function is

$$\phi_*(s) = \sup_{x \in \mathrm{dom}\,\phi} [\langle s, x\rangle - \phi(x)],$$

   where $\phi$ is an *arbitrary* function on $\mathbb{R}^n$. The function $\phi_*$ is called the *Fenchel dual* of $\phi$.
2. Let $\lambda = (\lambda^{(1)}, \ldots, \lambda^{(m)})$, and let $\Delta$ be a set in $\mathbb{R}^m_+$. Consider the function

$$f(x) = \sup_{\lambda \in \Delta} \left\{ \sum_{i=1}^{m} \lambda^{(i)} f_i(x) \right\},$$

   where all $f_i$ are closed and convex. In view of Theorem 3.1.5, the epigraphs of the functions

$$\phi_\lambda(x) = \sum_{i=1}^{m} \lambda^{(i)} f_i(x)$$

   are convex and closed. Thus, $f(\cdot)$ is closed and convex in view of Theorem 3.1.8. Note that we have not assumed anything about the structure of the set $\Delta$.

3. Let $Q$ be an arbitrary set. Consider the function

$$\xi_Q(x) = \sup\{\langle g, x\rangle \mid g \in Q\}.$$

The function $\xi_Q(\cdot)$ is called the *support function* of the set $Q$. Note that $\xi_Q(\cdot)$ is closed and convex in view of Theorem 3.1.8. This function is positively homogeneous of degree one:

$$\xi_Q(\tau x) = \tau \xi_Q(x), \quad x \in \operatorname{dom} Q,\ \tau \geq 0.$$

If the set $Q$ is bounded then $\operatorname{dom} \xi_Q = \mathbb{R}^n$.

The support function is a very useful tool in Convex Analysis with many interesting properties. We will present them later in the appropriate places. Here we mention only one of them.

**Lemma 3.1.3** *For two sets* $Q_1$ *and* $Q_2$ *define* $Q = \operatorname{Conv}\{Q_1, Q_2\}$. *Then*

$$\xi_Q(x) = \max\{\xi_{Q_1}(x), \xi_{Q_2}(x)\}, \quad x \in \mathbb{R}^n.$$

*Proof* Indeed, since the sets $Q_1$ and $Q_2$ are subsets of $Q$, for any $x \in \mathbb{R}^n$ we have

$$\xi_Q(x) \geq \max\{\xi_{Q_1}(x), \xi_{Q_2}(x)\}.$$

On the other hand,

$$\begin{aligned} \xi_Q(x) &= \sup_{\alpha, g_1, g_2} \{\langle \alpha g_1 + (1-\alpha) g_2, x\rangle : \ g_1 \in Q_1,\ g_2 \in Q_2,\ \alpha \in [0,1]\} \\ &\leq \sup_{\alpha \in [0,1]} \{\alpha \xi_{Q_1}(x) + (1-\alpha)\xi_{Q_2}(x)\} \ = \ \max\{\xi_{Q_1}(x), \xi_{Q_2}(x)\}. \qquad \square \end{aligned}$$

4. Another important example of a convex homogeneous function related to a convex set is the *Minkowski function*. Let $Q$ be a bounded closed convex set, and $0 \in \operatorname{int} Q$. Then we can define

$$\psi_Q(x) = \min_{\tau \geq 0}\{\tau : \ x \in \tau Q\}.$$

Denote the unique solution of this problem by $\tau(x)$. Then $\frac{x}{\tau(x)} \in \partial Q$. It is easy to see that $\psi_Q$ is a positively homogeneous convex function with $\operatorname{dom} \psi_Q = \mathbb{R}^n$. Indeed, for arbitrary $x_1, x_2 \in \mathbb{R}^n \setminus \{0\}$ and $\alpha \in [0,1]$, we have

$$\frac{\alpha x_1 + (1-\alpha) x_2}{\alpha\tau(x_1) + (1-\alpha)\tau(x_2)} = \frac{\alpha\tau(x_1)\frac{x_1}{\tau(x_1)} + (1-\alpha)\tau(x_2)\frac{x_2}{\tau(x_2)}}{\alpha\tau(x_1) + (1-\alpha)\tau(x_2)} \ \in \ Q.$$

Therefore, $\psi_Q(\alpha x_1 + (1-\alpha)x_2) \leq \alpha\tau(x_1) + (1-\alpha)\tau(x_2)$.

5. Let $Q$ be a set in $\mathbb{R}^n$. Consider the function $\psi(g,\gamma) = \sup\limits_{y\in Q} \phi(y,g,\gamma)$, where

$$\phi(y,g,\gamma) = \langle g, y\rangle - \tfrac{\gamma}{2} \parallel y \parallel^2 .$$

The function $\psi(g,\gamma)$ is closed and convex in $(g,\gamma)$ in view of Theorem 3.1.8. Let us look at its properties.

If $Q$ is bounded, then $\mathrm{dom}\,\psi = \mathbb{R}^{n+1}$. Let us describe the domain of $\psi$ for the case $Q = \mathbb{R}^n$. If $\gamma < 0$, then for any $g \neq 0$ we can set $y_\alpha = \alpha g$. Clearly, along this line, $\phi(y_\alpha, g, \gamma) \to \infty$ as $\alpha \to \infty$. Thus, $\mathrm{dom}\,\psi$ contains only points with $\gamma \geq 0$.

If $\gamma = 0$, the only possible value for $g$ is zero since otherwise the function $\phi(y,g,0)$ is unbounded. Finally, if $\gamma > 0$, then the point maximizing $\phi(y,g,\gamma)$ with respect to $y$ is $y^*(g,\gamma) = \frac{1}{\gamma} g$, and we get the following expression for $\psi$:

$$\psi(g,\gamma) = \tfrac{\|g\|^2}{2\gamma}.$$

Thus,

$$\psi(g,\gamma) = \begin{cases} 0, & \text{if } g = 0, \gamma = 0, \\ \frac{\|g\|^2}{2\gamma}, & \text{if } \gamma > 0, \end{cases}$$

with domain $\mathrm{dom}\,\psi = (\mathbb{R}^n \times \{\gamma > 0\}) \bigcup (0,0)$. This is a convex set which is neither closed nor open. Nevertheless, $\psi$ is a closed convex function. At the same time, this function is discontinuous at the origin:

$$\psi(\sqrt{\gamma} g, \gamma) \equiv \tfrac{1}{2} \parallel g \parallel^2, \quad \gamma \neq 0.$$

Considering the closed convex set $Q = \{(g,\gamma) : \ \gamma \geq \|g\|^2\}$, we can see that $\psi$ is a closed convex function on $Q$ (see Lemma 3.1.2), with bounded values. However, it is still discontinuous at the origin.

6. Similar constructions can be obtained by *homogenization*. Let $f$ be convex on $\mathbb{R}^n$. Consider the function

$$\hat{f}(\tau, x) = \tau f\left(\tfrac{x}{\tau}\right).$$

This function is well defined for all $x \in \mathbb{R}^n$ and $\tau > 0$. Note that $\hat{f}$ is a positively homogeneous function. Therefore, it is natural to define its value at the origin as follows:

$$\hat{f}(0,0) = 0.$$

Let us prove that this function is convex. Consider $z_1 = (\tau_1, x_1)$ and $z_2 = (\tau_2, x_2)$ with $\tau_1, \tau_2 > 0$. Then, for any $\alpha \in [0, 1]$ we have:

$$\begin{aligned}
\hat{f}(\alpha z_1 + (1-\alpha) z_2) &= (\alpha\tau_1 + (1-\alpha)\tau_2) f\left(\tfrac{\alpha x_1 + (1-\alpha)x_2}{\alpha\tau_1 + (1-\alpha)\tau_2}\right) \\
&= (\alpha\tau_1 + (1-\alpha)\tau_2) f\left(\tfrac{\alpha\tau_1 \frac{x_1}{\tau_1} + (1-\alpha)\tau_2 \frac{x_2}{\tau_2}}{\alpha\tau_1 + (1-\alpha)\tau_2}\right) \\
&\leq \alpha\tau_1 f\left(\tfrac{x_1}{\tau_1}\right) + (1-\alpha)\tau_2 f\left(\tfrac{x_2}{\tau_2}\right) \\
&= \alpha \hat{f}(z_1) + (1-\alpha)\hat{f}(z_2).
\end{aligned}$$

However, in general, $\hat{f}(\cdot)$ is not closed. In order to ensure closedness, it is enough to assume that

$$\lim_{\tau\to\infty} \tfrac{1}{\tau} f(\tau x) = +\infty \quad \forall x \in \mathbb{R}^n. \tag{3.1.11}$$

Note that the function $\psi$ in Item 5 can be obtained from $f(x) = \frac{1}{2}\|x\|^2$, which satisfies condition (3.1.11). □

As we have seen in Example 3.1.2(5), a closed convex function can be discontinuous at some points of its domain. However, there exists one very exceptional case when this cannot happen.

**Lemma 3.1.4** *Any univariate closed convex function is continuous on its domain.*

*Proof* Let $f$ be closed and convex and $\bar{x} \in \operatorname{dom} f \subseteq \mathbb{R}$. We have proved in Item 1 of Theorem 3.1.4 that $f$ is lower-semicontinuous at $\bar{x}$. On the other hand, if $x_k = (1-\alpha_k)\bar{x} + \alpha_k \bar{y}$, for certain $\bar{y} \in \operatorname{dom} f$ and $\alpha_k \in [0, 1]$, then

$$f(x_k) \overset{(3.1.2)}{\leq} (1-\alpha_k) f(\bar{x}) + \alpha_k f(\bar{y}).$$

Thus, if $x_k \to \bar{x}$, then $\alpha_k \to 0$ and $\limsup\limits_{k\to\infty} f(x_k) \leq f(\bar{x})$. Hence, $f$ is also upper-semicontinuous at $\bar{x}$. Consequently, it is continuous at $\bar{x}$. □

Thus, it is not surprising that a restriction of the discontinuous function $\psi$ in Item 5 of Example 3.1.2 onto the ray $\{(\gamma g, \gamma), \gamma \geq 0\}$ is a continuous convex function.

As for any other exception, the statement of Lemma 3.1.4 is sometimes very useful.

**Theorem 3.1.10** *Let functions $f_1$ and $f_2$ be closed and convex on $Q$ and their constrained level sets be bounded. Then there exists some $\lambda^* \in [0, 1]$ such that*

$$\min_{x\in Q}\left( f(x) \overset{\text{def}}{=} \max\{f_1(x), f_2(x)\}\right) = \min_{x\in Q}\{\lambda^* f_1(x) + (1-\lambda^*) f_2(x)\}. \tag{3.1.12}$$

*Proof* Define $\phi(\lambda) = \min\limits_{x\in Q}\{\lambda f_1(x) + (1-\lambda) f_2(x)\}$. In view of Theorem 3.1.8, this function is closed and convex, and by Lemma 3.1.4 it is continuous for $\lambda \in [0, 1]$. Thus, its maximal value $\phi^*$ is well defined and

$$\phi^* = \phi(\lambda^*) \;=\; \max_{\lambda\in[0,1]} \phi(\lambda) \;\le\; f^* \;=\; \min_{x\in Q} f(x).$$

Our goal is to show that $\phi^* = f^*$.

For each $\lambda \in [0, 1]$, we fix an arbitrary point

$$x(\lambda) \in \operatorname{Arg}\min_{x\in Q}\{\lambda f_1(x) + (1-\lambda) f_2(x)\}.$$

Define $g(\lambda) = f_1(x(\lambda)) - f_2(x(\lambda))$. Note that for arbitrary $\lambda_1, \lambda_2 \in [0, 1]$ we have

$$\phi(\lambda_1) \le \lambda_1 f_1(x(\lambda_2)) + (1-\lambda_1) f_2(x(\lambda_2)) \;=\; \phi(\lambda_2) + g(\lambda_2)(\lambda_1 - \lambda_2). \tag{3.1.13}$$

Adding two variants of this inequality with $\lambda_1$ and $\lambda_2$ interchanged, we get

$$(g(\lambda_2) - g(\lambda_1))(\lambda_1 - \lambda_2) \ge 0, \quad \lambda_1, \lambda_2 \in [0, 1].$$

Thus, $g(\cdot)$ is a non-increasing function on $[0, 1]$.

Define $f_i^* = \min\limits_{x\in Q} f_i(x)$, $i = 1, 2$. If $\lambda^* = 1$, then taking in (3.1.13) $\lambda_1 = 1$ and $\lambda_2 = \lambda \in (1, 0]$, we get $g(\lambda) \ge 0$. Therefore, in view of Lemma 3.1.4 we have

$$\begin{aligned}\phi^* &= \lim_{\lambda\to 1}\{\lambda f_1(x(\lambda)) + (1-\lambda) f_2(x(\lambda))\}\\ &\ge \lim_{\lambda\to 1}\{\lambda f(x(\lambda)) + (1-\lambda) f_2^*\} \;\ge\; f^*.\end{aligned}$$

Thus, $\phi^* = f^*$ and in this case equality (3.1.12) is proved. By a symmetric reasoning, we can justify this equality for $\lambda^* = 0$.

Consider now the case $\lambda^* \in (0,1)$. Assume first that there exists a sequence $\{\lambda_k\}_{k\geq 0} \subset [0,1]$ such that

$$\lambda_k \to \lambda^*, \quad g(\lambda_k) \to 0, \tag{3.1.14}$$

as $k \to \infty$. Then, in view of Lemma 3.1.4,

$$\begin{aligned}\phi^* &= \lim_{k\to\infty} \{\lambda_k f_1(x(\lambda_k)) + (1-\lambda_k) f_2(x(\lambda_k))\} = \lim_{k\to\infty} \{f_2(x(\lambda_k)) + \lambda_k g(\lambda_k)\} \\ &= \lim_{k\to\infty} f_2(x(\lambda_k)).\end{aligned}$$

Similarly, we can prove that $\phi^* = \lim\limits_{k\to\infty} f_1(x(\lambda_k))$. Since $\max\{\cdot,\cdot\}$ is a continuous function, we conclude that

$$\phi^* = \lim_{k\to\infty} f(x(\lambda_k)) \geq f^*,$$

which proves (3.1.12) under assumption (3.1.14).

Finally, let us assume that there is no sequence satisfying conditions (3.1.14). Consider two sequences:

$$\{\alpha_k\}_{k\geq 0} : \alpha_k \uparrow \lambda^*, \quad \{\beta_k\}_{k\geq 0} : \beta_k \downarrow \lambda^*.$$

Since the condition (3.1.14) is not satisfied and the function $g$ is monotone, there exist two positive values $a$ and $b$ such that

$$\lim_{k\to\infty} g(\alpha_k) = a, \quad \lim_{k\to\infty} g(\beta_k) = -b.$$

Let $\gamma = \frac{b}{a+b}$. Then, in view of Lemma 3.1.4, we have

$$\begin{aligned}\phi^* &= \lim_{k\to\infty} \{\gamma\phi(\alpha_k) + (1-\gamma)\phi(\beta_k)\} \\ &= \lim_{k\to\infty} \Big\{\gamma[f_2(x(\alpha_k)) + \alpha_k g(\alpha_k)] + (1-\gamma)[f_2(x(\beta_k)) + \beta_k g(\beta_k)]\Big\} \\ &= \lim_{k\to\infty} \Big\{\gamma f_2(x(\alpha_k)) + (1-\gamma) f_2(x(\beta_k))\Big\} \\ &\geq \limsup_{k\to\infty} f_2(\gamma x(\alpha_k) + (1-\gamma)x(\beta_k)).\end{aligned}$$

Similarly,

$$\begin{aligned}\phi^* &= \lim_{k\to\infty}\Big\{\gamma[f_1(x(\alpha_k)) - (1-\alpha_k)g(\alpha_k)]\\ &\quad +(1-\gamma)[f_1(x(\beta_k)) - (1-\beta_k)g(\beta_k)]\Big\}\\ &= \lim_{k\to\infty}\Big\{\gamma f_1(x(\alpha_k)) + (1-\gamma)f_1(x(\beta_k))\Big\}\\ &\geq \limsup_{k\to\infty} f_1(\gamma x(\alpha_k) + (1-\gamma)x(\beta_k)).\end{aligned}$$

Choosing subsequences convergent in the function values, we can see that

$$\phi^* \geq \lim_{k\to\infty} f(\gamma x(\alpha_k) + (1-\gamma)x(\beta_k)) \;\geq\; f^*. \qquad \square$$

**Corollary 3.1.3** *Let functions $f_i$, $i = 1,\dots,m$, be closed and convex on $Q$ and their constrained level sets be bounded. Then there exists some $\lambda_* \in \Delta_m$ such that*

$$\min_{x\in Q}\left(F(x) \overset{\text{def}}{=} \max_{1\leq i\leq m} f_i(x)\right) = \min_{x\in Q}\left\{\sum_{i=1}^m \lambda_*^{(i)} f_i(x)\right\}. \tag{3.1.15}$$

*Proof* In view of the cumbersome notation, we do only the first two steps in our proof by induction. Let $F_k(x) = \max\limits_{k\leq i\leq m} f_i(x)$. Then

$$F(x) = \max\{f_1(x), F_2(x)\}, \quad F_k(x) = \max\{f_k(x), F_{k+1}((x)\},\ k = 2,\dots,m-1.$$

Therefore, by Theorem 3.1.10 there exists a $\lambda_*^{(1)} \in [0,1]$ such that

$$\begin{aligned}F^* \overset{\text{def}}{=} \min_{x\in Q} F(x) \;&=\; \min_{x\in Q}\Big\{\psi_1(x) \;\overset{\text{def}}{=}\; \lambda_*^{(1)} f_1(x) + (1-\lambda_*^{(1)})F_2(x)\Big\}\\ &= \min_{x\in Q}\max\Big\{\lambda_*^{(1)} f_1(x) + (1-\lambda_*^{(1)})f_2(x), \lambda_*^{(1)} f_1(x) + (1-\lambda_*^{(1)})F_3(x)\Big\}.\end{aligned}$$

Again, by Theorem 3.1.10, there exists a $\xi^* \in [0,1]$ such that $F^* = \min\limits_{x\in Q}\psi_2(x)$, where

$$\begin{aligned}\psi_2(x) &= \xi^*(\lambda_*^{(1)} f_1(x) + (1-\lambda_*^{(1)})f_2(x))\\ &\quad +(1-\xi^*)(\lambda_*^{(1)} f_1(x) + (1-\lambda_*^{(1)})F_3(x))\\ &= \lambda_*^{(1)} f_1(x) + \xi^*(1-\lambda_*^{(1)})f_2(x) + (1-\xi^*)(1-\lambda_*^{(1)})F_3(x).\end{aligned}$$

Defining $\lambda_*^{(2)} = \xi^*(1-\lambda_*^{(1)})$, observe that

$$\psi_2(x) = \lambda_*^{(1)} f_1(x) + \lambda_*^{(2)} f_2(x) + (1-\lambda_*^{(1)}-\lambda_*^{(2)})F_3(x).$$

And we can continue. $\square$

Note that the functions $f_i$, $i = 1,\dots,m$, in Corollary 3.1.3 may be discontinuous.

### 3.1.3 Continuity and Differentiability

In the previous sections, we have seen that a behavior of convex function on the boundary of its domain can be unpredictable (see Examples 3.1.1(6) and 3.1.2(5)). Fortunately, this is the only bad thing which can happen. In this section, we will see that the local structure of a convex function in the *interior* of its domain is very simple.

**Theorem 3.1.11** *Let $f$ be convex and $x_0 \in \mathrm{int}\,(\mathrm{dom}\, f)$. Then $f$ is locally bounded and locally Lipschitz continuous at $x_0$.*

*Proof* Let us first prove that $f$ is locally bounded. Let us choose some $\epsilon > 0$ such that $x_0 \pm \epsilon e_i \in \mathrm{int}\,(\mathrm{dom}\, f)$, $i = 1,\dots,n$. Define

$$\Delta = \mathrm{Conv}\,\{x_0 \pm \epsilon e_i,\ i = 1\dots n\} \overset{(3.1.8)}{=} B_1(x_0,\epsilon).$$

Clearly, $\Delta \subseteq \mathrm{dom}\, f$ and, in view of Corollary 3.1.2, we have

$$\max_{x\in\Delta} f(x) = \max_{1\le i\le n} f(x_0 \pm \epsilon e_i) \overset{\text{def}}{=} M. \tag{3.1.16}$$

Consider now a point $y \in B_1(x_0,\epsilon)$, $y \neq x_0$. Let

$$\alpha = \tfrac{1}{\epsilon} \parallel y - x_0 \parallel_{(1)}, \quad z = x_0 + \tfrac{1}{\alpha}(y - x_0).$$

It is clear that $\parallel z - x_0 \parallel_{(1)} = \frac{1}{\alpha} \parallel y - x_0 \parallel_{(1)} = \epsilon$. Therefore, $\alpha \le 1$ and

$$y = \alpha z + (1-\alpha)x_0.$$

Hence,

$$\begin{aligned} f(y) &\le \alpha f(z) + (1-\alpha) f(x_0) \overset{(3.1.16)}{\le} f(x_0) + \alpha(M - f(x_0)) \\ &= f(x_0) + \tfrac{M - f(x_0)}{\epsilon} \parallel y - x_0 \parallel_{(1)} . \end{aligned}$$

Further, let $u = x_0 + \frac{1}{\alpha}(x_0 - y)$. Then $\| u - x_0 \|_{(1)} = \epsilon$ and $y = x_0 + \alpha(x_0 - u)$. Therefore, in view of Theorem 3.1.1, we have

$$f(y) \geq f(x_0) + \alpha(f(x_0) - f(u)) \overset{(3.1.16)}{\geq} f(x_0) - \alpha(M - f(x_0))$$

$$= f(x_0) - \frac{M - f(x_0)}{\epsilon} \| y - x_0 \|_{(1)} .$$

Thus, $| f(y) - f(x_0) | \leq \frac{M - f(x_0)}{\epsilon} \| y - x_0 \|_{(1)}$. □

Let us show that all convex functions possess a property which is very close to differentiability.

**Definition 3.1.3** Let $x \in \mathrm{dom}\, f$. We call $f$ *differentiable* at the point $x$ *in direction* $p \neq 0$ if the following limit exists:

$$f'(x; p) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha}[f(x + \alpha p) - f(x)]. \tag{3.1.17}$$

The value $f'(x; p)$ is called the *directional derivative* of $f$ at $x$.

**Theorem 3.1.12** *A convex function $f$ is differentiable in any direction at any interior point of its domain.*

*Proof* Let $x \in \mathrm{int}\,(\mathrm{dom}\, f)$. Consider the function

$$\phi(\alpha) = \frac{1}{\alpha}[f(x + \alpha p) - f(x)], \quad \alpha > 0.$$

Let $\beta \in (0, 1]$, and the value $\alpha \in (0, \epsilon]$ be small enough to have $x + \epsilon p \in \mathrm{dom}\, f$. Then,

$$f(x + \alpha\beta p) = f((1 - \beta)x + \beta(x + \alpha p)) \leq (1 - \beta) f(x) + \beta f(x + \alpha p).$$

Therefore,

$$\phi(\alpha\beta) = \frac{1}{\alpha\beta}[f(x + \alpha\beta p) - f(x_0)] \leq \frac{1}{\alpha}[f(x + \alpha p) - f(x)] = \phi(\alpha).$$

Thus, $\phi(\alpha)$ decreases as $\alpha \downarrow 0$. Let us choose $\gamma > 0$ small enough to have the point $x - \gamma p$ inside the domain. Then, $x + \alpha p = x + \frac{\alpha}{\gamma}(x - (x - \gamma p))$. Therefore, in view of inequality (3.1.5), we have

$$\phi(\alpha) \geq \frac{1}{\gamma}[f(x) - f(x - \gamma p)].$$

Hence, the limit in the right-hand side of (3.1.17) exists. □

Let us prove that the directional derivative provides us with a global lower support of the initial convex function.

**Lemma 3.1.5** *Let the function $f$ be convex and $x \in \mathrm{int}\,(\mathrm{dom}\, f)$. Then $f'(x;\cdot)$ is a convex function which is positively homogeneous of degree one. For any $y \in \mathrm{dom}\, f$, we have*

$$f(y) \geq f(x) + f'(x; y-x). \tag{3.1.18}$$

*Proof* Let us prove that the directional derivative is homogeneous. Indeed, for any $p \in \mathbb{R}^n$ and $\tau > 0$, we have

$$f'(x;\tau p) = \lim_{\alpha\downarrow 0} \tfrac{1}{\alpha}[f(x+\tau\alpha p) - f(x)]$$

$$= \tau \lim_{\beta\downarrow 0} \tfrac{1}{\beta}[f(x+\beta p) - f(x)] = \tau f'(x_0; p).$$

Further, for any $p_1, p_2 \in \mathbb{R}^n$ and $\beta \in [0,1]$, we obtain

$$\begin{aligned} f'(x;\beta p_1 + (1-\beta)p_2) &= \lim_{\alpha\downarrow 0} \tfrac{1}{\alpha}[f(x+\alpha(\beta p_1 + (1-\beta)p_2)) - f(x)] \\ &\leq \lim_{\alpha\downarrow 0} \tfrac{1}{\alpha}\{\beta[f(x+\alpha p_1) - f(x)] \\ &\quad +(1-\beta)[f(x+\alpha p_2) - f(x)]\} \\ &= \beta f'(x;p_1) + (1-\beta) f'(x;p_2). \end{aligned}$$

Thus, $f'(x;p)$ is convex in $p$. Finally, let $\alpha \in (0,1]$, $y \in \mathrm{dom}\, f$, and $y_\alpha = x + \alpha(y-x)$. Then in view of Theorem 3.1.1, we have

$$f(y) = f(y_\alpha + \tfrac{1}{\alpha}(1-\alpha)(y_\alpha - x)) \;\geq\; f(y_\alpha) + \tfrac{1}{\alpha}(1-\alpha)[f(y_\alpha) - f(x)],$$

and we get (3.1.18) taking the limit as $\alpha \downarrow 0$. □

### *3.1.4 Separation Theorems*

Up to now, we have looked at the properties of convex functions in terms of function values. We have not yet introduce any *directions*, which could be used by minimization schemes. In Convex Analysis, such directions are defined by *separation theorems*, which are presented in this section.

**Definition 3.1.4** Let $Q$ be a convex set. We say that the hyperplane

$$\mathscr{H}(g,\gamma) = \{x \in \mathbb{R}^n \mid \langle g, x\rangle = \gamma\}, \quad g \neq 0,$$

is *supporting* to $Q$ if any $x \in Q$ satisfies inequality $\langle g, x\rangle \leq \gamma$. The hyperplane $\mathscr{H}(g,\gamma) \not\supseteq Q$ *separates* a point $x_0$ from $Q$ if

$$\langle g, x\rangle \leq \gamma \leq \langle g, x_0\rangle \tag{3.1.19}$$

for all $x \in Q$. If one of the inequalities in (3.1.19) is strict, the we call the separation *strong*. □

In a similar way, we define separability of convex sets. Two sets $Q_1$ and $Q_2$ are called *separable* if there exist $g \in \mathbb{R}^n$, $g \neq 0$, and $\gamma \in \mathbb{R}$ such that

$$\langle g, x\rangle \leq \gamma \ \leq \langle g, y\rangle \quad \forall x \in Q_1,\ y \in Q_2. \tag{3.1.20}$$

The separation is *strict* if one of the inequalities in (3.1.20) is strict. We call the separation *strong* if

$$\sup_{x\in Q_1} \langle g, x\rangle < \gamma \ < \inf_{y\in Q_2} \langle g, y\rangle. \tag{3.1.21}$$

All separation theorems in $\mathbb{R}^n$ can be derived from the properties of Euclidean projection. Let us first describe the possibilities for strong separation.

**Theorem 3.1.13** *Let $Q_1$ and $Q_2$ be closed convex sets in $\mathbb{R}^n$ such that $Q_1 \bigcap Q_2 = \emptyset$. These sets are strongly separable provided that one of them is bounded.*

*Proof* Suppose that $Q_1$ is bounded. Consider the following minimization problem:

$$\rho^* = \min_{x\in Q_1} \ \rho_{Q_2}(x).$$

Note that the optimal value of this problem is positive and its optimal set $X^*$ is not empty. Moreover, for all $x^* \in X^*$, we have

$$\nabla\rho_{Q_2}(x^*) \overset{(2.2.41)}{=} g^*, \quad \langle g^*, x^*\rangle \overset{(2.2.41)}{=} \gamma^*.$$

Therefore, for all $x_1 \in Q_1$ we have

$$\langle g^*, x_1\rangle - \gamma^* \overset{(2.2.41)}{=} \langle \nabla\rho_Q(x^*), x_1 - x^*\rangle \overset{(2.2.39)}{\geq} 0.$$

On the other hand, for all $x_2 \in Q_2$ we have

$$\begin{aligned}\langle g^*, x_2\rangle - \gamma^* &\overset{(2.2.41)}{\leq} \langle x^* - \pi_{Q_2}(x^*), x_2 - x^*\rangle \overset{(2.2.47)}{\leq} -\|x^* - \pi_{Q_2}(x^*)\|^2 \\ &= -(\rho^*)^2.\end{aligned}$$

□

*Remark 3.1.1* The assumption of boundedness of one of the sets in Theorem 3.1.13 cannot be omitted. To see why, consider the separation problem for sets $Q$ and $\mathbb{R}^{1,2}_{+}$ in Example 2.2.1. □

**Corollary 3.1.4** *Let $Q$ be a closed convex set and $x \notin Q$. Then $x$ is strongly separable from $Q$.* □

Let us give an example of application of this important fact.

**Corollary 3.1.5** *Let $Q_1$ and $Q_2$ be two closed convex sets.*

*1. If $\xi_{Q_1}(g) \leq \xi_{Q_2}(g)$ for all $g \in \mathrm{dom}\, \psi_{Q_2}$, then $Q_1 \subseteq Q_2$.*

*2. Let $\mathrm{dom}\, \xi_{Q_1} = \mathrm{dom}\, \xi_{Q_2}$, and for any $g \in \mathrm{dom}\, \xi_{Q_1}$ we have $\xi_{Q_1}(g) = \xi_{Q_2}(p)$. Then $Q_1 \equiv Q_2$.*

*Proof*

1. Assume that there exists an $x_0 \in Q_1$ which does not belong to $Q_2$. Then, in view of Corollary 3.1.5, there exists a direction $g$ such that

$$\langle g, x_0 \rangle > \gamma \geq \langle g, x \rangle$$

for all $x \in Q_2$. Hence, $g \in \mathrm{dom}\, \xi_{Q_2}$ and $\xi_{Q_1}(g) > \xi_{Q_2}(g)$. This is a contradiction.

2. In view of the first statement, $Q_1 \subseteq Q_2$ and $Q_2 \subseteq Q_1$. Therefore, $Q_1 \equiv Q_2$. □

The next separation theorem deals with boundary points of convex sets.

**Theorem 3.1.14** *Let $Q$ be a closed convex set. If the point $x_0$ belongs to the boundary of $Q$, then there exists a supporting to $Q$ hyperplane $\mathscr{H}(g, \gamma)$ which contains $x_0$.*

(Such a vector $g$ is called *supporting to $Q$* at the point $x_0$.)

*Proof* Consider a sequence $\{y_k\}$ such that $y_k \notin Q$ and $y_k \to x_0$. Let

$$g_k = \tfrac{y_k - \pi_Q(y_k)}{\|y_k - \pi_Q(y_k)\|}, \quad \gamma_k = \langle g_k, \pi_Q(y_k) \rangle.$$

In view of Corollary 3.1.5, for all $x \in Q$ we have

$$\langle g_k, x \rangle \leq \gamma_k \leq \langle g_k, y_k \rangle. \tag{3.1.22}$$

However, $\| g_k \|= 1$ and, in view of Lemma 2.2.8, the sequence $\{\gamma_k\}$ is bounded:

$$\begin{aligned} \mid \gamma_k \mid &= \mid \langle g_k, \pi_Q(y_k) - x_0 \rangle + \langle g_k, x_0 \rangle \mid \leq \| \pi_Q(y_k) - x_0 \| + \| x_0 \| \\ &\leq \| y_k - x_0 \| + \| x_0 \| . \end{aligned}$$

Therefore, without loss of generality, we can assume that there exist $g^* = \lim\limits_{k\to\infty} g_k$ and $\gamma^* = \lim\limits_{k\to\infty} \gamma_k$. It remains to take the limit in inequalities (3.1.22). □

### *3.1.5 Subgradients*

Now we are ready to introduce a generalization of the notion of the gradient.

**Definition 3.1.5** A vector $g$ is called a *subgradient* of the function $f$ at the point $x_0 \in \mathrm{dom}\, f$ if for any $y \in \mathrm{dom}\, f$ we have

$$f(y) \geq f(x_0) + \langle g, y - x_0 \rangle. \tag{3.1.23}$$

The set of all subgradients of $f$ at $x_0$, $\partial f(x_0)$, is called the *subdifferential* of the function $f$ at the point $x_0$.

If inequality (3.1.23) is valid only for points $y \in Q$, we use notation $g \in \partial_Q f(x_0)$. The latter set is called *constrained subdifferential*. Clearly, $\partial f(x_0) \subseteq \partial_Q f(x_0)$ for any convex set $Q \subseteq \mathrm{dom}\, f$.

For concave functions, we define *super-gradients* and *super-differentials* by changing the sign in inequality (3.1.23). Note that $\partial f(x_0)$ can be nonempty even for nonconvex $f$.

A simple consequence of Definition 3.1.5 is as follows:

$$\langle g_1 - g_2, x_1 - x_2 \rangle \geq 0 \quad \forall x_1, x_2 \in \mathrm{dom}\, f,\ \ g_1 \in \partial f(x_1),\ \ g_2 \in \partial f(x_2). \tag{3.1.24}$$

The necessity of introducing the notion of subdifferential is clear from the following example.

*Example 3.1.3* Consider the function $f(x) = (x)_+ \overset{\mathrm{def}}{=} \max\{x, 0\}$, $x \in \mathbb{R}$. For all $y \in \mathbb{R}$ and $g \in [0, 1]$, we have

$$f(y) = \max\{y, 0\} \ \geq \ g \cdot y \ = \ f(0) + g \cdot (y - 0).$$

Therefore subgradient of $f$ at $x = 0$ is not uniquely defined. In our example, this is an arbitrary value from the interval [0, 1]. □

The whole set of conditions (3.1.23) parameterized by $y \in Q$ can be seen as a set of linear *inequality constraints* for $g$, defining the set $\partial_Q f(x_0)$. Therefore, by definition, any subdifferential is a *closed convex* set.

Let us prove that subdifferentiability of function $f$ at all points of some convex set implies convexity and closedness of the function.

**Lemma 3.1.6** *Let $Q$ be a convex set. Assume that, for any $x \in Q \subseteq \mathrm{dom}\, f$, the constrained subdifferential $\partial_Q f(x)$ is nonempty. Then $f$ is a closed convex function on $Q$.*

*Proof* For any $x \in Q$, define $\hat{f}(x) = \sup\limits_{y}\{f(y) + \langle g(y), x - y\rangle : \ y \in Q\} \geq f(x)$, where $g(y)$ is an arbitrary subgradient from $\partial_Q f(y)$. By Theorem 3.1.8, $\hat{f}$ is a closed convex function, and $f(x) \overset{(3.1.23)}{\geq} \hat{f}(x)$ for all $x \in Q$. □

On the other hand, we can prove a relaxed converse statement.

**Theorem 3.1.15** *Let the function $f$ be convex. If $x_0 \in \mathrm{int}\,(\mathrm{dom}\, f)$, then $\partial f(x_0)$ is a nonempty bounded set.*

*Proof* Since the point $(f(x_0), x_0)$ belongs to the boundary of epi $(f)$, in view of Theorem 3.1.14, there exists a hyperplane supporting to epi $(f)$ at $(f(x_0), x_0)$:

$$-\alpha\tau + \langle d, x\rangle \leq -\alpha f(x_0) + \langle d, x_0\rangle \tag{3.1.25}$$

for all $(\tau, x) \in$ epi $(f)$. Let us normalize the coefficients of hyperplane in order to satisfy the condition

$$\| \, d \, \|^2 + \alpha^2 = 1, \tag{3.1.26}$$

where the norm is standard Euclidean. Since the point $(\tau, x_0)$ belongs to epi $(f)$ for all $\tau \geq f(x_0)$, we conclude that $\alpha \geq 0$.

In view of Theorem 3.1.11 a convex function is locally Lipschitz continuous in the interior of its domain. This means that there exist some $\epsilon > 0$ and $M > 0$ such that $B_2(x_0, \epsilon) \subseteq \mathrm{dom}\, f$ and

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

for all $x \in B_2(x_0, \epsilon)$.[2] Therefore, in view of (3.1.25), for any $x$ from this ball

$$\langle d, x - x_0\rangle \leq \alpha(f(x) - f(x_0)) \leq \alpha M \| \, x - x_0 \, \| \, .$$

Choosing $x = x_0 + \epsilon d$, we get $\| \, d \, \|^2 \leq M\alpha \| \, d \, \|$. Thus, in view of normalizing condition (3.1.26), we get $\alpha \geq [1 + M^2]^{-1/2}$. Hence, choosing $g = d/\alpha$, we obtain

$$f(x) \overset{(3.1.25)}{\geq} f(x_0) + \langle g, x - x_0\rangle$$

for all $x \in \mathrm{dom}\, f$.

[2] In the proof of Theorem 3.1.11, we worked with the $\ell_1$-norm. However, the result remains valid for any norm in $\mathbb{R}^n$, since in finite dimensions all norms are topologically equivalent.

Finally, if $g \in \partial f(x_0)$, $g \neq 0$, then choosing $x = x_0 + \epsilon g / \parallel g \parallel$ we obtain

$$\epsilon \parallel g \parallel = \langle g, x - x_0 \rangle \ \leq \ f(x) - f(x_0) \ \leq \ M \parallel x - x_0 \parallel = \ M\epsilon.$$

Thus, $\partial f(x_0)$ is bounded. □

The next example shows that the statement of Theorem 3.1.15 cannot be strengthened.

*Example 3.1.4* Consider function $f(x) = -\sqrt{x}$ with domain $\mathbb{R}_+$. This function is convex and closed, but the subdifferential does not exist at $x = 0$. □

Sub-differentiability at $x \in \operatorname{dom} f$ is an important characteristic of the local structure of the function $f$ around this point. Let us prove the following fact.

**Theorem 3.1.16** *For the function $f$, define its Fenchel dual*

$$f_*(s) = \sup_{y \in \operatorname{dom} f} [\langle s, y \rangle - f(y)], \tag{3.1.27}$$

*and the dual of the Fenchel dual:*

$$f_{**}(x) = \sup_{s \in \operatorname{dom} f_*} [\langle s, x \rangle - f_*(s)].$$

*Then $f(x) \geq f_{**}(x)$ for all $x \in \operatorname{dom} f$. Moreover, if $\partial f(x) \neq \emptyset$ for some $x \in \operatorname{dom} f$, then $\partial f(x) \subseteq \operatorname{dom} f_*$ and $f(x) = f_{**}(x)$.*

*Proof* Indeed, for any $x \in \operatorname{dom} f$ we have

$$\begin{aligned} f_{**}(x) \quad &= \sup_{s \in \operatorname{dom} f_*} [\langle s, x \rangle - f_*(s)] \overset{(3.1.27)}{=} \sup_{s \in \operatorname{dom} f_*} \inf_{y \in \operatorname{dom} f} [\langle s, x \rangle - \langle s, y \rangle + f(y)] \\ &\overset{(1.3.6)}{\leq} \inf_{y \in \operatorname{dom} f} \sup_{s \in \operatorname{dom} f_*} [\langle s, x - y \rangle + f(y)] \overset{y=x}{\leq} f(x). \end{aligned}$$

Let us choose now an arbitrary $g \in \partial f(x)$. Then for any $y \in \operatorname{dom} f$ we have

$$\langle g, y \rangle - f(y) \overset{(3.1.23)}{\leq} \langle g, y \rangle - f(x) - \langle g, y - x \rangle \ = \ \langle g, x \rangle - f(x).$$

Thus, $g \in \operatorname{dom} f_*$. Therefore,

$$\begin{aligned} f_{**}(x) &= \sup_{s \in \operatorname{dom} f_*} \inf_{y \in \operatorname{dom} f} [\langle s, x \rangle - \langle s, y \rangle + f(y)] \\ &\geq \inf_{y \in \operatorname{dom} f} [\langle g, x \rangle - \langle g, y \rangle + f(y)] \overset{(3.1.23)}{=} f(x). \qquad \square \end{aligned}$$

Let us prove an important relation between subdifferential and directional derivatives of a convex function.

**Theorem 3.1.17** *Let the function $f$ be convex, and $x_0 \in \operatorname{int}(\operatorname{dom} f)$. Then*

$$\partial_2 f'(x_0; 0) = \partial f(x_0),$$

*where the subdifferential $\partial_2$ corresponds to the second argument of the function $f(x_0; \cdot)$. Moreover, for any $p \in \mathbb{R}^n$, we have*

$$f'(x_0; p) = \max\{\langle g, p\rangle \mid g \in \partial f(x_0)\}. \tag{3.1.28}$$

*Proof* Note that

$$f'(x_0; p) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha}[f(x_0 + \alpha p) - f(x_0)] \;\geq\; \langle g, p\rangle, \tag{3.1.29}$$

where $g$ is an arbitrary vector from $\partial f(x_0)$. Therefore, the subdifferential of the function $f'(x_0; \cdot)$ at $p = 0$ is not empty and $\partial f(x_0) \subseteq \partial_2 f'(x_0; 0)$. On the other hand, since $f'(x_0; p)$ is convex in $p$, in view of Lemma 3.1.5, for any $y \in \operatorname{dom} f$ we have

$$f(y) \geq f(x_0) + f'(x_0; y - x_0) \;\geq\; f(x_0) + \langle g, y - x_0\rangle,$$

where $g \in \partial_2 f'(x_0; 0)$. Thus, $\partial_2 f'(x_0; 0) \subseteq \partial f(x_0)$ and we see that these two sets coincide.

Consider $g \in \partial_2 f'(x_0; p)$. Then, in view of inequality (3.1.18), for all $v \in \mathbb{R}^n$ and $\tau > 0$ we have

$$\tau f'(x_0; v) = f'(x_0; \tau v) \;\geq\; f'(x_0; p) + \langle g, \tau v - p\rangle.$$

Considering $\tau \to \infty$ we get

$$f'(x_0; v) \geq \langle g, v\rangle, \tag{3.1.30}$$

and, considering $\tau \to 0$, we obtain

$$f'(x_0; p) - \langle g, p\rangle \leq 0. \tag{3.1.31}$$

However, inequality (3.1.30) implies that $g \in \partial_2 f'(x_0; 0)$. Therefore, comparing (3.1.29) and (3.1.31), we conclude that $\langle g, p\rangle = f'(x_0; p)$. □

Let us mention some properties of subgradients, which are of central importance for Convex Optimization. The next result forms the basis for the *cutting plane* optimization schemes.

**Theorem 3.1.18** *For any $x_0 \in \mathrm{dom}\, f$, all vectors $g \in \partial f(x_0)$ are supporting to the level set $\mathscr{L}_f(f(x_0))$:*

$$\langle g, x_0 - x\rangle \geq 0 \quad \forall x \in \mathscr{L}_f(f(x_0)) \;=\; \{x \in \mathrm{dom}\, f : \; f(x) \leq f(x_0)\}.$$

*Proof* Indeed, if $f(x) \leq f(x_0)$ and $g \in \partial f(x_0)$, then

$$f(x_0) + \langle g, x - x_0\rangle \leq f(x) \leq f(x_0). \qquad \square$$

**Corollary 3.1.6** *Let $Q \subseteq \mathrm{dom}\, f$ be a closed convex set, $x_0 \in Q$, and*

$$x^* \in \operatorname{Arg}\min_{x \in Q} \; f(x).$$

*Then for any $g \in \partial f(x_0)$, we have $\langle g, x_0 - x^*\rangle \geq 0$.* $\square$

In some situations, the following objects are very useful.

**Definition 3.1.6** Let the set $X \subseteq \mathrm{dom}\, f$ be closed and convex. The set

$$\widehat{\partial f}(X) = \bigcap_{x \in X} \partial f(x) \tag{3.1.32}$$

is called the *epigraph facet* of the set $X$.

This definition is motivated by the following statement.

**Theorem 3.1.19** *Let the set $X$ be closed and convex, and $\widehat{\partial f}(X) \neq \emptyset$. Then*

$$f((1-\alpha)x_0 + \alpha x_1) = (1-\alpha)f(x_0) + \alpha f(x_1), \;\; \forall x_0, x_1 \in X, \;\; \alpha \in [0,1]. \tag{3.1.33}$$

*Moreover, for any $g \in \widehat{\partial f}(X)$ and all $x_0$, $x_1$ from $X$, we have*

$$f(x_1) = f(x_0) + \langle g, x_1 - x_0\rangle. \tag{3.1.34}$$

*Proof* Indeed, let $g \in \widehat{\partial f}(X) \subseteq \partial f(x_0) \bigcap \partial f(x_1)$. Then,

$$f(x_0) + \langle g, x_1 - x_0\rangle \overset{(3.1.23)}{\leq} f(x_1) \overset{(3.1.23)}{\leq} f(x_0) + \langle g, x_1 - x_0\rangle.$$

Thus, (3.1.34) is proved. Consequently, for $x_\alpha = (1-\alpha)x_0 + \alpha x_1$ with $\alpha \in [0,1]$, we have

$$\begin{aligned}(1-\alpha)f(x_0) + \alpha f(x_1) &\overset{(3.1.2)}{\geq} f(x_\alpha) \overset{(3.1.23)}{\geq} f(x_0) + \langle g, x_\alpha - x_0\rangle \\ &= f(x_0) + \alpha\langle g, x_1 - x_0\rangle \overset{(3.1.34)}{=} (1-\alpha)f(x_0) + \alpha f(x_1).\end{aligned}$$

Thus, we have proved equality (3.1.33). $\square$

Let us show how the epigraph facets arise in optimality conditions for Unconstrained Optimization.

**Theorem 3.1.20** *Let* $X^* = Arg \min\limits_{x \in \mathrm{dom}\, f} f(x)$. *Then a closed convex set* $X_*$ *is a subset of* $X^*$ *if and only if*

$$0 \in \widehat{\partial f}(X_*).$$

*Proof* Indeed, if $0 \in \widehat{\partial f}(X_*)$, then for any $x^* \in X_*$ and all $x \in \mathrm{dom}\, f$ we have

$$f(x) \geq f(x^*) + \langle 0, x - x^* \rangle \; = \; f(x^*).$$

Thus, $x^* \in X^*$.

On the other hand, if $f(x) \geq f(x^*)$ for all $x \in \mathrm{dom}\, f$ and $x^* \in X_*$, then by Definition 3.1.5, $0 \in \bigcap\limits_{x^* \in X_*} \partial f(x^*)$. □

In what follows, for a set-valued mapping $\mathscr{S}(\cdot)$ and arbitrary set $X \subseteq \mathbb{R}^n$, we use the notation $\widehat{\mathscr{S}}(X) \overset{\mathrm{def}}{=} \bigcap\limits_{x \in X} \mathscr{S}(x)$.

### *3.1.6 Computing Subgradients*

In the previous section, we introduced subgradients, the objects which we are going to use in minimization methods. However, in order to apply such methods for solving real-life problems, we need to be sure that subgradients are computable. In this section, we present the corresponding computational rules. Note that for the majority of minimization methods, it is enough to be able to compute a single subgradient from the set $\partial f(x)$.

Let us first establish some relations between gradients and subgradients.

**Lemma 3.1.7** *Let a function* $f$ *be convex. Assume that it is differentiable at a point* $x \in \mathrm{int}\,(\mathrm{dom}\, f)$. *Then* $\partial f(x) = \{\nabla f(x)\}$.

*Proof* Indeed, for any direction $p \in \mathbb{R}^n$ , we have

$$f'(x; p) = \langle \nabla f(x), p \rangle.$$

It remains to use Theorem 3.1.17 and Item 2 of Corollary 3.1.5. □

**Lemma 3.1.8** *Let a function* $\psi(\cdot)$ *be convex and* $\varphi$ *be a univariate convex function, which is non-decreasing on the set*

$$\mathrm{Im}\, \psi = \{\tau = \psi(x), \;\; x \in \mathrm{dom}\, \psi\}.$$

*Then the function* $f(\cdot) = \varphi(\psi(\cdot))$ *is convex and for any* $x$ *from* $\mathrm{int}\,(\mathrm{dom}\,\psi)$ *we have*

$$\partial f(x) = \mathrm{Conv}\,\{\lambda\partial\psi(x),\ \lambda \in \partial\varphi(\psi(x))\}.$$

*Proof* Indeed the function $f$ is convex in view of Theorem 3.1.9. Let us fix an arbitrary $x \in \mathrm{int}\,(\mathrm{dom}\,\psi)$ and any direction $h$. Then, by the chain rule for directional derivatives, we have

$$\begin{aligned} f'(x;p) = \varphi'(\psi(x);\psi'(x;p)) \ &= \ \max_{\lambda}\{\lambda\psi'(x;p) :\ \lambda \in \partial\varphi(\psi(x))\} \\ &= \max_{\lambda,g}\{\langle g,p\rangle :\ g \in \lambda\partial\psi(x),\ \lambda \in \partial\varphi(\psi(x))\}. \end{aligned}$$

It remains to use Theorem 3.1.17 and Item 2 of Corollary 3.1.5. □

Consider now a mixed situation when the function $f(x,y)$ depends on two variables $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$.

**Lemma 3.1.9** *Let a function* $f$ *be convex, and*

$$\bar{z} = (\bar{x},\bar{y}) \in \mathrm{int}\,(\mathrm{dom}\,f) \ \subseteq \ \mathbb{R}^n \times \mathbb{R}^m.$$

*Assume that* $f$ *is differentiable in the first variable, and the corresponding partial gradient* $\nabla_1 f(\cdot,\cdot) \in \mathbb{R}^n$ *is continuous at* $\bar{z}$ *along any direction in* $\mathbb{R}^{n+m}$. *Then*

$$\partial f(\bar{z}) = (\nabla_1 f(\bar{x},\bar{y}), \partial_2 f(\bar{x},\bar{y})),$$

*where* $\partial_2 f(x,y) \subset \mathbb{R}^m$ *is the partial subdifferential of* $f$ *with respect to the second variable, when the first variable is fixed.*

*Proof* Let us fix an arbitrary direction $h = (h_x, h_y) \in \mathbb{R}^n \times \mathbb{R}^m$. Then for $\alpha > 0$ small enough, we have

$$\begin{aligned} \tfrac{1}{\alpha}(f(\bar{x}+\alpha h_x, \bar{y}+\alpha h_y) - f(\bar{x},\bar{y})) &= \tfrac{1}{\alpha}(f(\bar{x}+\alpha h_x, \bar{y}+\alpha h_y) - f(\bar{x}, \bar{y}+\alpha h_y)) \\ &\quad +\tfrac{1}{\alpha}(f(\bar{x}, \bar{y}+\alpha h_y) - f(\bar{x},\bar{y})). \end{aligned}$$

Since $f$ is convex, we have

$$\begin{aligned} \alpha\langle\nabla_1 f(\bar{x}, \bar{y}+\alpha h_y), h_x\rangle &\overset{(2.1.2)}{\leq} f(\bar{x}+\alpha h_x, \bar{y}+\alpha h_y) - f(\bar{x}, \bar{y}+\alpha h_y) \\ &\overset{(2.1.2)}{\leq} \alpha\langle\nabla_1 f(\bar{x}+\alpha h_x, \bar{y}+\alpha h_y), h_x\rangle. \end{aligned}$$

Hence, in view of the directional continuity of $\nabla_1 f$, we have

$$\begin{aligned} f'(\bar{z}, h) \quad &= \quad \langle \nabla_1 (f(\bar{x}, \bar{y}), h_x\rangle + f'(\bar{z}, (0, h_y)) \\ &\overset{(3.1.28)}{=} \langle \nabla_1 (f(\bar{x}, \bar{y}), h_x\rangle + \max_g\{\langle g, h_y\rangle :\ g \in \partial_2 f(\bar{x}, \bar{y})\}. \end{aligned}$$

It remains to use Corollary 3.1.5. □

Finally, let us present a converse statement, which derives differentiability from a kind of continuous subdifferentiability.

**Lemma 3.1.10** *Let $f$ be convex and $x_0 \in \operatorname{int}(\operatorname{dom} f)$. Assume that there exists a vector function $g(x) \in \partial f(x)$ which is continuous at $x_0$. Then $f$ is differentiable at $x_0$ and $\nabla f(x_0) = g(x_0)$.*

*Proof* Indeed, for any direction $h \in \mathbb{R}^n$ and small enough positive $\alpha$, we have

$$\langle g(x_0), h\rangle \overset{(3.1.23)}{\leq} \tfrac{1}{\alpha}[f(x_0 + \alpha h) - f(x_0)] \overset{(3.1.23)}{\leq} \langle g(x_0 + \alpha h), h\rangle.$$

Thus, taking the limit as $\alpha \downarrow 0$, we get $f'(x_0; h) = \langle g(x_0), h\rangle$ for all $h \in \mathbb{R}^n$. Hence, $g(x_0) = \nabla f(x_0)$. □

Let us provide all operations for convex functions, described in Sect. 3.1.2, with corresponding chain rules for updating subgradients.

**Lemma 3.1.11** *Let the function $f$ be closed and convex on the bounded set $S \subseteq \operatorname{dom} f \subseteq \mathbb{R}^m$. Consider a linear operator*

$$\mathscr{A}(x) = Ax + b : \quad \mathbb{R}^n \to \mathbb{R}^m.$$

*Then $\phi(x) = f(\mathscr{A}(x))$ is a closed convex function on the set*

$$Q = \{x \mid \mathscr{A}(x) \in S\}.$$

*For any $x \in Q$ with nonempty $\partial f(\mathscr{A}(x))$ we have*

$$\partial \phi(x) = A^T \partial f(\mathscr{A}(x)).$$

*Proof* We have already proved the first part of this lemma in Theorem 3.1.6. Let us prove the relation for the subdifferential. Let $y_0 = \mathscr{A}(x_0)$. Then for all $p \in \mathbb{R}^n$, we have

$$\begin{aligned} \phi'(x_0, p) = f'(y_0; Ap) \ &= \ \max\{\langle g, Ap\rangle \mid g \in \partial f(y_0)\} \\ &= \max\{\langle \bar{g}, p\rangle \mid \bar{g} \in A^T \partial f(y_0)\}. \end{aligned}$$

Using Theorem 3.1.17 and Corollary 3.1.5, we get $\partial \phi(x_0) = A^T \partial f(\mathscr{A}(x_0))$. □

**Lemma 3.1.12** *Let functions $f_1$ and $f_2$ be closed and convex, and $\alpha_1, \alpha_2 \geq 0$. Then the function $f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x)$ is also closed and convex and*

$$\partial f(x) = \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x) \tag{3.1.35}$$

*for any $x$ from* $\operatorname{int}(\operatorname{dom} f) = \operatorname{int}(\operatorname{dom} f_1) \bigcap \operatorname{int}(\operatorname{dom} f_2)$.

*Proof* In view of Theorem 3.1.5, we need to prove only the relation for the subdifferentials. Consider $x_0 \in \operatorname{int}(\operatorname{dom} f_1) \bigcap \operatorname{int}(\operatorname{dom} f_2)$. In view of Theorem 3.1.15, at this point both subdifferentials are bounded. For any $p \in \mathbb{R}^n$, we have

$$\begin{aligned} f'(x_0; p) &= \alpha_1 f_1'(x_0; p) + \alpha_2 f_2'(x_0; p) \\ &= \max\{\langle g_1, \alpha_1 p\rangle \mid g_1 \in \partial f_1(x_0)\} \\ &\quad + \max\{\langle g_2, \alpha_2 p\rangle \mid g_2 \in \partial f_2(x_0)\} \\ &= \max\{\langle \alpha_1 g_1 + \alpha_2 g_2, p\rangle \mid g_1 \in \partial f_1(x_0), \; g_2 \in \partial f_2(x_0)\} \\ &= \max\{\langle g, p\rangle \mid g \in \alpha_1 \partial f_1(x_0) + \alpha_2 \partial f_2(x_0)\}. \end{aligned}$$

Hence, using Theorem 3.1.17 and Corollary 3.1.5, we get (3.1.35). □

**Lemma 3.1.13** *Let functions $f_i$, $i = 1 \ldots m$, be closed and convex. Then the function $f(x) = \max\limits_{1 \leq i \leq m} f_i(x)$ is closed and convex. For any $x \in$* $\operatorname{int}(\operatorname{dom} f) = \bigcap\limits_{i=1}^{m} \operatorname{int}(\operatorname{dom} f_i)$*, we have*

$$\partial f(x) = \operatorname{Conv}\{\partial f_i(x) \mid i \in I(x)\}, \tag{3.1.36}$$

*where $I(x) = \{i : f_i(x) = f(x)\}$.*

*Proof* Again, in view of Theorem 3.1.5, we need to justify only the rules for subdifferentials. Consider $x \in \bigcap\limits_{i=1}^{m} \operatorname{int}(\operatorname{dom} f_i)$. In view of Theorem 3.1.15, at this point, subdifferentials of all functions $f_i$ are bounded.

For the sake of notation, assume that $I(x) = \{1, \ldots, k\}$. Then for any $p \in \mathbb{R}^n$, we have

$$f'(x; p) = \max_{1 \leq i \leq k} f_i'(x; p) = \max_{1 \leq i \leq k} \max\{\langle g_i, p\rangle \mid g_i \in \partial f_i(x)\}.$$

Note that for any set of values $a_1, \dots, a_k$ we have

$$\max_{1 \le i \le k} a_i = \max \left\{ \sum_{i=1}^{k} \lambda_i a_i \mid \{\lambda_i\} \in \Delta_k \right\},$$

where $\Delta_k = \{\lambda_i \ge 0, \sum\limits_{i=1}^{k} \lambda_i = 1\}$ is the standard $k$-dimensional *simplex*. Therefore,

$$\begin{aligned} f'(x; p) &= \max_{\{\lambda_i\} \in \Delta_k} \{\sum_{i=1}^{k} \lambda_i \max\{\langle g_i, p\rangle \mid g_i \in \partial f_i(x)\}\} \\ &= \max\{\langle \sum_{i=1}^{k} \lambda_i g_i, p\rangle \mid g_i \in \partial f_i(x), \ \{\lambda_i\} \in \Delta_k\} \\ &= \max\{\langle g, p\rangle \mid g = \sum_{i=1}^{k} \lambda_i g_i, g_i \in \partial f_i(x), \{\lambda_i\} \in \Delta_k\} \\ &= \max\{\langle g, p\rangle \mid g \in \mathrm{Conv}\,\{\partial f_i(x), i \in I(x)\}\}. \qquad \square \end{aligned}$$

The last rule can be useful for computing some elements from subdifferentials.

**Lemma 3.1.14** *Let $\Delta$ be an arbitrary set, and $f(x) = \sup\{\phi(x, y) \mid \ y \in \Delta\}$. Suppose that for any $y \in \Delta$ the function $\phi(\cdot, y)$ is closed and convex on some convex set $Q$. Then $f$ is closed convex on the set*

$$\hat{Q} = \left\{ x \in Q \mid \sup_{y \in \Delta} \phi(x, y) < +\infty \right\}.$$

*Moreover, for any $x \in \hat{Q}$ we have*

$$\partial_{\hat{Q}} f(x) \supseteq \mathrm{Conv}\,\{\partial_{Q,x}\phi(x, y) \mid y \in I(x)\},$$

*where $I(x) = \{y \in \Delta \mid \ \phi(x, y) = f(x)\}$.*

*Proof* In view of Theorem 3.1.8, we have to prove only the inclusion. Indeed, for any $x \in \hat{Q}$, $y_0 \in I(x_0)$, and $g_0 \in \partial_{Q,x}\phi(x_0, y_0)$, we have

$$f(x) \ge \phi(x, y_0) \ \ge \ \phi(x_0, y_0) + \langle g_0, x - x_0\rangle \ = \ f(x_0) + \langle g_0, x - x_0\rangle. \qquad \square$$

Now we can look at some examples of subdifferentials.

*Example 3.1.5*

1. Let $f(x) = (x)_+, x \in \mathbb{R}$. Then $\partial f(0) = [0,1]$ since $f(x) = \max\limits_{g \in [0,1]} g\,x$.
2. Consider the function $f(x) = \sum\limits_{i=1}^{m} \mid \langle a_i, x \rangle \mid$. Define
$$I_-(x) = \{i : \langle a_i, x \rangle < 0\},$$
$$I_+(x) = \{i : \langle a_i, x \rangle > 0\},$$
$$I_0(x) \;= \{i : \langle a_i, x \rangle = 0\}.$$
Then $\partial f(x) = \sum\limits_{i \in I_+(x)} a_i - \sum\limits_{i \in I_-(x)} a_i + \sum\limits_{i \in I_0(x)} [-a_i, a_i]$.
3. Consider the function $f(x) = \max\limits_{1 \le i \le n} x^{(i)}$. Define $I(x) = \{i : \; x^{(i)} = f(x)\}$. Then
$$\partial f(x) = \text{Conv}\,\{e_i \mid \; i \in I(x)\}.$$
For $x = 0$, we have $\partial f(0) = \text{Conv}\,\{e_i \mid \; 1 \le i \le n\} \equiv \Delta_n$.
4. For the Euclidean norm $f(x) = \parallel x \parallel$, we have
$$\partial f(0) = B_2(0,1) = \{x \in \mathbb{R}^n \mid \parallel x \parallel \le 1\},$$
$$\partial f(x) = \{x / \parallel x \parallel\}, \; x \neq 0.$$
5. For the $\ell_1$-norm, $f(x) = \parallel x \parallel_1 = \sum\limits_{i=1}^{n} \mid x^{(i)} \mid$, we have
$$\partial f(0) = B_\infty(0,1) = \{x \in \mathbb{R}^n \mid \max_{1 \le i \le n} \mid x^{(i)} \mid \le 1\},$$
$$\partial f(x) = \sum_{i \in I_+(x)} e_i - \sum_{i \in I_-(x)} e_i + \sum_{i \in I_0(x)} [-e_i, e_i], \; x \neq 0,$$
where $I_+(x) = \{i \mid x^{(i)} > 0\}$, $I_-(x) = \{i \mid x^{(i)} < 0\}$ and $I_0(x) = \{i \mid x^{(i)} = 0\}$.
6. In the case of the Minkowski function, we need to introduce a *polar* of the set $Q$:
$$\mathscr{P}_Q = \{g \in \mathbb{R}^n : \; \langle g, x \rangle \le 1 \, \forall x \in Q\}. \tag{3.1.37}$$

Then

$$\partial\psi_Q(0) = \mathscr{P}_Q, \quad \partial\psi_Q(x) \;=\; \mathrm{Arg}\max_{g\in\mathscr{P}_Q} \langle g, x\rangle.$$

We leave the justification of these examples as an exercise for the reader. □

Finally, let us describe subgradients of homogeneous functions.

**Definition 3.1.7** A function $f$ is called (positively) homogeneous of degree $p \geq 0$ if dom $f$ is a cone and

$$f(\tau x) = \tau^p f(x) \quad \forall x \in \mathrm{dom}\, f,\ \forall \tau \geq 0. \tag{3.1.38}$$

Note that all functions in Example 3.1.5 are homogeneous of degree one.

**Theorem 3.1.21 (Euler's Homogeneous Function Theorem)** *Let the function $f$ be convex and subdifferentiable on its domain. If it is homogeneous of degree $p \geq 1$, then*

$$\langle g, x\rangle = pf(x) \quad \forall x \in \mathrm{dom}\, f,\ \forall g \in \partial f(x). \tag{3.1.39}$$

*Proof* Indeed, let $x \in \mathrm{dom}\, f$ and $g \in \partial f(x)$. Then for any $\tau \geq 0$ we have

$$\tau^p f(x) \overset{(3.1.38)}{=} f(\tau x) \overset{(3.1.23)}{\geq} f(x) + (\tau - 1)\langle g, x\rangle.$$

For $\tau > 1$, this implies that $\frac{\tau^p-1}{\tau-1} f(x) \geq \langle g, x\rangle$. Therefore, taking the limit as $\tau \downarrow 1$, we get $pf(x) \geq \langle g, x\rangle$.

For $\tau < 1$, the above inequality implies $\frac{1-\tau^p}{1-\tau} f(x) \leq \langle g, x\rangle$. Hence, taking the limit as $\tau \uparrow 1$, we get $pf(x) \leq \langle g, x\rangle$. □

In Convex Analysis, the most important homogeneous functions have degree of homogeneity one. For such functions,

$$\langle g, x\rangle \overset{(3.1.39)}{=} f(x) \quad \forall x \in \mathrm{dom}\, f,\ \forall g \in \partial f(x). \tag{3.1.40}$$

From now on, let us assume that dom $f = \mathbb{R}^n$. Then, for all $x \in \mathbb{R}^n$ we have

$$f(x) = f'(0, x) \overset{(3.1.28)}{=} \max_g\{\langle g, x\rangle : \ g \in \partial f(0)\}. \tag{3.1.41}$$

The simplest example of the homogeneous function is a linear function $f(x) = \langle a, x\rangle$. A more important case is a general norm. For $f(x) = \|x\|$, we have

$$\|x\| = \max_g\{\langle g, x\rangle : \ \|g\|_* \leq 1\},$$

where $\|g\|_* = \max\limits_x \{\langle g, x\rangle : \ \|x\| \leq 1\}$ is the *dual norm*. Thus,

$$\partial \|x\| \Big|_{x=0} = \{g \in \mathbb{R}^n : \ \|g\|_* \leq 1\}. \tag{3.1.42}$$

**Lemma 3.1.15** *Let a function $f$ be convex and homogeneous of degree one with* dom $f = \mathbb{R}^n$. *Then for all $x \in \mathbb{R}^n$, we have*

$$\partial f(x) = \{g \in \partial f(0) : \ \langle g, x\rangle = f(x)\}. \tag{3.1.43}$$

*Proof* Denote the right-hand side of equality (3.1.43) by $G(x)$. If $g \in \partial f(x)$, then for any $y \in \mathbb{R}^m$ we have

$$f(y) \overset{(3.1.23)}{\geq} f(x) + \langle g, y - x\rangle \overset{(3.1.40)}{=} \langle g, y\rangle.$$

Thus, $g \in \partial f(0)$. Consequently, $g \overset{(3.1.40)}{\in} G(x)$. On the other hand, if $g \in G(x)$, then for any $y \in \mathbb{R}^n$ we have

$$f(y) \overset{(3.1.23)}{\geq} \langle g, y\rangle = f(x) + \langle g, y - x\rangle.$$

Therefore, $g \in \partial f(x)$. □

Thus, in view of equality (3.1.41), $\partial f(x)$ is a *facet* of $\partial f(0)$.

Let us give an example of application for the machinery developed so far.

**Theorem 3.1.22** *Let $Q_1$ and $Q_2$ be bounded closed convex sets with intersection $Q = Q_1 \bigcap Q_2$, which has nonempty interior. Then*

$$\xi_Q(x) = \min_{y \in \mathbb{R}^n} \left\{\xi_{Q_1}(x + y) + \xi_{Q_2}(-y)\right\}, \quad x \in \mathbb{R}^n. \tag{3.1.44}$$

*Proof* Let us first prove that the optimization problem in (3.1.44) is solvable. If $g \in Q_1 \bigcap Q_2$, then for any $y \in \mathbb{R}^n$ we have

$$\phi_x(y) \overset{\text{def}}{=} \xi_{Q_1}(x + y) + \xi_{Q_2}(-y) \geq \langle g, x + y\rangle + \langle g, -y\rangle = \langle g, x\rangle.$$

Thus the objective function in (3.1.44) is bounded below and for its infimum $\phi_x^*$ we have $\phi_x^* \geq \xi_Q(x)$. Consider a sequence $\{y_k\}$ such that $\phi_x(y_k) \to \phi_x^*$. If this sequence is bounded, then the infimum is attained. If not, then we can have $t_k \overset{\text{def}}{=} \|y_k\| \to \infty$. Let $\bar{y}_k = \frac{1}{t_k} y_k$. Then

$$\lim_{k\to\infty} \phi_x(\bar{y}_k) = \lim_{k\to\infty} [\xi_{Q_1}(\tfrac{1}{t_k} x + \bar{y}_k) + \xi_{Q_2}(-\bar{y}_k)] = \lim_{k\to\infty} \tfrac{1}{t_k} \phi_x(y_k) = 0.$$

Since the sequence $\{\bar{y}_k\}$ is bounded, we can assume that it is convergent to a point $\bar{y}$ with $\|\bar{y}\| = 1$ and $\phi_x(\bar{y}) = 0$. In this case, we have

$$\langle g_1, \bar{y}\rangle \le \xi_{Q_1}(\bar{y}) \;\le\; -\xi_{Q_2}(-\bar{y}) \;=\; \langle g_2, \bar{y}\rangle, \quad \forall g_1 \in Q_1,\; \forall g_2 \in Q_2.$$

Hence, $\langle g, \bar{y}\rangle = 0$ for all $g \in Q$, and we get a contradiction with the assumptions.

Denote by $y^*$ the solution of the optimization problem in (3.1.44). In view of Theorem 3.1.20, we have

$$0 \in \partial\phi_x(y^*) \overset{(3.1.35)}{=} \partial\xi_{Q_1}(x + y^*) + \partial\xi_{-Q_2}(y^*).$$

In view of Lemma 3.1.15 this means that there exists a vector $g$ such that

$$g \in Q_1, \quad \langle g, x + y^*\rangle \;=\; \xi_{Q_1}(x + y^*),$$

$$-g \in -Q_2, \quad \langle -g, y^*\rangle \;= \xi_{-Q_2}(y^*).$$

Thus, $\phi_x^* = \xi_{Q_1}(x + y^*) + \xi_{Q_2}(-y^*) = \xi_{Q_1}(x + y^*) + \xi_{-Q_2}(y^*) = \langle g, x\rangle$. Since $g \in Q$, we conclude that $\phi_x^* \le \xi_Q(x)$. □

Finally, let us describe subgradients of superpositions of convex functions and differentiable convex functions.

**Lemma 3.1.16** *Consider $\psi(g) = \max\limits_{\lambda\in\Lambda}\langle\lambda, g\rangle$, where $\Lambda \subset \mathbb{R}^m_+$ is a bounded convex set. Let the vector function $F(x) = (f_1(x), \ldots, f_m(x))$, $x \in \mathbb{R}^n$, have differentiable convex components. Then the superposition $f(x) = \psi(F(x))$ is convex and*

$$\partial f(x) = \left\{\sum_{i=1}^{m} \lambda^{(i)}\nabla f_i(x) :\; \lambda \in Arg\max_{\lambda\in\Lambda}\langle\lambda, F(x)\rangle\right\}. \tag{3.1.45}$$

*Proof* Indeed the function $\psi(\cdot)$ is monotone: if $g_1 \le g_2$ in the component-wise sense, then $\psi(g_1) \le \psi(g_2)$. Therefore, for any $x, y$ from $\mathbb{R}^n$ and $\alpha \in [0, 1]$ we have

$$f(\alpha x + (1-\alpha)y) \le \psi(\alpha F(x) + (1-\alpha)F(y)) \;\le\; \alpha f(x) + (1-\alpha)f(y).$$

Relation (3.1.45) follows from the representation of directional derivatives. Define $F'(x) = (\nabla f_1(x), \ldots, \nabla f_m(x)) \in \mathbb{R}^{n\times m}$. Then for any direction $h \in \mathbb{R}^n$ we have

$$\begin{aligned} f'(x; h) \quad &= \quad \psi'(F(x); (F'(x))^T h) \\ &\overset{(3.1.43)}{=} \max\{\langle\lambda, (F'(x))^T h\rangle :\; \lambda \in \operatorname{Arg}\max_{\lambda\in\Lambda}\langle\lambda, F(x)\rangle\}. \end{aligned}$$ □

**Lemma 3.1.17** *Let $F$ be a differentiable convex and monotone function on $\mathbb{R}^m$ and suppose the functions $f_i$ are convex on a convex open set $Q$. Then the function*

$$\phi(x) = F(f_1(x), \dots, f_m(x))$$

*is convex on $Q$ and*

$$\partial\phi(x) = \sum_{i=1}^{m} \nabla_i F(f(x)) \cdot \partial f_i(x), \quad x \in Q, \tag{3.1.46}$$

*where $f(x) = (f_1(x), \dots, f_m(x))^T \in \mathbb{R}^m$.*

*Proof* Indeed, for $x, y \in Q$ and $\alpha \in [0, 1]$ we have

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

Further, for any direction $p \in \mathbb{R}^n$,

$$\phi'(x; p) = \sum_{i=1}^{m} \nabla_i F(f(x)) f_i'(x; p) \overset{(3.1.28)}{=} \sum_{i=1}^{m} \nabla_i F(f(x)) \xi_{\partial f_i(x)}(p).$$

It remains to use Corollary 3.1.5. □

**Corollary 3.1.7** *If all $f_i$, $i = 1, \dots, m$, are convex, then the function*

$$\phi(x) = \ln\left(\sum_{i=1}^{m} e^{f_i(x)}\right) \tag{3.1.47}$$

*is also convex.*

*Proof* Indeed, we have seen in Example 2.1.1(4) that the function

$$F(s) = \ln\left(\sum_{i=1}^{n} e^{s^{(i)}}\right)$$

is convex and monotone on $\mathbb{R}^n$. □

### 3.1.7 Optimality Conditions

Let us apply the developed technique to derive different optimality conditions. We start with a simple minimization problem, where the objective function has a *composite form*:

$$\min_{x \in Q} \left\{ \tilde{f}(x) \overset{\text{def}}{=} f(x) + \Psi(x) \right\}, \tag{3.1.48}$$

where $Q$ is a closed convex set, $f \in C^1(Q)$ is a continuously differentiable convex function and $\Psi$ is a closed convex function defined on the set $Q$.

**Theorem 3.1.23** *A point $x^*$ is a solution to problem (3.1.48) if and only if for every $x \in Q$ we have*

$$\langle \nabla f(x^*), x - x^* \rangle + \Psi(x) \geq \Psi(x^*). \tag{3.1.49}$$

*Proof* Indeed, if condition (3.1.49) is satisfied, then

$$\begin{aligned} \tilde{f}(x) &= f(x) + \Psi(x) \overset{(2.1.2)}{\geq} f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \Psi(x) \\ &\overset{(3.1.49)}{\geq} f(x^*) + \Psi(x^*) = \tilde{f}(x^*). \end{aligned}$$

Assume now that $x^*$ is an optimal solution of the minimization problem (3.1.48). Suppose that there exists an $x \in Q$ such that

$$\langle \nabla f(x^*), x - x^* \rangle + \Psi(x) < \Psi(x^*).$$

Note that $\lim\limits_{\alpha \downarrow 0} \frac{1}{\alpha}[f(\alpha x + (1-\alpha)x^*) - f(x^*)] = \langle \nabla f(x^*), x - x^* \rangle$. Thus, for a positive $\alpha$ small enough we have

$$\begin{aligned} f(\alpha x + (1-\alpha)x^*) &< f(x^*) + \alpha[\Psi(x^*) - \Psi(x)] \\ &= \tilde{f}(x^*) + \alpha[\Psi(x^*) - \Psi(x)] - \Psi(x^*) \\ &\overset{(3.1.2)}{\leq} \tilde{f}(x^*) - \Psi(\alpha x + (1-\alpha)x^*). \end{aligned}$$

Hence, $\tilde{f}(\alpha x + (1-\alpha)x^*) < \tilde{f}(x^*)$ and we get a contradiction. □

In view of Definition 3.1.5, condition (3.1.49) is equivalent to the inclusion

$$-\nabla f(x^*) \in \partial_Q \Psi(x^*).$$

Let us now look at optimization problems with general objective functions. Consider the problem

$$\min_{x \in Q} f(x), \tag{3.1.50}$$

where $Q \subseteq \mathbb{R}^n$ is a closed convex set and $f$ is a closed convex function, $\mathrm{dom}\, f \supset Q$. For a point $\bar{x} \in Q$, define the *normal cone*:

$$\mathscr{N}(\bar{x}) = \{g \in \mathbb{R}^n \mid \langle g, x - \bar{x} \rangle \geq 0, \ \forall x \in Q\}. \tag{3.1.51}$$

Since inclusion $g \in \mathscr{N}(\bar{x})$ implies $\tau g \in \mathscr{N}(\bar{x})$ for any $\tau \geq 0$, this is indeed a cone. It is closed and convex as an intersection of closed convex sets, the half-spaces

$$\{g : \langle g, x - \bar{x} \rangle \geq 0\}, \quad x \in Q.$$

Clearly, $\mathscr{N}(\bar{x}) = \{0_n\}$ for all $\bar{x} \in \operatorname{int} Q$. Thus, this cone is nontrivial only at the boundary points $\bar{x} \in \partial Q$.

For $\bar{x} \in Q$, define the *tangent cone*

$$\mathscr{T}(\bar{x}) = \{p \in \mathbb{R}^n | \ \langle g, p \rangle \geq 0, \ \forall g \in \mathscr{N}(\bar{x})\}. \tag{3.1.52}$$

Thus, this is a standard *dual cone* to $\mathscr{N}(\bar{x})$. Again, this cone is closed and convex as the intersection of the system of half-spaces. Clearly, for $\bar{x} \in \operatorname{int} Q$ we have $\mathscr{T}(\bar{x}) = \mathbb{R}^n$.

The name of the cone $\mathscr{T}(\cdot)$ is justified by the following property.

**Lemma 3.1.18** *Let $\bar{x} \in \partial Q$. Then $Q - \bar{x} \subset \mathscr{T}(\bar{x})$. Moreover,*

$$\mathscr{T}(\bar{x}) = \operatorname{cl}\left(\mathscr{K}(Q - \bar{x})\right). \tag{3.1.53}$$

*Thus, $\mathscr{T}(\bar{x})$ is the closure of the conic hull of the set $Q - \bar{x}$.*

*Proof* Indeed, in view of the definition of normal cone (3.1.51), we have

$$\langle g, x - \bar{x} \rangle \geq 0, \quad \forall x \in Q, \ g \in \mathscr{N}(\bar{x}).$$

Therefore, $Q - \bar{x} \overset{(3.1.52)}{\subset} \mathscr{T}(\bar{x})$. Since $\mathscr{T}(\bar{x})$ is a closed cone, this means that

$$\bar{\mathscr{K}} \overset{\text{def}}{=} \operatorname{cl}\left(\mathscr{K}(\bar{x})\right) \subseteq \mathscr{T}(\bar{x}).$$

Let us assume that there exists a point $\bar{p} \in \mathscr{T}(\bar{x})$ such that $\bar{p} \notin \bar{\mathscr{K}}$. Then, by Corollary 3.1.4, there exists a direction $\bar{g}$ which strongly separates $\bar{p}$ from $\bar{\mathscr{K}}$:

$$\langle \bar{g}, \bar{p} \rangle < \gamma \ \leq \ \langle \bar{g}, \alpha(x - \bar{x}) \rangle, \quad \forall x \in Q, \ \alpha \geq 0.$$

Letting $\alpha \to +\infty$ in this inequality, we get $\langle \bar{g}, x - \bar{x} \rangle \geq 0$ for all $x \in Q$. Thus, direction $\bar{g}$ belongs to the cone $\mathscr{N}(\bar{x})$. On the other hand, taking $\alpha = 0$, we get $\gamma \leq 0$. Thus, $\langle \bar{g}, \bar{p} \rangle < 0$. This means that $\bar{p} \overset{(3.1.52)}{\notin} \mathscr{T}(\bar{x})$. Hence, we get a contradiction. □

*Remark 3.1.2* For the special case $Q = \{x \in \mathbb{R}^n : \ Ax = b\}$, where $A$ is an $(m \times n)$-matrix, standard arguments from Linear Algebra prove the following representation:

$$\begin{aligned} \mathscr{N}(\bar{x}) &= \{g \in \mathbb{R}^n : \ g = A^T y, \ y \in \mathbb{R}^m\}, \\ \mathscr{T}(\bar{x}) &= \{h \in \mathbb{R}^n : \ Ah = 0\}, \end{aligned} \tag{3.1.54}$$

which is valid for all $\bar{x} \in Q$.

The next statement gives us an optimality condition for a linearized version of problem (3.1.50).

**Lemma 3.1.19** *Let $x^*$ be an optimal solution to problem (3.1.50). Then*

$$f'(x^*; p) \geq 0 \quad \forall p \in \mathscr{T}(x^*). \tag{3.1.55}$$

*Proof* Assume that there exists a point $\bar{p} \in \mathscr{T}(x^*)$ such that $f'(x^*, \bar{p}) < 0$. In view of Lemma 3.1.18, there exist two sequences $\{\alpha_k\} \subset \mathbb{R}_+$ and $\{x_k\} \subset Q$ such that

$$\bar{p} = \lim_{k \to \infty} \alpha_k (x_k - x^*).$$

Since the function $f'(x^*; \cdot)$ is continuous, in view of Lemma 3.1.5, we have

$$\begin{aligned} 0 > f'(x^*; \bar{p}) \ &= \ \lim_{k \to \infty} \alpha_k f'(x^*; x_k - x^*) \\ &= \lim_{k \to \infty} \lim_{\beta \downarrow 0} \tfrac{\alpha_k}{\beta} [f(x^* + \beta(x_k - x^*)) - f(x^*)] \ \geq \ 0. \end{aligned}$$

Thus, we come to a contradiction. □

Now we can justify an optimality condition for problem (3.1.50). Define

$$X^* = \operatorname{Arg} \min_{x \in Q} f(x).$$

**Theorem 3.1.24** *A point $x^*$ from $Q$ belongs to $X^*$ if and only if there exists a $g^* \in \partial f(x^*)$ such that*

$$\langle g^*, x - x^* \rangle \geq 0 \quad \forall x \in Q. \tag{3.1.56}$$

*In this case, $g^* \in \widehat{\partial f}(X^*) \bigcap \widehat{\mathscr{N}}(X^*)$ (see Definition 3.1.6).*

*Proof* Indeed, from the condition (3.1.56) and definition of $\partial f(x^*)$, we have

$$f(x) \overset{(3.1.23)}{\geq} f(x^*) + \langle g^*, x - x^* \rangle \overset{(3.1.56)}{\geq} f(x^*) \quad \forall x \in Q.$$

Thus, $x^* \in X^*$.

Let us prove the converse statement. Let $x^* \in X^*$ be an optimal solution of problem (3.1.50). Assume that there is no $g \in \partial f(x^*)$ such that

$$\langle g, x - x^* \rangle \geq 0 \quad \forall x \in Q.$$

In view of definition (3.1.51), this means that $\partial f(x^*) \bigcap \mathscr{N}(x^*) = \emptyset$. Consider the following auxiliary optimization problem:

$$\min_{g_1, g_2} \left\{ \phi(g_1, g_2) = \tfrac{1}{2} \| g_1 - g_2 \|^2 : \; g_1 \in \partial f(x^*), \; g_2 \in \mathscr{N}(x^*) \right\},$$

where the norm is standard Euclidean. Since the set $\partial f(x^*)$ is bounded, there exists its optimal solution $(g_1^*, g_2^*)$ and the optimal value $\rho^* \overset{\text{def}}{=} \phi(g_1^*, g_2^*)$ is positive. Let us write down optimality conditions for this auxiliary problem. By Theorem 2.2.9, we obtain

$$\langle \nabla_{g_1} \phi(g_1^*, g_2^*), g_1 - g_1^* \rangle = \langle g_1^* - g_2^*, g_1 - g_1^* \rangle \;\geq\; 0 \quad \forall g_1 \in \partial f(x^*), \tag{3.1.57}$$

$$\langle \nabla_{g_2} \phi(g_1^*, g_2^*), g_2 - g_2^* \rangle = \langle g_2^* - g_1^*, g_2 - g_2^* \rangle \;\geq\; 0 \quad \forall g_2 \in \mathscr{N}(x^*). \tag{3.1.58}$$

Taking in (3.1.58) $g_2 = 0$ and $g_2 = \alpha g_2^*$ as $\alpha \to +\infty$, we get

$$\langle g_2^* - g_1^*, g_2^* \rangle \leq 0 \;\leq\; \langle g_2^* - g_1^*, g_2^* \rangle.$$

Thus, for $p^* \overset{\text{def}}{=} g_2^* - g_1^*$ we have $\langle g_2^*, p^* \rangle = 0$. Therefore,

$$\langle g_2, p^* \rangle \overset{(3.1.58)}{\geq} 0 \quad \forall g_2 \in \mathscr{N}(x^*),$$

which means $p^* \overset{(3.1.52)}{\in} \mathscr{T}(x^*)$. On the other hand, for all $g_1 \in \partial f(x^*)$ we have

$$\langle g_1, p^* \rangle \overset{(3.1.57)}{\leq} \langle g_1^*, p^* \rangle \;=\; \langle g_1^* - g_2^*, p^* \rangle \;=\; -2\rho^*.$$

This means that $f'(x^*; p^*) \overset{(3.1.28)}{=} -2\rho^* < 0$. Thus, we get a contradiction with Lemma 3.1.19 and prove the existence of a vector $g^* \in \partial f(x^*)$ such that

$$\langle g^*, x - x^* \rangle \geq 0 \quad \forall x \in Q.$$

Note that for any other point $x_1^* \in X^*$ we have

$$f(x^*) = f(x_1^*) \overset{(3.1.23)}{\geq} f(x^*) + \langle g^*, x_1^* - x^* \rangle \;\geq\; f(x^*).$$

Hence, $\langle g^*, x_1^* - x^* \rangle = 0$ and we conclude that $g^* \in \partial f(x_1^*)$. Consequently, $g^* \in \widehat{\partial f}(X_*)$. For the same reason, $g^*$ belongs both to $\mathscr{N}(x^*)$ and $\mathscr{N}(x_1^*)$. □

*Remark 3.1.3* For $x^* \in \operatorname{int} Q$, condition (3.1.56) is equivalent to the inclusion of Theorem 3.1.20.

*Remark 3.1.4* In the special case $Q = \{x \in \mathbb{R}^n : \; Ax = b\}$, where $A$ is an $(m \times n)$-matrix, in view of representation (3.1.54), the statement of Theorem 3.1.24 can be specified in the following way:

$$\begin{array}{l}\text{A point } x^* \text{ belongs to } X^* \text{ if and only if there exists a}\\ g^* \in \partial f(x^*) \text{ such that } g^* = A^T y^* \text{ for some } y^* \in \mathbb{R}^m.\end{array} \tag{3.1.59}$$

(Compare with the statement of Corollary 1.2.1.)

Theorem 3.1.24 is one of the most powerful tools of Convex Analysis. Let us demonstrate this with several important examples.

First of all, consider the differentiation rules for a partial minimum of a convex function (3.1.9).

**Theorem 3.1.25** *Let $\phi$ be a closed convex function, and $Q_1 \subseteq \mathbb{R}^n$ and $Q_2 \subseteq \mathbb{R}^m$ be two closed convex sets such that $Q_1 \times Q_2 \subseteq \operatorname{dom} \phi$. Define*

$$f(x) = \inf_{y \in Q_2} \phi(x, y).$$

*Then $f$ is convex on $Q_1$. Moreover, if $Y(x) \overset{\text{def}}{=} \operatorname{Arg} \min\limits_{y \in Q_2} \phi(x, y) \neq \emptyset$, then*

$$\begin{array}{rl}\partial_{Q_1} f(x) \supseteq & \{g_x \in \mathbb{R}^n : \exists g_y \text{ such that } (g_x, g_y) \in \bigcap\limits_{y \in Y(x)} \partial \phi(x, y), \\ & \quad \text{and } \langle g_y, y - y_x \rangle \geq 0 \quad \forall y \in Q_2, \; \forall y_x \in Y(x)\}.\end{array} \tag{3.1.60}$$

*Proof* The convexity of the function $f$ was already proved in Theorem 3.1.7. Let us fix a point $x \in Q_1$ with $Y(x) \neq \emptyset$. In view of Theorem 3.1.24, the right-hand side of inclusion (3.1.60) is not empty. Consider an arbitrary element $(g_x, g_y)$ from this set. Let $x_1 \in Q_1$ and $\epsilon > 0$. Choosing a point $y_1 \in Q_2$ such that $\phi(x_1, y_1) \leq f(x_1) + \epsilon$, we get

$$\begin{aligned} f(x_1) + \epsilon \geq \phi(x_1, y_1) \; & \geq \; \phi(x, y_x) + \langle g_x, x_1 - x \rangle + \langle g_y, y_1 - y_x \rangle \\ & \geq \phi(x, y_x) + \langle g_x, x_1 - x \rangle \; = \; f(x) + \langle g_x, x_1 - x \rangle. \end{aligned}$$

Since we can choose $\epsilon$ arbitrarily small, inclusion $g_x \in \partial_{Q_1} f(x)$ is proved. □

**Corollary 3.1.8** *If $Y(x) \neq \emptyset$ for all $x \in \mathrm{dom}\, f$, then $f$ is a closed convex function on $Q_1$.*

*Proof* By inclusion (3.1.60), $\partial f(x) \neq \emptyset$. Therefore, we can apply Lemma 3.1.6. □

Note that separability of the constraints $x \in Q_1$ and $y \in Q_2$ is essential for the validity of the rule (3.1.60). Simple examples show that in the general situation of Theorem 3.1.7, the set $\partial f(x)$ can be dependent also on the partial subgradients of function $\phi$ in $y$. Such a general case can be treated by Theorem 3.1.28.

Let us look now at optimality conditions for smooth minimization problem with functional constraints:

$$\min_{x \in Q} \{ f_0(x) \mid f_i(x) \leq 0,\ i = 1, \ldots, m \}, \tag{3.1.61}$$

where $Q$ is a closed convex set.

**Theorem 3.1.26 (Karush–Kuhn–Tucker)** *Let functions $f_i$, $i = 0 \ldots m$, be convex and differentiable with $\mathrm{int}\,(\mathrm{dom}\, f_i) \supset Q$ . Suppose that there exists a point $\bar{x} \in Q$ such that*

$$f_i(\bar{x}) < 0, \quad i = 1, \ldots, m. \quad \text{(Slater condition for inequalities)} \tag{3.1.62}$$

*A point $x^*$ is an optimal solution of problem (3.1.61) if and only if there exist nonnegative values $\lambda_i^*$, $i = 1 \ldots m$, satisfying the following conditions:*

$$\begin{aligned} \langle \nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*), x - x^* \rangle \geq 0, \quad \forall x \in Q, \\ \lambda_i^* f_i(x^*) = 0, \quad i = 1, \ldots, m. \end{aligned} \tag{3.1.63}$$

*Proof* In view of Lemma 2.3.4, $x^*$ is an optimal solution to problem (3.1.61) if and only if it is a global minimizer of the function

$$\phi(x) = \max\{ f_0(x) - f^*;\ f_i(x), i = 1 \ldots m \}$$

over the set $Q$. In view of Theorem 3.1.24, this is the case if and only if there exists a $g^* \in \partial \phi(x^*)$ such that

$$\langle g^*, x - x^* \rangle \geq 0 \quad \forall x \in Q.$$

Further, in view of Lemma 3.1.13, inclusion $g^* \in \partial f(x^*)$ is equivalent to the existence of nonnegative weights $\bar{\lambda}_i$, $i = 0, \dots, m$, such that

$$\bar{\lambda}_0 \nabla f_0(x^*) + \sum_{i \in I^*} \bar{\lambda}_i \nabla f_i(x^*) = g^*,$$

$$\bar{\lambda}_0 + \sum_{i \in I^*} \bar{\lambda}_i = 1,$$

where $I^* = \{i \in \{1, \dots, m\} : \ f_i(x^*) = 0\}$.

Thus, we need to prove only that $\bar{\lambda}_0 > 0$. Indeed, if $\bar{\lambda}_0 = 0$, then

$$\sum_{i \in I^*} \bar{\lambda}_i f_i(\bar{x}) \geq \sum_{i \in I^*} \bar{\lambda}_i [f_i(x^*) + \langle \nabla f_i(x^*), \bar{x} - x^* \rangle] \ \geq \ 0.$$

This contradicts the Slater condition. Therefore $\bar{\lambda}_0 > 0$ and we can take $\lambda_i^* = \bar{\lambda}_i / \bar{\lambda}_0$ for all $i \in I^*$ and $\lambda_i^* = 0$ for $i \notin I^*$. □

Theorem 3.1.26 is very useful for solving simple optimization problems.

**Lemma 3.1.20** *Let $A \succ 0$. Then*

$$\max_x \{\langle c, x \rangle : \ \langle Ax, x \rangle \leq 1\} = \langle c, A^{-1} c \rangle^{1/2}. \tag{3.1.64}$$

*Proof* Note that all conditions of Theorem 3.1.26 are satisfied and the solution $x^*$ of the above problem is attained at the boundary of the feasible set. Therefore, in accordance with Theorem 3.1.26, we have to solve the following equations:

$$c = \lambda^* A x^*, \quad \langle A x^*, x^* \rangle = 1.$$

Thus, $\lambda^* = \langle c, A^{-1} c \rangle^{1/2}$ and $x^* = \frac{1}{\lambda^*} A^{-1} c$. □

The values $\lambda_i^* \geq 0$, $i = 1, \dots, m$, are called optimal *dual (Lagrange)* multipliers for problem (3.1.61). We can get some upper bounds for these values from the depth of the Slater condition (3.1.62).

**Lemma 3.1.21** *Any point $\bar{x}$, feasible for problem (3.1.61), generates the following upper bound on the magnitude of optimal dual multipliers:*

$$f_0(\bar{x}) - f_0(x^*) \geq \sum_{i=1}^m (-f_i(\bar{x})) \lambda_i^*. \tag{3.1.65}$$

*Proof* Indeed,

$$\begin{aligned}
& f_0(\bar{x}) + \sum_{i=1}^{m} \lambda_i^* f_i(\bar{x}) \\
\overset{(2.1.2)}{\geq} \; & f_0(x^*) + \langle \nabla f_0(x^*), \bar{x} - x^* \rangle + \sum_{i=1}^{m} \lambda_i^* [f_i(x^*) + \langle \nabla f_i(x^*), \bar{x} - x^* \rangle] \\
= \; & f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \langle \nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*), \bar{x} - x^* \rangle \\
\overset{(3.1.63)}{\geq} \; & f_0(x^*). \qquad \square
\end{aligned}$$

The statement of Lemma 3.1.21 can be used to construct an *exact penalty function* for problem (3.1.61). Let the point $\bar{x} \in Q$ satisfy Slater condition (3.1.62). Assume that we know some upper bound $D$ for the gap $f_0(\bar{x}) - f_0(x^*)$. For example, it can be found by the following optimization problem:

$$D = \max_{x \in Q} \langle \nabla f_0(\bar{x}), \bar{x} - x \rangle.$$

Consider the set $\Lambda = \{\lambda \in \mathbb{R}^m_+ : \sum\limits_{i=1}^{m} (-f_i(\bar{x}))\lambda_i \leq D\}$. In view of Lemma 3.1.21, we have $\lambda^* \in \Lambda$. Define the following nonsmooth penalty function:

$$\Psi(g) = \max_{\lambda \in \Lambda} \langle \lambda, g \rangle \; = \; D \left( \max_{1 \leq i \leq m} \frac{g^{(i)}}{-f_i(\bar{x})} \right)_+, \quad g \in \mathbb{R}^m, \tag{3.1.66}$$

where $(a)_+ = \max\{0, a\}$.

Consider the following minimization problem:

$$\min_{x \in Q} \left\{ \phi(x) \overset{\text{def}}{=} f_0(x) + \Psi(f(x)) \right\}, \tag{3.1.67}$$

where $f(x) = (f_1(x), \ldots, f_m(x))$. Let us compute its subdifferential at the point $x^*$, the solution of problem (3.1.61).

Note that $\max\limits_{\lambda \in \Lambda} \langle \lambda, f(x^*) \rangle = 0$. In accordance with the rules of Lemma 3.1.16, we can form the set

$$\Lambda_+ = \{\lambda \in \Lambda : \langle \lambda, f(x^*) \rangle = 0\} \; = \; \{\lambda \in \Lambda : \lambda_i = 0, \; i \notin I^*(x)\},$$

where $I(x^*) = \{i : \ f_i(x^*) = 0\}$. Since $\lambda^* \in \Lambda_+$, in view of Lemma 3.1.16 we have

$$g^* = \nabla f_0(x^*) + \sum_{i \in I(x^*)} \lambda_i^* \nabla f_i(x^*) \ \in \partial \phi(x^*).$$

Hence, by Theorem 3.1.26 and Theorem 3.1.24, $x^* \in \operatorname{Arg}\min\limits_{x \in Q} \phi(x)$. Thus, the optimal values of problems (3.1.67) and (3.1.61) coincide.

Let $\hat{x}$ be an arbitrary optimal solution to problem (3.1.67). Then, by Theorem 3.1.24 and Lemma 3.1.16, there exists a vector $\hat{\lambda} \in \operatorname{Arg}\max\limits_{\lambda \in \Lambda} \langle \lambda, f(\hat{x}) \rangle$ such that

$$\langle \nabla f_0(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla f_i(\hat{x}), x - \hat{x} \rangle \geq 0, \quad \forall x \in Q.$$

Let us assume that $\Psi(f(\hat{x})) > 0$. Then the inequality constraint in the definition of the set $\Lambda$ is active and we have $\langle \hat{\lambda}, -f(\bar{x}) \rangle = D$. However,

$$\begin{aligned} D \geq f_0(\bar{x}) - f_0(\hat{x}) \ &\geq \ \langle \nabla f_0(\hat{x}), \bar{x} - \hat{x} \rangle \ \geq \ \sum_{i=1}^{m} \hat{\lambda}_i \langle \nabla f_i(\hat{x}), \hat{x} - \bar{x} \rangle \\ &\geq \langle \hat{\lambda}, f(\hat{x}) - f(\bar{x}) \rangle \ = \ \Psi(f(\hat{x})) + D. \end{aligned}$$

This contradiction proves that $\Psi(f(\hat{x})) = 0$. Therefore, this point is feasible for problem (3.1.61) and it attains the optimal value of the objective function.

In some situations, the optimization methods based on the exact penalty may look more attractive than the two-level procedure described in Sect. 2.3.5. However, note that for these methods it is necessary to know the point $\bar{x}$ satisfying the Slater condition (3.1.62). If this condition is not "deep" enough, the resulting penalty function can have bad bounds on the derivatives. This slows down the minimization schemes.

The Slater condition in the form (3.1.62) cannot work for equality constraints. Let us show how it can be modified in order to justify the Karush–Kuhn–Tucker condition for a minimization problem of the following form:

$$\min_{x \in Q} \{ f(x) : \ Ax = b \}, \tag{3.1.68}$$

where $Q$ is a closed convex set and the matrix $A \in \mathbb{R}^{m \times n}$ has full row rank.

**Theorem 3.1.27** *Let a function $f$ be convex on $Q \subset \operatorname{int}(\operatorname{dom} f)$ and its level sets on $Q$ be bounded. Suppose that there exist a point $\bar{x}$ and $\epsilon > 0$ such that*

$$A\bar{x} = b, \quad B(\bar{x}, \epsilon) \subseteq Q. \quad \text{(Slater condition for equalities)} \tag{3.1.69}$$

*A point $x^* \in Q$ is an optimal solution for problem (3.1.68) if and only if $Ax^* = b$ and there exist $y^* \in \mathbb{R}^m$ and $g^* \in \partial f(x^*)$ such that*

$$\langle g^* - A^T y^*, x - x^* \rangle \geq 0 \quad \forall x \in Q. \tag{3.1.70}$$

*The magnitude of the vector $y^*$ can be estimated as follows:*

$$\| A^T y^* \| \leq \tfrac{1}{\epsilon} \left( \max_{x \in B(\bar{x},\epsilon)} f(x) - \min_{x \in Q} f(x) \right). \tag{3.1.71}$$

*Proof* Indeed, if condition (3.1.70) is satisfied, then for any $x \in Q$ with $Ax = b$ we have

$$f(x) - f(x^*) \overset{(3.1.23)}{\geq} \langle g^*, x - x^* \rangle \overset{(3.1.70)}{\geq} \langle y^*, A(x - x^*) \rangle \; = \; 0.$$

To prove the converse statement, consider the function

$$\phi(x) = f(x) + K \| b - Ax \|,$$

where the norm is standard Euclidean and $K > 0$ is a constant, which will be specified later. In view of our assumptions, $\phi$ attains its minimum on $Q$ at some point $x_*$. Therefore, by Theorem 3.1.24, there exists a vector $g_\phi^* \in \partial \phi(x_*)$ such that

$$\langle g_\phi^*, x - x_* \rangle \geq 0, \quad \forall x \in Q. \tag{3.1.72}$$

In view of Lemma 3.1.12, Lemma 3.1.11 and representation (3.1.42), there exist $g^* \in \partial f(x_*)$ and $\bar{y} \in \mathbb{R}^m$ with $\| \bar{y} \| \leq 1$ such that

$$g_\phi^* = g^* - K A^T \bar{y}.$$

Moreover, in view of Lemma 3.1.15, $\langle \bar{y}, b - Ax_* \rangle = \| b - A\bar{x} \|$.

On the other hand, for any $\delta \in B(0, \epsilon)$, we have $x_\delta \overset{\text{def}}{=} \bar{x} + \delta \overset{(3.1.70)}{\in} Q$. Therefore,

$$\begin{aligned} \langle g^*, x_\delta - x_* \rangle & \overset{(3.1.72)}{\geq} K \langle A^T \bar{y}, \bar{x} + \delta - x^* \rangle \; = \; K \langle \bar{y}, A\delta + b - Ax_* \rangle \\ & = \; K \| b - Ax_* \| + K \langle A^T \bar{y}, \delta \rangle. \end{aligned} \tag{3.1.73}$$

In view of Theorem 3.1.11, $M = \max\limits_{x} \{ f(x) : \; x \in B(\bar{x}, \epsilon) \} < +\infty$. Then

$$\langle g^*, x_\delta - x_* \rangle \overset{(3.1.23)}{\leq} f(x_\delta) - f(x_*) \; \leq \; M - f_*,$$

where $f_* = \min\limits_{x \in Q} f(x)$. Therefore, maximizing the right-hand side of inequality (3.1.73) in $\delta \in B(0, \epsilon)$, we get

$$M - f_* \geq K\epsilon \| A^T \bar{y} \| \;\geq K\epsilon\mu \|\bar{y}\|,$$

where $\mu = \lambda_{\min}^{1/2}(AA^T) > 0$. Defining $y^* = K\bar{y}$, we get from the first inequality the bound (3.1.71). On the other hand, choosing $K > \frac{1}{\epsilon\mu}(M - f_*)$, from the second inequality we necessarily get $\|\bar{y}\| < 1$. By Lemma 3.1.15, this implies that $Ax_* = b$. Consequently, $x_*$ is an optimal solution for problem (3.1.68).

As we can see now, for $K$ big enough, any solution $x^*$ of problem (3.1.68) is a global minimum of the function $\phi$. Repeating the above reasoning, we can justify the condition (3.1.70). □

In view of its simplicity, Theorem 3.1.27 has many interesting applications. Here we present only one of them, related to the rules for differentiating a partial minimization of a convex function. The new statement significantly extends a particular case of Theorem 3.1.25.

**Theorem 3.1.28** *Let a function $f$ be convex and $Q$ be a closed convex set belonging to* int (dom $f$). *Assume that the level sets of $f$ are bounded on $Q$.*

*Let a matrix $A \in \mathbb{R}^{m \times n}$ with $n > m$ have a full row rank. Consider the function*

$$\phi(u) = \min_{x \in Q} \{ f(x) : \; Ax = u \}.$$

*Then $\phi$ is convex and for any $u \in \mathbb{R}^m$ such that $\{x \in \operatorname{int}(Q) : \; Ax = u\} \neq \emptyset$, we have*

$$\begin{aligned} &\{y^* : \; \exists x^* \in Q, \; Ax^* = u, \; \textit{and} \; g^* \in \partial f(x^*) \\ &\qquad \text{such that } \langle g^* - A^T y^*, x - x^* \rangle \geq 0 \; \forall x \in Q\} \;\subseteq\; \partial\phi(u). \end{aligned} \tag{3.1.74}$$

*Proof* Let $Q(u) = \{x \in Q : \; Ax = u\}$. Then $\operatorname{dom}\phi = \{u \in \mathbb{R}^m : \; Q(u) \neq \emptyset\}$. In view of the conditions of the theorem, for any $u \in \operatorname{dom}\phi$ there exists at least one point $x(u)$ in the set $\operatorname{Arg} \min\limits_{x \in Q(u)} f(x)$. Let $u_1, u_2 \in \operatorname{dom}\phi$ and $\alpha \in [0, 1]$. Then

$$x_\alpha \overset{\text{def}}{=} \alpha x(u_1) + (1 - \alpha) x(u_2) \; \in Q(\alpha u_1 + (1 - \alpha) u_2).$$

Therefore,

$$\begin{aligned} \phi(\alpha u_1 + (1 - \alpha) u_2) \leq f(x_\alpha) &\overset{(3.1.2)}{\leq} \alpha f(x(u_1)) + (1 - \alpha) f(x(u_2)) \\ &= \alpha\phi(u_1) + (1 - \alpha)\phi(u_2). \end{aligned}$$

Further, in view of Theorem 3.1.27, the set in the left-hand side of inclusion (3.1.74) is nonempty. Let the triple $(x^*, y^*, g^*)$ be an element of this set for some $u = u_1 \in \operatorname{dom} \phi$. Then for another $u_2 \in \operatorname{dom} \phi$ we have

$$\phi(u_2) = f(x(u_2)) \overset{(3.1.23)}{\geq} f(x^*) + \langle g^*, x(u_2) - x^* \rangle \overset{(3.1.74)}{\geq} \langle A^T y^*, x(u_2) - x^* \rangle$$

$$= \phi(u_1) + \langle y^*, u_2 - u_1 \rangle.$$

Therefore, $y^* \overset{(3.1.23)}{\in} \partial \phi(u_1)$. □

Thus, the rules for differentiating the function $\phi$ at a point $u$ are very simple. We need to solve the corresponding minimization problem and extract from the solver the optimal Lagrange multipliers for equality constraints. This vector is an element of the subdifferential $\partial \phi(u)$.

## *3.1.8 Minimax Theorems*

Consider a function $\Psi(\cdot, \cdot)$ defined on the direct product of two convex sets, $P \subseteq \mathbb{R}^n$ and $S \subseteq \mathbb{R}^m$. We assume that the functions $\Psi(\cdot, u)$ are closed and convex on $P \subseteq \operatorname{dom} \Psi(\cdot, u)$ for all $u \in S$. Similarly, all functions $\Psi(x, \cdot)$ are closed and concave on $S \subseteq \operatorname{dom} \Psi(x, \cdot)$ for all $x \in P$. The main goal of this section is the justification of the sufficient conditions for the equality

$$\inf_{x \in P} \sup_{u \in S} \Psi(x, u) = \sup_{u \in S} \inf_{x \in P} \Psi(x, u). \tag{3.1.75}$$

Note that in general, we can guarantee only that the right-hand side of this relation does not exceed its left-hand side (see (1.3.6)).

Define $f(x) = \sup\limits_{u \in S} \Psi(x, u) \geq \phi(u) = \inf\limits_{x \in P} \Psi(x, u)$. We will see that in many situations

$$\min_{x \in P} f(x) = \max_{u \in S} \phi(u).$$

Let us start from a simple observation.

**Lemma 3.1.22** *Assume that for any $u \in S$, the level sets of the function $\Psi(\cdot, u)$ are bounded on $P$, and the function $\phi$ attains its maximum on $S$ at some point $u^*$. Then for any $u \in S$ we have*

$$\min_{x \in P} \max\{\Psi(x, u), \Psi(x, u^*)\} = \phi(u^*). \tag{3.1.76}$$

*Proof* Let us choose an arbitrary $u \in S$. For $x \in P$, consider the function

$$f_u(x) = \max\{\Psi(x,u), \Psi(x,u^*)\} \ \geq \ \max\{\phi(u), \phi(u^*)\} \ = \ \phi(u^*). \qquad (3.1.77)$$

In view of Theorem 3.1.10, there exists a $\lambda^* \in [0,1]$ such that

$$\begin{aligned}\min_{x\in P} f_u(x) &= \min_{x\in P} \left\{\lambda^*\Psi(x,u) + (1-\lambda^*)\Psi(x,u^*)\right\}\\ &\leq \min_{x\in P} \Psi(x, \lambda^* u + (1-\lambda^*)u^*)\\ &= \phi(\lambda^* u + (1-\lambda^*)u^*).\end{aligned}$$

Hence, $\phi(u^*) \overset{(3.1.77)}{\leq} \min\limits_{x\in P} f_u(x) \leq \phi(\lambda u + (1-\lambda)u^*) \leq \phi(u^*)$. □

Now we can prove the first variant of the Minimax Theorem.

**Theorem 3.1.29** *Let each of the functions $\Psi(\cdot,u)$ attain a unique minimum on $P$, and let the function $\phi$ attain its maximum on $S$. Then*

$$\min_{x\in P} \ f(x) = \max_{u\in S} \phi(u). \qquad (3.1.78)$$

*Proof* Since the point $x(u) = \arg\min\limits_{x\in P} \Psi(x,u)$ is uniquely defined, the level sets of all functions $\Psi(\cdot,u)$, $u \in S$ are bounded (see Theorem 3.1.4(5)). Thus, by Lemma 3.1.22, relation (3.1.76) is valid for all $u \in S$.

Since $\phi(u^*) = \Psi(x(u^*), u^*)$, the minimum of problem (3.1.76) can be achieved only at the point $x(u^*)$. But then for any $u \in S$ we have

$$\Psi(x(u^*), u) \ \overset{(3.1.76)}{\leq} \ \Psi(x(u^*), u^*) \ \leq \ \Psi(x, u^*), \quad x \in P.$$

Thus, $f(x(u^*)) \leq \phi(u^*)$, and we get (3.1.78) by (1.3.6). □

Relaxation of the uniqueness condition for the minimizers of the functions $\Psi(\cdot,u)$, $u \in S$, gives us a variant of von Neuman's Theorem.[3]

**Theorem 3.1.30** *Assume that both sets $P$ and $S$ are bounded. Then*

$$\min_{x\in P} \ f(x) = \max_{u\in S} \phi(u). \qquad (3.1.79)$$

[3] As compared with the standard version of this theorem, we replace the continuity assumptions by assumptions on closedness of the epigraphs.

*Proof* Let us fix some $\epsilon > 0$. For the standard Euclidean norm $\|\cdot\|$, consider the function

$$\Psi_\epsilon(x,u) = \Psi(x,u) + \tfrac{1}{2}\epsilon\|x\|^2, \quad x \in P,\ u \in S.$$

Since for each $u \in S$ the function $\Psi_\epsilon(\cdot,u)$ is strongly convex, it attains a unique minimum on $P$. Therefore the function $\phi_\epsilon(u) = \min\limits_{x\in P} \Psi_\epsilon(x,u)$ is well defined, and in view of Theorem 3.1.8, it is concave and closed on $S$. Therefore, by Theorem 3.1.29, there exist points $u_\epsilon^* \in S$ and $x_\epsilon^* = \arg\min\limits_{x\in P} \Psi_\epsilon(x,u_\epsilon^*)$, such that

$$\Psi_\epsilon(x_\epsilon^*,u) \le \Psi_\epsilon(x_\epsilon^*,u_\epsilon^*) \ \le\ \Psi_\epsilon(x,u_\epsilon^*), \quad x \in P,\ u \in S.$$

The first inequality is $\Psi(x_\epsilon^*,u) \le \Psi(x_\epsilon^*,u_\epsilon^*)$ for all $u \in S$. Thus,

$$f(x_\epsilon^*) = \sup_{u\in S} \Psi(x_\epsilon^*,u) \ \le\ \Psi(x_\epsilon^*,u_\epsilon^*).$$

On the other hand, for all $x \in P$ we have

$$\Psi(x_\epsilon^*,u_\epsilon^*) \le \Psi_\epsilon(x_\epsilon^*,u_\epsilon^*) \ \le\ \Psi(x,u_\epsilon^*) + \tfrac{1}{2}\epsilon D^2,$$

where $D \ge \sup\limits_{x\in P}\|x\|$. Hence,

$$f(x_\epsilon^*) \le \phi(u_\epsilon^*) + \tfrac{1}{2}\epsilon D^2, \quad \epsilon > 0.$$

In view of the boundedness of the sets $P$ and $S$, letting $\epsilon \to 0$ in this inequality, we get the relation (3.1.79) (see Item 4 of Theorem 3.1.4). □

Finally, let us show that sometimes it is possible to derive the no-gap property (3.1.78) from the local optimality conditions.

**Theorem 3.1.31** *Let a function $f$ attain its minimum on $P$ at the point $x^*$. Suppose that for some $g_* \in \partial_P f(x^*)$, yielding the first-order optimality condition*

$$\langle g_*, x - x^*\rangle \overset{(3.1.56)}{\ge} 0, \quad x \in P,$$

*there exists a representation*

$$g_* = \sum_{i=1}^{k} \lambda^{(i)} g_i, \tag{3.1.80}$$

*for certain $k \ge 1$, $\lambda \in \Delta_k$, and some $g_i$ belonging to the sets $\partial_{P,x}\Psi(x^*,u_i)$, where $u_i \in I(x^*)$, $i = 1,\dots,k$, and $I(x^*) = \{u \in S :\ \Psi(x^*,u) = f(x^*)\}$. Then the relation (3.1.78) is satisfied.*

*Proof* Indeed, let $\bar{u} = \sum\limits_{i=1}^{k} u_i$. Then, for any $x \in P$, we have

$$\begin{aligned} f(x^*) &\leq f(x^*) + \langle g_*, x - x^* \rangle \overset{(3.1.80)}{=} f(x^*) + \sum_{i=1}^{k} \lambda^{(i)} \langle g_i, x - x^* \rangle \\ &\overset{(3.1.23)}{\leq} f(x^*) + \sum_{i=1}^{k} \lambda^{(i)} [\Psi(x, u_i) - \Psi(x^*, u_i)] = \sum_{i=1}^{k} \lambda^{(i)} \Psi(x, u_i) \\ &\leq \Psi(x, \bar{u}). \end{aligned}$$

Thus, $f(x^*) \leq \phi(\bar{u})$, and by (1.3.6) we see that $\phi(\bar{u}) = \max\limits_{u \in S} \phi(u)$. □

Note that the right-hand side of representation (3.1.80) belongs to $\partial_P f(x^*)$ (see Lemma 3.1.14). Therefore, a sufficient condition for the existence of this representation is

$$\partial_P f(x^*) = \text{Conv}\, \{\partial_{P,x} \Psi(x^*, u) : \; u \in I(x^*)\}. \tag{3.1.81}$$

### 3.1.9 Basic Elements of Primal-Dual Methods

Very often, the possibility of applying *primal-dual* optimization methods comes out from direct access to the internal structure of the objective function. Consider the problem

$$f^* = \min_{x \in P} f(x), \tag{3.1.82}$$

where the function $f$ is closed and convex on $P$. Suppose that the objective function $f$ has a *max-representation*:

$$f(x) = \max_{u \in S} \Psi(x, u), \tag{3.1.83}$$

where the function $\Psi$ satisfies all our assumptions made in the beginning of Sect. 3.1.8. From this representation, we derive the *dual problem*[4]

$$\phi^* = \max_{u \in S} \phi(u), \quad \phi(u) \overset{\text{def}}{=} \min_{x \in P} \Psi(x, u). \tag{3.1.84}$$

[4] In Chap. 6 we call it the *adjoint problem* due to the fact that very often representation (3.1.83) is not unique.

From the mathematical point of view, the pair of primal-dual problems (3.1.82) and (3.1.84) looks completely symmetric. However, this is not true for numerical methods. Indeed, our initial intention was to solve problem (3.1.82). Hence, it is implicitly assumed that the maximization problem in definition (3.1.83) is relatively *easy*. It should be possible to solve it either in a closed form, or by a simple numerical procedure (which defines the *complexity of the oracle*). At the same time, the complexity of computing the value of the objective function in problem (3.1.84) can be very high. It can easily reach the complexity of our initial problem (3.1.82). Therefore, it seems that the dual problem has a good chance of being much more difficult than the initial primal problem (3.1.82).

Fortunately this is not the case provided that we have an access to the internal structure of the oracle (3.1.83). Indeed, in order to compute the value $f(x)$ the oracle needs to compute a point

$$u(x) \in \operatorname{Arg}\max_{u \in S} \Psi(x, u).$$

Let us assume that this point is used to compute the subgradient $g(x)$ (or, when $f$ is smooth, the gradient) of the objective function (see Lemma 3.1.14):

$$g(x) \in \partial_{P,x} \Psi(x, u(x)).$$

Thus, we assume that the oracle returns three objects: $f(x)$, $g(x)$, and $u(x) \in S$. Let us show how this information can be used in numerical methods.

In Smooth Optimization, we often use the *functional model* of the objective function. Assume that some method accumulated the information from the oracle at points $\{y_k\}_{k=0}^N \subset P$. Then, for some scaling coefficients

$$\alpha_k > 0, \quad k = 0, \ldots, N, \quad \sum_{k=0}^{N} \alpha_k \ = \ 1,$$

we can construct a *linear model* of the objective function:

$$\ell_N(x) = \sum_{k=0}^{N} \alpha_k [f(y_k) + \langle g(y_k), x - y_k \rangle] \overset{(3.1.23)}{\leq} f(x), \quad x \in P.$$

In some methods (see, for example, (2.2.3), (2.2.4)), for points of minimizing sequence $\{x_k\}_{k \geq 0}$, it is possible to ensure the following relation:

$$f(x_N) \leq \min_{x \in P} \ \ell_N(x) + r_N, \tag{3.1.85}$$

where $r_N \to 0$ as $N \to \infty$. In fact, this relation can be used not only for justifying the quality of point $x_N$, but also for estimating the primal-dual gap with respect to the dual solution

$$\hat{u}_N = \sum_{k=0}^{N} \alpha_k u(y_k) \ \in \ S. \tag{3.1.86}$$

**Lemma 3.1.23** *Let the point $x_N$ satisfy (3.1.85). Then*

$$0 \leq (f(x_N) - f^*) + (\phi^* - \phi(\hat{u}_N)) \ \leq \ f(x_N) - \phi(\hat{u}_N) \ \leq \ r_N.$$

*Proof* Indeed, $g(y_k) \in \partial_{P,x} \Psi(y_k, u(y_k))$. Therefore,

$$\begin{aligned} \min_{x \in P} \ \ell_N(x) \quad &= \quad \min_{x \in P} \sum_{k=0}^{N} \alpha_k [\Psi(y_k, u(y_k)) + \langle g(y_k), x - y_k \rangle] \\ &\overset{(3.1.23)}{\leq} \quad \min_{x \in P} \sum_{k=0}^{N} \alpha_k \Psi(x, u(y_k)) \ \leq \ \min_{x \in P} \Psi(x, \hat{u}_N) \ = \ \phi(\hat{u}_N). \end{aligned}$$

It remains to use inequality (3.1.85). □

Since we have ensured $r_N \to 0$, for our problem we have managed to prove the no-gap property *algorithmically*. Note that our way of generating the good dual solution (3.1.86) does not require a single computation of the dual function.

In Nonsmooth Optimization, we use another certificate of optimality based on the *gap function*. It is defined by a sequence of test points $\{y_k\}_{k=0}^{N}$ and scaling coefficients as follows:

$$\delta_N(x) = \sum_{k=0}^{N} \alpha_k \langle g(y_k), y_k - x \rangle.$$

Define $\hat{f}_N = \sum\limits_{k=0}^{N} \alpha_k f(y_k)$.

**Lemma 3.1.24** *Assume that* $\max\limits_{x \in P} \delta_N(x) \leq r_N \to 0$. *Then*

$$0 \leq (\hat{f}_N - f^*) + (\phi^* - \phi(\hat{u}_N)) \ \leq \ \hat{f}_N - \phi(\hat{u}_N) \ \leq \ r_N \to 0.$$

*Proof* Indeed

$$\begin{aligned}
\max_{x\in P} \delta_N(x) &= \max_{x\in P} \sum_{k=0}^{N} \alpha_k \langle g(y_k), y_k - x\rangle \\
&\overset{(3.1.23)}{\geq} \min_{x\in P} \sum_{k=0}^{N} \alpha_k [\Psi(y_k, u(y_k)) - \Psi(x, u(y_k))] \\
&\overset{(3.1.4)}{\geq} \hat{f}_N - \min_{x\in P} \Psi(x, \hat{u}_N) = \hat{f}_N - \phi(\hat{u}_N). \qquad \square
\end{aligned}$$

Again, for nonsmooth problems, computation of the good dual solution $\hat{u}_N$ does not require significant computational resources.

## 3.2 Methods of Nonsmooth Minimization

(General lower complexity bounds; Main lemma; Localization sets; The subgradient method; Minimization with functional constraints; Approximation of optimal Lagrange multipliers; Strongly convex functions; Optimization in finite dimensions and lower complexity bounds; Cutting plane schemes; The center of gravity method; The ellipsoid method and others.)

### *3.2.1 General Lower Complexity Bounds*

In Sect. 3.1, we introduced a class of general convex functions. These functions can be nonsmooth and therefore the corresponding minimization problem can be quite difficult. As for smooth problems, let us try to derive lower complexity bounds, which will help us to evaluate the performance of numerical methods.

In this section, we derive such bounds for the following unconstrained minimization problem

$$\min_{x\in\mathbb{R}^n} f(x), \tag{3.2.1}$$

where $f$ is a convex function. Denote by $x^* \in \mathbb{R}^n$ one of its optimal solutions. Thus, our problem class is as follows.

| | |
|---|---|
| **Model**: | 1. Unconstrained minimization.<br>2. $f$ is convex on $\mathbb{R}^n$ and Lipschitz continuous on a bounded set. |
| **Oracle**: | First-order Black Box:<br>at each point $\hat{x}$, we can compute<br>$f(\hat{x}), \quad g(\hat{x}) \in \partial f(\hat{x})$,<br>$g(\hat{x})$ is an *arbitrary* subgradient. |
| **Approximate solution**: | Find $\bar{x} \in \mathbb{R}^n : \ f(\bar{x}) - f^* \leq \epsilon$. |
| **Methods**: | Generate a sequence $\{x_k\}$ :<br>$x_k \in x_0 + \mathrm{Lin}\,\{g(x_0), \dots, g(x_{k-1})\}$. |

(3.2.2)

As in Sect. 2.1.2, to derive lower complexity bounds for our problem class, we will study the behavior of numerical methods on some function, which appears to be very difficult for all schemes.

Let us fix some parameters $\mu > 0$ and $\gamma > 0$. Consider the family of functions

$$f_k(x) = \gamma \max_{1 \leq i \leq k} x^{(i)} + \tfrac{\mu}{2} \parallel x \parallel^2, \quad k = 1 \dots n, \tag{3.2.3}$$

where the norm is standard Euclidean. Using the rules of subdifferential calculus, described in Sect. 3.1.6, we can write down a closed-form expression for the subdifferential of $f_k$ at $x$. This is

$$\partial f_k(x) = \mu x + \gamma \, \mathrm{Conv}\,\{e_i \mid i \in I(x)\},$$

$$I(x) = \{j \mid \ 1 \leq j \leq k, \ x^{(j)} = \max_{1 \leq i \leq k} x^{(i)}\}.$$

Let $x_k^*$ be the global minimum of the function $f_k$. Then, for any $x, y \in B_2(x^*, \rho)$, $\rho > 0$, and $g_k(y) \in \partial f_k(y)$, we have

$$\begin{aligned} f_k(y) - f_k(x) &\leq \langle g_k(y), y - x \rangle \ \leq \parallel g_k(y) \parallel \cdot \parallel y - x \parallel \\ &\leq \left(\mu \| x_k^* \| + \mu\rho + \gamma\right) \parallel y - x \parallel . \end{aligned} \tag{3.2.4}$$

Thus, $f_k$ is Lipschitz continuous on $B_2(x_k^*, \rho)$ with Lipschitz constant

$$M = \mu \|x_k^*\| + \mu\rho + \gamma.$$

Further, by Theorem 3.1.20, it is easy to check that the optimal point $x_k^*$ has the following coordinates:

$$(x_k^*)^{(i)} = \begin{cases} -\frac{\gamma}{\mu k}, \; 1 \le i \le k, \\ \\ \quad 0, \; k+1 \le i \le n. \end{cases}$$

Now we have all the important characteristics of our problem:

$$\begin{aligned} R_k &\overset{\text{def}}{=} \| \, x_k^* \, \| = \tfrac{\gamma}{\mu\sqrt{k}}, \quad f_k^* \;=\; -\tfrac{\gamma^2}{\mu k} + \tfrac{\mu}{2} R_k^2 \;=\; -\tfrac{\gamma^2}{2\mu k}, \\ M &= \mu \|x_k^*\| + \mu\rho + \gamma \;=\; \mu\rho + \gamma \, \tfrac{\sqrt{k}+1}{\sqrt{k}}. \end{aligned} \tag{3.2.5}$$

Let us describe now a resisting oracle for the function $f_k(\cdot)$. Since the analytical form of this function is fixed, the resistance of this oracle consists in providing us with the worst possible subgradient at each test point. The algorithmic scheme of this oracle is as follows.

```
Input:      $x \in \mathbb{R}^n$.

MainLoop:   $f := -\infty; \quad i^* := 0;$

            for $j := 1$ to $k$ do
            if $x^{(j)} > f$ then { $f := x^{(j)}; \quad i^* := j$ };

            $f := \gamma f + \frac{\mu}{2} \| \, x \, \|^2; \quad g := \gamma e_{i^*} + \mu x;$

Output :    $f_k(x) := f, \quad g_k(x) := g \in \mathbb{R}^n$.
```
(3.2.6)

At first glance, there is nothing special in this procedure. Its main loop is just a standard process for finding the maximal coordinate of a vector from $\mathbb{R}^k$. However, the main feature of this loop is that we always form the subgradient of the nonsmooth part of the objective proportional to a coordinate vector. Moreover, the active coordinate $i^*$ always corresponds to the *first* maximal component of vector $x$. Let us see what happens with a minimizing sequence based on such an oracle.

Let us choose the starting point $x_0 = 0$. Define

$$\mathbb{R}^{p,n} = \{x \in \mathbb{R}^n \mid x^{(i)} = 0,\ p+1 \le i \le n\}.$$

Since $x_0 = 0$, the answer of the oracle is $f_k(x_0) = 0$ and $g_k(x_0) = e_1$. Therefore, the next point of the sequence, $x_1$, necessarily belongs to $\mathbb{R}^{1,n}$. Assume now that the current test point of the sequence, $x_i$, belongs to $\mathbb{R}^{p,n}$, $1 \le p \le k$. Then the oracle returns a subgradient

$$g = \mu x_i + \gamma e_{i^*},$$

where $i^* \le p+1$. Therefore, the next test point $x_{i+1}$ belongs to $\mathbb{R}^{p+1,n}$.

This simple reasoning proves that for all $i$, $1 \le i \le k$, we have $x_i \in \mathbb{R}^{i,n}$. Consequently, for $i$: $1 \le i \le k-1$, we cannot improve the starting value of the objective function:

$$f_k(x_i) \ge \gamma \max_{1 \le j \le k} x_i^{(j)} = 0.$$

Let us convert this observation into a lower complexity bound. Let us fix some parameters of our problem class $\mathscr{P}(x_0, R, M)$, that is, $R > 0$ and $M > 0$. In addition to (3.2.2) we assume the following.

- The point $x_0$ is close enough to the solution of problem (3.2.1):

$$\|x_0 - x^*\| \le R. \tag{3.2.7}$$

- The function $f$ is Lipschitz continuous on $B_2(x^*, R)$ with constant $M > 0$.

**Theorem 3.2.1** *For any class* $\mathscr{P}(x_0, R, M)$ *and any* $k$, $0 \le k \le n-1$, *there exists a function* $f \in \mathscr{P}(x_0, R, M)$ *such that*

$$f(x_k) - f^* \ge \tfrac{MR}{2(2+\sqrt{k+1})}$$

*for any optimization scheme, which generates a sequence* $\{x_k\}$ *satisfying the condition*

$$x_k \in x_0 + \text{Lin}\,\{g(x_0), \ldots, g(x_{k-1})\}.$$

*Proof* Without loss of generality, we can assume that $x_0 = 0$. Let us choose $f(x) = f_{k+1}(x)$ with the following values of parameters:

$$\gamma = \frac{\sqrt{k+1}M}{2+\sqrt{k+1}}, \quad \mu = \frac{M}{(2+\sqrt{k+1})R}.$$

Then

$$f^* = f^*_{k+1} \overset{(3.2.5)}{=} -\frac{\gamma^2}{2\mu(k+1)} = -\frac{MR}{2(2+\sqrt{k+1})},$$

$$\| x_0 - x^* \| = R_{k+1} \overset{(3.2.5)}{=} \frac{\gamma}{\mu\sqrt{k+1}} = R.$$

Moreover, $f$ is Lipschitz continuous on $B_2(x^*, R)$ with constant $\mu R + \gamma \frac{\sqrt{k+1}+1}{\sqrt{k+1}} = M$. Note that $x_k \in \mathbb{R}^{k,n}$. Hence, $f(x_k) - f^* \geq -f^*$. □

The lower complexity bound presented in Theorem 3.2.1 does not depend on the dimension of the space of variables. As for the lower bound of Theorem 2.1.7, it can be applied to problems with very large dimension, or to the efficiency analysis of starting iterations of a minimization scheme ($k \leq n-1$).

We will see that our lower estimate is exact: There exist minimization methods which have a rate of convergence proportional to this lower bound. Comparing this bound with the lower bound for smooth minimization problems, we can see that now the possible convergence rate is much slower. However, we should remember that we are working with one of the most general classes of convex problems.

### 3.2.2 Estimating Quality of Approximate Solutions

We are now interested in the following optimization problem:

$$\min_{x \in Q} f(x), \tag{3.2.8}$$

where $Q$ is a closed convex set, and the function $f$ is convex on $\mathbb{R}^n$. We are going to study numerical methods for solving (3.2.8), which employ subgradients $g(x)$ of the objective function, computed at $x \in Q$. As compared with the smooth problem, our goal is more challenging. Indeed, even in the simplest situation, when $Q \equiv \mathbb{R}^n$, the subgradient seems to be a poor replacement for the gradient of a smooth function. For example, we cannot be sure that the value of the objective function is decreasing in the direction $-g(x)$. We cannot expect that $g(x) \to 0$ as $x$ approaches the solution of our problem, etc.

Fortunately, there is one property of subgradients which makes our goals reachable. We have justified this property in Corollary 3.1.6:

$$\boxed{\begin{array}{c}\text{At any } x \in Q, \text{ the following inequality holds:}\\[1em] \langle g(x), x - x^* \rangle \geq 0.\end{array}} \qquad (3.2.9)$$

This simple inequality leads to two important consequences, which form the basis for the majority of nonsmooth minimization methods. Namely:

- The distance between $x$ and $x^*$ decreases along the direction $-g(x)$.
- Inequality (3.2.9) cuts $\mathbb{R}^n$ in two half-spaces, and it is known which of them contains the optimal point $x^*$.

Nonsmooth minimization methods cannot employ the idea of relaxation or approximation. There is another concept underlying all these schemes. This is the concept of *localization*. However, to go forward with this concept, we have to develop a special technique which allows us to estimate the quality of an approximate solution to problem (3.2.8). This is the main goal of this section.

Let us fix some $\bar{x} \in \mathbb{R}^n$. For $x \in \mathbb{R}^n$ with $g(x) \neq 0$ define

$$v_f(\bar{x}, x) = \tfrac{1}{\|g(x)\|} \langle g(x), x - \bar{x} \rangle. \qquad (3.2.10)$$

If $g(x) = 0$, then define $v_f(\bar{x}; x) = 0$. Clearly, by the Cauchy-Schwarz inequality,

$$v_f(\bar{x}, x) \leq \| x - \bar{x} \| .$$

The values $v_f(\bar{x}, x)$ have a natural geometric interpretation. Consider a point $x$ such that $g(x) \neq 0$ and $\langle g(x), x - \bar{x} \rangle \geq 0$. Let us look at the point

$$\bar{y} = \bar{x} + v_f(\bar{x}, x) \tfrac{g(x)}{\|g(x)\|}.$$

Then

$$\langle g(x), x - \bar{y} \rangle = \langle g(x), x - \bar{x} \rangle - v_f(\bar{x}, x) \parallel g(x) \parallel \overset{(3.2.10)}{=} 0,$$

and $\| \bar{y} - \bar{x} \| = v_f(\bar{x}, x)$. Thus, $v_f(\bar{x}, x)$ is a *distance* from point $\bar{x}$ to the hyperplane $\{y : \langle g(x), x - y \rangle = 0\}$.

Let us introduce a function which measures the growth of the function $f$ around the point $\bar{x}$. For $t \geq 0$, define

$$\omega_f(\bar{x}; t) = \max_x \{ f(x) - f(\bar{x}) : \| x - \bar{x} \| \leq t \}.$$

If $t < 0$, we set $\omega_f(\bar{x}; t) = 0$.

Clearly function $\omega_f$ possesses the following properties.

- $\omega_f(\bar{x}; t) = 0$ for all $t \leq 0$.
- $\omega_f(\bar{x}; t)$ is a nondecreasing function of $t \in \mathbb{R}$.
- $f(x) - f(\bar{x}) \leq \omega_f(\bar{x}; \| x - \bar{x} \|)$.

It is important that under a convexity assumption the last inequality can be significantly strengthened.

**Lemma 3.2.1** *For any $x \in \mathbb{R}^n$ we have*

$$f(x) - f(\bar{x}) \leq \omega_f(\bar{x}; v_f(\bar{x}; x)). \tag{3.2.11}$$

*If $f(\cdot)$ is Lipschitz continuous on $B_2(\bar{x}, R)$ with constant $M$, then*

$$f(x) - f(\bar{x}) \leq M(v_f(\bar{x}; x))_+ \tag{3.2.12}$$

*for all $x \in \mathbb{R}^n$ with $v_f(\bar{x}; x) \leq R$.*

*Proof* If $\langle g(x), x - \bar{x} \rangle < 0$, then $f(\bar{x}) \geq f(x) + \langle g(x), \bar{x} - x \rangle \geq f(x)$. Since $v_f(\bar{x}; x)$ is negative, we have $\omega_f(\bar{x}; v_f(\bar{x}; x)) = 0$ and (3.2.11) holds.

Let $\langle g(x), x - \bar{x} \rangle \geq 0$. For

$$\bar{y} = \bar{x} + v_f(\bar{x}; x) \tfrac{g(x)}{\|g(x)\|},$$

we have $\langle g(x), \bar{y} - x \rangle = 0$ and $\| \bar{y} - \bar{x} \| = v_f(\bar{x}; x)$. Therefore,

$$f(\bar{y}) \geq f(x) + \langle g(x), \bar{y} - x \rangle \; = \; f(x),$$

and

$$f(x) - f(\bar{x}) \leq f(\bar{y}) - f(\bar{x}) \; \leq \; \omega_f(\bar{x}; \| \bar{y} - \bar{x} \|) \; = \; \omega_f(\bar{x}; v_f(\bar{x}; x)).$$

If $f$ is Lipschitz continuous on $B_2(\bar{x}, R)$ and $0 \leq v_f(\bar{x}; x) \leq R$, then $\bar{y} \in B_2(\bar{x}, R)$. Hence,

$$f(x) - f(\bar{x}) \leq f(\bar{y}) - f(\bar{x}) \; \leq \; M \| \bar{y} - \bar{x} \| = \; M v_f(\bar{x}; x). \qquad \square$$

Let us fix some optimal solution $x^*$ of problem (3.2.8). The values $v_f(x^*; x)$ allow us to estimate the quality of so-called *localization sets*.

**Definition 3.2.1** Let $\{x_i\}_{i=0}^{\infty}$ be a sequence in $Q$. Define

$$S_k = \{x \in Q \mid \; \langle g(x_i), x_i - x \rangle \geq 0, \; i = 0 \dots k\}.$$

We call $S_k$ the *localization set* of problem (3.2.8) generated by the sequence $\{x_i\}_{i=0}^{\infty}$.

In view of inequality (3.2.9), for all $k \geq 0$, we have $x^* \in S_k$.

Let

$$v_i = v_f(x^*; x_i) \ (\geq 0), \quad v_k^* = \min_{0 \leq i \leq k} v_i.$$

Thus,

$$v_k^* = \max\{r : \langle g(x_i), x_i - x \rangle \geq 0, \ i = 0 \ldots k, \ \forall x \in B_2(x^*, r)\}.$$

This is the radius of the maximal ball centered at $x^*$, which is contained in the localization set $S_k$.

**Lemma 3.2.2** *Let* $f_k^* = \min\limits_{0 \leq i \leq k} f(x_i)$. *Then*

$$f_k^* - f^* \leq \omega_f(x^*; v_k^*).$$

*Proof* Using Lemma 3.2.1, we have

$$\omega_f(x^*; v_k^*) = \min_{0 \leq i \leq k} \omega_f(x^*; v_i) \ \geq \ \min_{0 \leq i \leq k} [f(x_i) - f^*] \ = \ f_k^* - f^*. \qquad \square$$

### 3.2.3 The Subgradient Method

Now we are ready to analyze the behavior of some minimization methods. Consider the problem

$$\min_{x \in Q} f(x), \tag{3.2.13}$$

where the function $f$ is convex on $\mathbb{R}^n$, and $Q$ is a *simple* closed convex set. The term "simple" means that we can solve *explicitly* some simple minimization problems over $Q$. In this section, we need to find in a reasonably cheap way the Euclidean projection of any point onto the set $Q$.

We assume that problem (3.2.13) is equipped with a first-order oracle, which at any test point $\bar{x}$ provides us with the value of the objective function $f(\bar{x})$ and one of its subgradients $g(\bar{x})$.

As usual, we first try a version of the Gradient Method. Note that for nonsmooth problems the norm of the subgradient, $\| \ g(x) \ \|$, is not very informative. Therefore, in the subgradient scheme we use a *normalized* direction $\frac{g(\bar{x})}{\|g(\bar{x})\|}$.

**Subgradient Method for Simple Sets**

**0.** Choose $x_0 \in Q$ and a sequence $\{h_k\}_{k=0}^{\infty}$:

$$h_k > 0, \quad h_k \to 0, \quad \sum_{k=0}^{\infty} h_k = \infty. \tag{3.2.14}$$

**1.** ***k*****th iteration (*k* ≥ 0).**
Compute $f(x_k)$, $g(x_k)$ and set $x_{k+1} = \pi_Q \left( x_k - h_k \frac{g(x_k)}{\| g(x_k) \|} \right)$.

Let us estimate the rate of convergence of this scheme.

**Theorem 3.2.2** *Let a function $f$ be Lipschitz continuous on $B_2(x^*, R)$ with constant $M$, where $R \geq \| x_0 - x^* \|$. Then*

$$f_k^* - f^* \leq M \frac{R^2 + \sum\limits_{i=0}^{k} h_i^2}{2 \sum\limits_{i=0}^{k} h_i}. \tag{3.2.15}$$

*Proof* Let $r_i = \| x_i - x^* \|$. Then, in view of Lemma 2.2.8, we have

$$\begin{aligned} r_{i+1}^2 &= \left\| \pi_Q \left( x_i - h_i \frac{g(x_i)}{\| g(x_i) \|} \right) - x^* \right\|^2 \\ &\leq \left\| x_i - h_i \frac{g(x_i)}{\| g(x_i) \|} - x^* \right\|^2 = r_i^2 - 2 h_i v_i + h_i^2. \end{aligned}$$

Summing up these inequalities for $i = 0 \dots k$, we get

$$r_0^2 + \sum_{i=0}^{k} h_i^2 \geq 2 \sum_{i=0}^{k} h_i v_i + r_{k+1}^2 \geq 2 v_k^* \sum_{i=0}^{k} h_i.$$

Thus,

$$v_k^* \leq \frac{R^2 + \sum\limits_{i=0}^{k} h_i^2}{2 \sum\limits_{i=0}^{k} h_i}.$$

Since $v_k^* \leq v_0 \leq \| x_0 - x^* \| \leq R$, we can use Lemma 3.2.2. □

Thus, by Theorem 3.2.2, the rate of convergence of the *Subgradient Method* (3.2.14) depends on the values

$$\Delta_k = \frac{R^2 + \sum\limits_{i=0}^{k} h_i^2}{2\sum\limits_{i=0}^{k} h_i}.$$

We can easily see that $\Delta_k \to 0$ if the series $\sum\limits_{i=0}^{\infty} h_i$ diverges. However, let us try to choose $h_k$ in an optimal way.

Let us assume that we have to perform a fixed number of steps, say $N \geq 1$, of the Subgradient Method. Then, minimizing $\Delta_k$ as a function of $\{h_k\}_{k=0}^{N}$, we can see that the optimal strategy is as follows[5]:

$$h_i = \frac{R}{\sqrt{N+1}}, \quad i = 0 \ldots N. \tag{3.2.16}$$

In this case, $\Delta_N = \frac{R}{\sqrt{N+1}}$ and we obtain the following rate of convergence:

$$f_N^* - f^* \leq \frac{MR}{\sqrt{N+1}}. \tag{3.2.17}$$

Another possibility for defining the step sizes in the Subgradient Method (3.2.14) consists in using the final accuracy $\epsilon > 0$ as a parameter of the algorithm. Indeed, let us find $N$ from the equation

$$\frac{MR}{\sqrt{N+1}} \overset{(3.2.17)}{=} \epsilon \quad \Rightarrow \quad N+1 = \frac{M^2R^2}{\epsilon^2}. \tag{3.2.18}$$

Then, in accordance with (3.2.16), we have

$$h_i = \frac{\epsilon}{M}, \quad i \geq 0. \tag{3.2.19}$$

In view of the upper bound (3.2.15), in this case we have

$$f_N^* - f^* \leq \frac{MR^2}{2\epsilon N} + \frac{1}{2}\epsilon. \tag{3.2.20}$$

Thus, we get an $\epsilon$-solution of the problem (3.2.1) as far as

$$N \geq \frac{M^2R^2}{\epsilon^2}. \tag{3.2.21}$$

[5] From Example 3.1.2(5), we can see that $\Delta_k$ is a symmetric convex function of $\{h_i\}$. Therefore, its minimum is achieved at the point having same values for all variables.

The main advantage of the step size rule (3.2.19) consists in its independence of the parameters $R$ and $N$, which usually are not known in advance. Parameter $M$ is an upper bound on the norm of subgradients of the objective function, which are easily observable during the minimization process.

Comparing inequality (3.2.17) with the lower bound of Theorem 3.2.1, we come to the following conclusion.

Subgradient Method (3.2.14), (3.2.16) is optimal for the problem (3.2.13) uniformly in the number of variables $n$.

If we are not going to fix the number of iterations in advance, we can choose

$$h_i = \tfrac{r}{\sqrt{i+1}}, \quad i = 0, \ldots .$$

Then it is easy to see that $\Delta_k$ is proportional to

$$\tfrac{R^2 + r^2 \ln(k+1)}{4r\sqrt{k+1}},$$

and we can classify this rate of convergence as *sub-optimal*.

Thus, the simplest method for solving problem (3.2.8) appears to be optimal. Usually, this indicates that the problems from our class cannot be solved very efficiently. However, we should remember that our conclusion is valid *uniformly* in the dimension of the problem. We will see later that a moderate dimension of the problem, taken into account in a proper way, helps in developing much faster schemes.

### 3.2.4 Minimization with Functional Constraints

Let us show how we can use the Subgradient Method to solve minimization problems with functional constraints. Consider the problem

$$\min_{x \in Q} \{ f(x) : \; f_j(x) \leq 0, \; j = 1, \ldots, m \}, \tag{3.2.22}$$

with closed and convex functions $f$ and $f_j$, and a simple closed convex set $Q$.

Let us form an aggregate constraint $\bar{f}(x) = \max\limits_{1 \leq j \leq m} f_j(x)$. Then our problem can be written in the following way:

$$\min_{x \in Q} \{ f(x) : \; \bar{f}(x) \leq 0 \}. \tag{3.2.23}$$

Note that we can easily compute a subgradient $\bar{g}(x)$ of the function $\bar{f}$, provided that we can do so for the functions $f_j$ (see Lemma 3.1.13).

Let us fix some $x^*$, an optimal solution to problem (3.2.22). Let $\epsilon > 0$ be the desired accuracy of the approximate solution of problem (3.2.22). Consider the following method.

**Subgradient Method with Functional Constraints**

```
0. Choose a starting point $x_0 \in Q$.
1. kth iteration ($k \geq 0$).
    (a) Compute $f(x_k)$ with $g(x_k) \in \partial f(x_k)$, and $\bar{f}(x_k)$ with
        $\bar{g}(x_k) \in \partial \bar{f}(x_k)$.
    (b) If $\bar{f}(x_k) \leq \epsilon$, then set

        $x_{k+1} = \pi_Q \left( x_k - \frac{\epsilon}{\|g(x_k)\|^2} g(x_k) \right)$.   (Case A)

    Else, set $x_{k+1} = \pi_Q \left( x_k - \frac{\bar{f}(x_k)}{\|\bar{g}(x_k)\|^2} \bar{g}(x_k) \right)$.   (Case B)
```

(3.2.24)

For method (3.2.24), denote by $\mathscr{I}_A(N)$ the set of iterations of type $A$, and by $\mathscr{I}_B(N)$ the set of iterations of type $B$, which occurred during the first $N$ steps of this scheme. Clearly,

$$\bar{f}(x_k) \leq \epsilon, \quad \forall k \in \mathscr{I}_A(N). \tag{3.2.25}$$

**Theorem 3.2.3** *Let functions $f$ and $f_j$, $j = 1, \ldots, m$, be Lipschitz continuous on the ball $B_2(x^*, \|x_0 - x^*\|)$ with constant $M$. If the number of steps $N$ in method (3.2.24) is big enough,*

$$N \geq \frac{M^2}{\epsilon^2} \|x_0 - x^*\|^2, \tag{3.2.26}$$

*then $\mathscr{F}_A(N) \neq \emptyset$ and*

$$f_N^* \overset{\text{def}}{=} \min_k \{ f(x_k) : \; k \in \mathscr{I}_A(N) \} \leq f(x^*) + \epsilon. \tag{3.2.27}$$

*Proof* Define $r_k = \|x_k - x^*\|$. Let us assume that $N$ satisfies (3.2.26), but

$$f(x_k) - f^* \geq \epsilon, \quad \forall k \in \mathscr{I}_A(N). \tag{3.2.28}$$

If $k \in \mathscr{I}_A(N)$, then

$$r_{k+1}^2 \overset{(2.2.49)}{\leq} \left\| x_k - \tfrac{\epsilon}{\|g(x_k)\|^2} g(x_k) \right\|^2 = r_k^2 - \tfrac{2\epsilon}{\|g(x_k)\|^2} \langle g(x_k), x_k - x^* \rangle + \tfrac{\epsilon^2}{\|g(x_k)\|^2}$$

$$\overset{(3.1.23)}{\leq} r_k^2 - \tfrac{2\epsilon}{\|g(x_k)\|^2}(f(x_k) - f^*) + \tfrac{\epsilon^2}{\|g(x_k)\|^2} \overset{(3.2.28)}{\leq} r_k^2 - \tfrac{\epsilon^2}{\|g(x_k)\|^2}.$$

In Case B, we have

$$r_{k+1}^2 \overset{(2.2.49)}{\leq} \left\| x_k - \tfrac{\bar{f}(x_k)}{\|\bar{g}(x_k)\|^2} \bar{g}(x_k) \right\|^2 = r_k^2 - \tfrac{2\bar{f}(x_k)}{\|\bar{g}(x_k)\|^2} \langle \bar{g}(x_k), x_k - x^* \rangle + \tfrac{\bar{f}(x_k)^2}{\|\bar{g}(x_k)\|^2}$$

$$\overset{(3.1.23)}{\leq} r_k^2 - \tfrac{\bar{f}(x_k)^2}{\|\bar{g}(x_k)\|^2} \overset{(3.2.28)}{\leq} r_k^2 - \tfrac{\epsilon^2}{\|\bar{g}(x_k)\|^2}.$$

Thus, in both cases, $r_{k+1} < r_k \leq \|x_0 - x^*\|$. Hence,

$$\|g(x_k)\| \leq M, \ k \in \mathscr{I}_A(N), \quad \|\bar{g}(x_k)\| \leq M, \ k \in \mathscr{I}_B(N).$$

Therefore, $r_{k+1}^2 \leq r_k^2 - \frac{\epsilon^2}{M^2}$ for any $k = 0, \ldots, N$. Summing up these inequalities, we get the inequality

$$0 \leq r_{N+1}^2 \leq r_0^2 - \tfrac{\epsilon^2}{M^2}(N+1),$$

which contradicts our assumption (3.2.26). □

Comparing the bound (3.2.26) with the result of Theorem 3.2.1, we see that the scheme (3.2.24) has an optimal worst-case performance guarantee. Recall, that the same lower complexity bound was obtained for an unconstrained minimization problem. Thus, we can see that, from the viewpoint of analytical complexity, Convex Unconstrained Minimization is not easier than Constrained Minimization.

### 3.2.5 Approximating the Optimal Lagrange Multipliers

Let us show now that a simple subgradient switching strategy can be used for approximating the optimal Lagrange multipliers of problem (3.2.22) (see Theorem 3.1.26).

For $\epsilon > 0$, denote by

$$\mathscr{F}(\epsilon) = \{x \in Q : \ f_j(x) \leq \epsilon, \ j = 1, \ldots, m\}$$

the extended feasible set of problem (3.2.22). Defining the Lagrangian

$$\mathscr{L}(x,\lambda) = f(x) + \sum_{j=1}^{m} \lambda^{(j)} f_j(x), \quad x \in Q, \quad \lambda = (\lambda^{(1)}, \dots, \lambda^{(m)}) \in \mathbb{R}^m_+,$$

we can introduce the *Lagrangian dual problem*

$$\phi^* \overset{\text{def}}{=} \sup_{\lambda \in \mathbb{R}^m_+} \phi(\lambda), \tag{3.2.29}$$

where $\phi(\lambda) \overset{\text{def}}{=} \min_{x \in Q} \mathscr{L}(x,\lambda)$. Clearly, $f^* \overset{(1.3.6)}{\geq} \phi^*$.

In order to approach an optimal solution of problems (3.2.22), (3.2.29), we apply the following switching strategy. It has only one input parameter, the step size $h > 0$. In what follows, we use the notation $\| \cdot \|$ for the standard Euclidean norm, $g(\cdot)$ denotes the subgradient of the objective function, and $g_j(\cdot)$ denotes the subgradient of the corresponding constraints.

```
Subgradient Method for Lagrange Multipliers

0. Choose a starting point $x_0 \in Q$.
1. $k$th iteration ($k \geq 0$).                                          (3.2.30)
    (a) Define $\mathscr{I}_k = \{j : f_j(x_k) > h\|g_j(x_k)\|\}$.
    (b) If $\mathscr{I}_k = \emptyset$, then compute $x_{k+1} = \pi_Q\left(x_k - \frac{hg(x_k)}{\|g(x_k)\|}\right)$.
    (c) If $\mathscr{I}_k \neq \emptyset$, then choose arbitrary $j_k \in \mathscr{I}_k$ and define
    $h_k = \frac{f_{j_k}(x_k)}{\|g_{j_k}(x_k)\|^2}$. Compute $x_{k+1} = \pi_Q(x_k - h_k g_{j_k}(x_k))$.
```

After $t \geq 0$ iterations, define $\mathscr{A}_0(t) = \{k \in \{0, \dots, t\} : \mathscr{I}_k = \emptyset\}$ and let

$$\mathscr{A}_j(t) = \{k \in \{0, \dots, t\} : j_k = j\}, \quad 1 \leq j \leq m.$$

Let $N(t) = |\mathscr{A}_0(t)|$. It is possible that $N(t) = 0$. However, if $N(t) > 0$, then we can define the approximate dual multipliers as follows:

$$\sigma_t = h \sum_{k \in \mathscr{A}_0(t)} \frac{1}{\|g(x_k)\|}, \quad \lambda_t^{(j)} = \frac{1}{\sigma_t} \sum_{k \in \mathscr{A}_j(t)} h_k, \quad j = 1, \dots, m. \tag{3.2.31}$$

Let $S_t = \sum_{k \in \mathscr{A}_0(t)} \frac{1}{\|g(x_k)\|}$. If $\mathscr{A}_0(t) = \emptyset$, then we define $S_t = 0$. Thus, $\sigma_t = hS_t$.

For proving convergence of the switching strategy (3.2.30), we are going to find an upper bound for the gap

$$\delta_t = \frac{1}{S_t} \sum_{k \in \mathscr{A}_0(t)} \frac{f(x_k)}{\|g(x_k)\|} - \phi(\lambda_t),$$

assuming that $N(t) > 0$. Here and in the sequel $\lambda_t$ denotes $(\lambda_t^{(1)}, \dots, \lambda_t^{(m)})$.

**Theorem 3.2.4** *Let the set $Q$ be bounded: $\|x - x_0\| \leq R$ for all $x \in Q$. If the number of iterations $t$ of method (3.2.30) is big enough,*

$$t > \frac{R^2}{h^2}, \tag{3.2.32}$$

*then $N(t) > 0$. Moreover, in this case*

$$\begin{aligned} \max_{1 \leq j \leq m} f_j(x_k) &\leq Mh, \quad k \in \mathscr{A}_0(t), \\ \delta_t &\leq Mh, \end{aligned} \tag{3.2.33}$$

*where $M = \max\limits_{0 \leq k \leq t} \max\limits_{0 \leq j \leq m} \|g_j(x_k)\|$.*

*Proof* Note that

$$\begin{aligned} \sigma_t \cdot \delta_t &\overset{(3.2.31)}{=} \max_{x \in Q} \left\{ \sum_{k \in \mathscr{A}_0(t)} \frac{hf(x_k)}{\|g(x_k)\|} - \sigma_t f(x) - \sum_{j=1}^{m} \sum_{k \in \mathscr{A}_j(t)} h_k f_j(x) \right\} \\ &= \max_{x \in Q} \left\{ \sum_{k \in \mathscr{A}_0(t)} \frac{h(f(x_k) - f(x))}{\|g(x_k)\|} - \sum_{k \notin \mathscr{A}_0(t)} h_k f_{j_k}(x) \right\} \\ &\leq \max_{x \in Q} \left\{ \sum_{k \in \mathscr{A}_0(t)} \frac{h\langle g(x_k), x_k - x \rangle}{\|g(x_k)\|_*} + \sum_{k \notin \mathscr{A}_0(t)} h_k [\langle g_{j_k}(x_k), x_k - x \rangle - f_{j_k}(x_k)] \right\}. \end{aligned} \tag{3.2.34}$$

Let us estimate from above the right-hand side of this inequality. For arbitrary $x \in Q$, let $r_k(x) = \|x - x_k\|$. Assume that $k \in \mathscr{A}_0(t)$. Then

$$\begin{aligned} r_{k+1}^2(x) &\overset{(2.2.48)}{\leq} \left\| x_k - x - \frac{hg(x_k)}{\|g(x_k)\|} \right\|^2 \\ &= r_k^2(x) - \frac{2h}{\|g(x_k)\|} \langle g(x_k), x_k - x \rangle + h^2. \end{aligned} \tag{3.2.35}$$

If $k \notin \mathscr{A}_0(t)$, then

$$\begin{aligned} r_{k+1}^2(x) &\overset{(2.2.48)}{\leq} \|x_k - x - h_k g_{j_k}(x_k)\|^2 \\ &= r_k^2(x) - 2h_k\langle g_{j_k}(x_k), x_k - x\rangle + h_k^2\|g_{j_k}(x_k)\|^2. \end{aligned}$$

Hence,

$$\begin{aligned} 2h_k[\langle g_{j_k}(x_k), x_k - x\rangle - f_{j_k}(x_k)] &\leq r_k^2(x) - r_{k+1}^2(x) - \frac{f_{j_k}^2(x_k)}{\|g_{j_k}(x_k)\|^2} \\ &\leq r_k^2(x) - r_{k+1}^2(x) - h^2. \end{aligned}$$

Summing up these inequalities and inequalities (3.2.35) for $k = 0, \ldots, t$, and taking into account that $r_{t+1}(x) \geq 0$, we get

$$\begin{aligned} \sigma_t\delta_t &\overset{(3.2.34)}{\leq} \tfrac{1}{2}r_0^2(x) + \tfrac{1}{2}N(t)h^2 - \tfrac{1}{2}(t - N(t))h^2 \\ &= \tfrac{1}{2}r_0^2(x) - \tfrac{1}{2}th^2 + N(t)h^2 \leq \tfrac{1}{2}R^2 - \tfrac{1}{2}th^2 + N(t)h^2. \end{aligned} \tag{3.2.36}$$

Assume now that $t$ satisfies the condition (3.2.32). In this case we cannot have $N(t) = 0$ since then $\sigma_t = 0$ and inequality (3.2.36) is violated. Thus, the first inequality in (3.2.33) follows from the conditions of Step (b) in method (3.2.30). Finally, $\sigma_t \overset{(3.2.31)}{\geq} \frac{h}{M}N(t)$. Therefore, if $N(t) > 0$ and the iteration counter $t$ satisfies inequality (3.2.32), then $\delta_t \overset{(3.2.36)}{\leq} \frac{N(t)h^2}{\sigma_t} \leq Mh$. □

### 3.2.6 Strongly Convex Functions

In Sect. 2.1.3, we introduced the notion of strong convexity for differentiable convex functions. We have seen that this additional assumption significantly accelerates optimization methods. Let us study the effect of this assumption on the class of non-differentiable convex functions. For the sake of simplicity, we work in this section with standard Euclidean norm.

**Definition 3.2.2** A function $f$ is called *strongly convex* on a convex set $Q$ if there exists a constant $\mu > 0$ such that for all $x, y \in Q$ and $\alpha \in [0, 1]$ we have

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \tfrac{1}{2}\mu\alpha(1-\alpha)\|x - y\|^2. \tag{3.2.37}$$

For such functions, we use the notation $f \in \mathscr{S}^0_\mu(Q)$. If in this inequality $\mu = 0$, we get definition (3.1.2) of the usual convex function.

Note that for smooth convex functions we proved this inequality as one of the equivalent definitions (2.1.23).

Let us present the most important properties of strongly convex functions.

**Lemma 3.2.3** *Let $f \in \mathscr{S}^0_\mu(Q)$. Then for any $x \in \operatorname{int} Q$ and $y \in W$, we have*

$$f(y) \geq f(x) + f'(x; y - x) + \tfrac{1}{2}\mu\|x - y\|^2. \tag{3.2.38}$$

*Proof* Indeed,

$$\begin{aligned} f(y) &\overset{(3.2.37)}{\geq} \tfrac{1}{\alpha}\left[f((1-\alpha)x + \alpha y) - (1-\alpha)f(x) + \tfrac{1}{2}\mu\alpha(1-\alpha)\|x - y\|^2\right] \\ &= f(x) + \tfrac{1}{\alpha}[f(x + \alpha(y - x)) - f(x)] + \tfrac{1}{2}\mu(1-\alpha)\|y - x\|^2. \end{aligned}$$

Taking in this inequality the limit as $\alpha \downarrow 0$, we get inequality (3.2.38). The limit exists in view of Theorem 3.1.12. □

**Corollary 3.2.1** *Let $f \in \mathscr{S}^0_\mu(Q)$. For any $g \in \partial f(x)$, we have*

$$f(y) \geq f(x) + \langle g, y - x\rangle + \tfrac{1}{2}\mu\|y - x\|^2. \tag{3.2.39}$$

*Proof* Indeed, in view of Theorem 3.1.17, for any $g \in \partial f(x)$ we have

$$f'(x; y - x) \geq \langle g, y - x\rangle. \qquad \square$$

**Corollary 3.2.2** *If in problem (3.2.13) the objective function belongs to the class $\mathscr{S}^0_\mu(Q)$, then its level sets are bounded. Hence, its optimal solution exists.* □

**Corollary 3.2.3** *Let $x^* \in \operatorname{int}\operatorname{dom} f$ be an optimal solution of problem (3.2.13) with $f \in \mathscr{S}^0_\mu$. Then for all $x \in Q$, we have*

$$f(x) \geq f^* + \tfrac{1}{2}\mu\|x - x^*\|^2. \tag{3.2.40}$$

*Hence, the solution of this problem is unique.*

*Proof* Indeed, in view of Theorem 3.1.24, there exists a $g^* \in \partial f(x^*)$ such that

$$\langle g^*, y - x^*\rangle \geq 0.$$

Thus, (3.2.40) follows from (3.2.39). □

Let us describe the results of some operations with strongly convex functions.

1. *Addition.* If $f_1 \in \mathscr{S}^0_{\mu_1}(Q)$ and $f_2 \in \mathscr{S}^0_{\mu_2}(Q)$, then for any $\alpha_1, \alpha_2 \ge 0$ we have

$$\alpha_1 f_1 + \alpha_2 f_2 \in \mathscr{S}^0_{\alpha_1\mu_1+\alpha_2\mu_2}(Q).$$

(The proof follows directly from definition (3.2.37).) In particular, if we add a convex function and a strongly convex function with parameter $\mu$, then we get a strongly convex function with the same value of parameter.

2. *Maximum.* If $f_1 \in \mathscr{S}^0_{\mu_1}(Q)$ and $f_2 \in \mathscr{S}^0_{\mu_2}(Q)$, then

$$f(x) = \max\{f_1(x), f_2(x)\} \ \in \ \mathscr{S}^0_{\mu}(Q)$$

with $\mu = \min\{\mu_1, \mu_2\}$. Indeed, for any $x_1, x_2 \in Q$ and $\alpha \in [0,1]$, we have

$$\begin{aligned} f(\alpha x_1 + (1-\alpha)x_2) \le \max\{&\alpha f_1(x_1) + (1-\alpha) f_1(x_2) \\ &- \frac{1}{2}\mu_1\alpha(1-\alpha)\|x_1 - x_2\|^2, \alpha f_2(x_1) + (1-\alpha) f_2(x_2) \\ &- \frac{1}{2}\mu_2\alpha(1-\alpha)\|x_1 - x_2\|^2\} \\ \le\ & \alpha f(x_1) + (1-\alpha) f(x_2) - \frac{1}{2}\mu\alpha(1-\alpha)\|x_1 - x_2\|^2. \end{aligned}$$

3. *Subtraction.* If $f \in \mathscr{S}^0_{\mu}(Q)$, then the function $\hat{f}(x) = f(x) - \frac{1}{2}\mu\|x\|^2$ is convex. This fact follows from definition (3.2.37) and the Euclidean identity

$$\tfrac{1}{2}\|\alpha x + (1-\alpha)y\|^2 \equiv \tfrac{1}{2}\alpha\|x\|^2 + \tfrac{1}{2}(1-\alpha)\|y\|^2 - \tfrac{1}{2}\alpha(1-\alpha)\|x - y\|^2, \tag{3.2.41}$$

which is valid for all $x, y \in \mathbb{R}^n$ and $\alpha \in [0,1]$.

Note also that any differentiable strongly convex function in the sense of (2.1.20) belongs to the class $\mathscr{S}^0_{\mu}(Q)$ (see Theorem 2.1.9).

Let us now derive the lower complexity bounds for problem (3.2.13) with a strongly convex objective function. For that, we are going to use the function $f_k(\cdot)$ defined by (3.2.3). We add to assumptions (3.2.2) on the problem class the following specification (compare with (3.2.7)).

- The function $f$ is Lipschitz continuous on $B_2(x^*, \|x_0 - x^*\|)$ with constant $M > 0$.
- $f \in \mathscr{S}^0_{\mu}(B_2(x^*, \|x_0 - x^*\|))$ with $\mu > 0$.

(3.2.42)

In what follows, we denote the class of problems satisfying assumptions (3.2.2), (3.2.42) by $\mathscr{P}_s(x_0, \mu, M)$.

**Theorem 3.2.5** *For any class $\mathscr{P}_s(x_0, \mu, M)$ and any $k$, $0 \le k \le n-1$, there exists a function $f \in \mathscr{P}_s(x_0, \mu, M)$ such that*

$$f(x_k) - f^* \ge \frac{M^2}{2\mu(2+\sqrt{k+1})^2} \tag{3.2.43}$$

*for any optimization scheme generating a sequence $\{x_k\}$, which satisfies the condition*

$$x_k \in x_0 + \mathrm{Lin}\,\{g(x_0), \ldots, g(x_{k-1})\}.$$

*Proof* In this proof, we use functions (3.2.3) with the resisting oracle (3.2.6).

Without loss of generality, we can take $x_0 = 0$. Let us choose $f(x) = f_{k+1}(x)$ with parameter

$$\gamma = \frac{M\sqrt{k+1}}{2+\sqrt{k+1}}. \tag{3.2.44}$$

In view of identity (3.2.41) function $f_k$ belongs to the class $\mathscr{S}^0_\mu(\mathbb{R}^n)$. At the same time,

$$R_k \overset{\mathrm{def}}{=} \|x_0 - x_k^*\| \overset{(3.2.5)}{=} \frac{\gamma}{\mu\sqrt{k+1}} \overset{(3.2.44)}{=} \frac{M}{\mu(2+\sqrt{k+1})}.$$

In view of (3.2.4), the Lipschitz constant of the function $f_k$ on the ball $B_2(x_k^*, R_k)$ is bounded by

$$2\mu R_k + \gamma \overset{(3.2.44)}{=} \frac{2M}{2+\sqrt{k+1}} + \frac{M\sqrt{k+1}}{2+\sqrt{k+1}} = M.$$

Thus, optimization problem (3.2.13) with $f = f_{k+1}$ belongs to the problem class $\mathscr{P}_s(x_0, \mu, M)$. At the same time, in view of the condition of the theorem,

$$f(x_k) - f^* \ge -f_{k+1}^* \overset{(3.2.5)}{=} \frac{\gamma^2}{2\mu(k+1)} = \frac{M^2}{2\mu(2+\sqrt{k+1})^2}. \qquad \square$$

It appears that for our problem class the simplest subgradient method is suboptimal.

**Theorem 3.2.6** *Assume that the objective function $f$ in problem (3.2.13) satisfies assumptions (3.2.42). Let $\epsilon > 0$ be the desired accuracy in the optimal value of this problem. Consider a sequence of points $\{x_k\} \subset Q$ generated by the following rule:*

$$x_{k+1} = \pi_Q\left(x_k - \frac{2\epsilon\, g(x_k)}{\|g(x_k)\|^2}\right), \quad k \ge 0, \tag{3.2.45}$$

*where $g(x_k) \in \partial f(x_k)$. Then, if the number of steps $N$ of this scheme is big enough,*

$$N \geq \tfrac{M^2}{\mu\epsilon} \ln \tfrac{M\|x_0 - x^*\|}{\epsilon}, \tag{3.2.46}$$

*we have $f_N^* \overset{\mathrm{def}}{=} \min\limits_{0 \leq k \leq N} f(x_k) \leq f^* + \epsilon$.*

*Proof* Let $r_k = \|x_k - x^*\|$ and $h_k = \frac{2\epsilon}{\|g(x_k)\|^2}$. Assume that $N$ satisfies the lower bound (3.2.46) and

$$f(x_k) - f^* > \epsilon, \quad k = 0, \dots, N. \tag{3.2.47}$$

Then

$$\begin{aligned} r_{k+1}^2 &\overset{(2.2.49)}{\leq} \|x_k - h_k g(x_k)\|^2 \;=\; r_k^2 - 2h_k \langle g(x_k), x_k - x^* \rangle + \tfrac{4\epsilon^2}{\|g(x_k)\|^2} \\ &\overset{(3.2.39)}{\leq} r_k^2 - \tfrac{4\epsilon}{\|g(x_k)\|^2} \left[ f(x_k) - f^* + \tfrac{1}{2}\mu r_k^2 \right] + \tfrac{4\epsilon^2}{\|g(x_k)\|^2} \\ &\overset{(3.2.47)}{\leq} \left(1 - \tfrac{2\mu\epsilon}{\|g(x_k)\|^2}\right) r_k^2. \end{aligned}$$

Thus, all $x_k \in B(x^*, r_0)$ and therefore $\|g(x_k)\| \leq M$. This implies that

$$\epsilon \overset{(3.2.47)}{<} f(x_N) - f^* \leq M r_N \leq M \left(1 - \tfrac{2\mu\epsilon}{M^2}\right)^{N/2} r_0 \leq M \exp\left\{-\tfrac{\mu\epsilon N}{M^2}\right\} r_0.$$

This contradicts the lower bound (3.2.46). □

In view of our assumptions,

$$\tfrac{1}{2}\mu\|x_0 - x^*\|^2 \overset{(3.2.40)}{\leq} f(x_0) - f^* \leq M\|x_0 - x^*\|.$$

Therefore, $\|x_0 - x^*\| \leq \frac{2M}{\mu}$. Thus, the lower bound on the number of iterations (3.2.46) can be rewritten in terms of the class parameters in the following way:

$$N \geq \tfrac{M^2}{\mu\epsilon} \ln \tfrac{2M^2}{\mu\epsilon}. \tag{3.2.48}$$

Comparing it with the lower complexity bound (3.2.43), we can see that the Subgradient Method (3.2.45) is suboptimal. Its main advantage is independence on the exact values of the class parameters $\mu$ and $M$.

Note that the step sizes of method (3.2.45) are twice as big as those of method (3.2.24). If we divide the step sizes in (3.2.45) by two, then, for strongly convex functions, this method will be twice as slow. At the same time, this new version will be identical to (3.2.24) with $m = 0$, which is able to minimize Lipschitz continuous functions with simple set constraints (see Theorem 3.2.3).

### 3.2.7 Complexity Bounds in Finite Dimension

Let us look at the problems of Unconstrained Minimization again, assuming that their dimension is relatively small. This means that our computational resources allow us to perform a number of iterations of minimization schemes proportional to the dimension of the space of variables. What will be the lower complexity bounds in this case?

In this section, we obtain a finite-dimensional lower complexity bound for a problem which is closely related to minimization problems. This is the *feasibility problem*:

$$\boxed{\text{Find } x^* \in S,} \tag{3.2.49}$$

where $S$ is a closed convex set. We assume that this problem is endowed with a *separation oracle*, which answers our request at a point $\bar{x} \in \mathbb{R}^n$ in the following way.

- Either it reports that $\bar{x} \in S$.
- Or, it returns a vector $\bar{g}$, separating $\bar{x}$ from S:

$$\langle \bar{g}, \bar{x} - x \rangle \geq 0 \quad \forall x \in S.$$

In order to measure the complexity of this problem, we introduce the following assumption.

**Assumption 3.2.1** *There exists a point $x^* \in S$ such that for some $\epsilon > 0$ the ball $B_2(x^*, \epsilon)$ belongs to $S$ .*

For example, if we know an optimal value $f^*$ for problem (3.2.8), we can treat this problem as a feasibility problem with

$$S = \{(t, x) \in \mathbb{R}^{n+1} \mid t \geq f(x),\ t \leq f^* + \bar{\epsilon},\ x \in Q\}.$$

The relation between accuracy parameters $\bar{\epsilon}$ and $\epsilon$ in (3.2.2) can be easily obtained, using the assumption that $f$ is Lipschitz continuous. We leave the corresponding reasoning as an exercise for the reader.

Let us describe now a *resisting oracle* for problem (3.2.49). Taking into account the requests of the numerical method, this oracle forms a sequence of boxes $\{B_k\}_{k=0}^{\infty}$, $B_{k+1} \subset B_k$, defined by their lower and upper bounds:

$$B_k = \{x \in \mathbb{R}^n \mid \; a_k \le x \le b_k\}.$$

For each box $B_k$, $k \ge 0$, denote by $c_k = \frac{1}{2}(a_k + b_k)$ its center. For each box $B_k$, $k \ge 1$, the oracle creates an individual separating vector $g_k$. Up to the choice of sign, this is always a coordinate vector.

In the scheme below, we use two dynamic counters:

- $m$ is the number of generated boxes.
- $i$ is the active coordinate.

Denote by $\bar{e}_n \in \mathbb{R}^n$ the vector of all ones. The oracle starts from the following settings:

$$a_0 := -R\bar{e}_n, \quad b_0 := R\bar{e}_n, \quad m := 0, \quad i := 1.$$

Its input is an arbitrary test point $x \in \mathbb{R}^n$.

**Resisting oracle for feasibility problem**

```
If $x \notin B_0$ then return a separator of $x$ from $B_0$ else

1. Find the maximal $k \in [0, \dots, m]$ : $x \in B_k$.
2. If $k < m$ then return $g_k$ else {Create a new box}:

        If $x^{(i)} \ge c_m^{(i)}$ then $a_{m+1} := a_m$,
            $b_{m+1} := b_m + (c_m^{(i)} - b_m^{(i)})e_i$,   $g_m := e_i$.
        else $a_{m+1} := a_m + (c_m^{(i)} - a_m^{(i)})e_i$,
            $b_{m+1} := b_m$,   $g_m := -e_i$.

        $m := m + 1$; $i := i + 1$; If $i > n$ then $i := 1$.

        Return $g_m$.
```

This oracle implements a very simple strategy. Note that the next box $B_{m+1}$ is always half of the last box $B_m$. The last generated box $B_m$ is divided into two equal parts by a hyperplane, defined by the coordinate vector $e_i$, which passes through $c_m$, the center of $B_m$. Depending on the part of the box $B_m$ containing the point $x$, we choose the sign of the separation vector: $g_{m+1} = \pm e_i$. The new box $B_{m+1}$ is always the half of the box $B_m$ which does not contain the test point $x$.

After creating a new box $B_{m+1}$, the index $i$ is increased by 1. If its value exceeds $n$, we set again $i = 1$. Thus, the sequence of boxes $\{B_k\}$ possesses two important properties:

- $\mathrm{vol}_n\, B_{k+1} = \frac{1}{2}\mathrm{vol}_n\, B_k$.
- For any $k \geq 0$ we have $b_{k+n} - a_{k+n} = \frac{1}{2}(b_k - a_k)$.

Note also that the number of generated boxes does not exceed the number of calls of the oracle.

**Lemma 3.2.4** *For all $k \geq 0$ we have the inclusion*

$$B_2(c_k, r_k) \subset B_k, \quad \textit{with} \quad r_k = \tfrac{R}{2}\left(\tfrac{1}{2}\right)^{\frac{k}{n}}. \tag{3.2.50}$$

*Proof* Indeed, for all $k \in [0, \dots, n-1]$ we have

$$B_k \supset B_n = \{x \mid \ c_n - \frac{1}{2}R\bar{e}_n \leq x \leq c_n + \frac{1}{2}R\bar{e}_n\} \supset B_2(c_n, \frac{1}{2}R).$$

Therefore, for such $k$ we have $B_k \supset B_2(c_k, \frac{1}{2}R)$ and (3.2.50) holds. Further, let $k = nl + p$ for some $p \in [0, \dots, n-1]$. Since

$$b_k - a_k = \left(\frac{1}{2}\right)^l (b_p - a_p),$$

we conclude that

$$B_k \supset B_2\left(c_k, \frac{1}{2}R\left(\frac{1}{2}\right)^l\right).$$

It remains to note that $r_k \leq \frac{1}{2}R\left(\frac{1}{2}\right)^l$. □

Lemma 3.2.4 immediately leads to the following complexity result.

**Theorem 3.2.7** *Consider a class of feasibility problems (3.2.49), which satisfy Assumption 3.2.1, and for which the feasible sets $S$ are subsets of $B_\infty(0, R)$. The lower analytical complexity bound for this class is*

$$n \ln \tfrac{R}{2\epsilon}$$

*calls of the separation oracle.*

*Proof* Indeed, we have seen that the number of generated boxes does not exceed the number of calls of the oracle. Moreover, in view of Lemma 3.2.4, after $k$ iterations the last box contains the ball $B_2(c_{m_k}, r_k)$. □

The lower complexity bound for minimization problem (3.2.8) can be obtained in a similar way. However, the corresponding reasoning is more complicated. Therefore we present here only the final result.

**Theorem 3.2.8** *A lower bound for the analytical complexity of the problem class formed by minimization problem (3.2.8) with $Q \subseteq B_\infty(0,R)$ and $f \in \mathscr{F}_M^{0,0}(B_\infty(0,R))$, is $n \ln \frac{MR}{8\epsilon}$ calls of the oracle.* □

### *3.2.8 Cutting Plane Schemes*

Let us look now at the following minimization problem with set constraint:

$$\min\{f(x) \mid x \in Q\}, \tag{3.2.51}$$

where the function $f$ is convex on $\mathbb{R}^n$, and $Q$ is a bounded closed convex set such that

$$\operatorname{int} Q \neq \emptyset, \quad \operatorname{diam} Q = D < \infty.$$

We assume that $Q$ is not simple and that our problem is equipped with a separation oracle. At any test point $\bar{x} \in \mathbb{R}^n$, this oracle returns a vector $g(x)$, which is either:

- a subgradient of $f$ at $\bar{x}$, if $x \in Q$,
- a separator of $\bar{x}$ from $Q$, if $x \notin Q$.

An important example of such a problem is a constrained minimization problem with functional constraints (3.2.22). We have seen that this problem can be rewritten as a problem with a single functional constraint (see (3.2.23)) defining the feasible set

$$Q = \{x \in \mathbb{R}^n \mid \bar{f}(x) \leq 0\}.$$

In this case, for $x \notin Q$ the oracle has to provide us with any subgradient $\bar{g} \in \partial \bar{f}(x)$. Clearly, $\bar{g}$ separates $x$ from $Q$ (see Theorem 3.1.18).

Let us present the main property of finite-dimensional localization sets.

Consider a sequence $X \equiv \{x_i\}_{i=0}^\infty$ belonging to the set $Q$. Recall that the localization sets generated by this sequence are defined as follows:

$$S_0(X) = Q,$$

$$S_{k+1}(X) = \{x \in S_k(X) \mid \langle g(x_k), x_k - x\rangle \geq 0\}.$$

Clearly, for any $k \geq 0$ we have $x^* \in S_k$. Define

$$v_i = v_f(x^*; x_i) \ (\geq 0), \quad v_k^* = \min_{0 \leq i \leq k} v_i.$$

Denote by $\mathrm{vol}_n\, S$ the $n$-dimensional volume of the set $S \subset \mathbb{R}^n$.

**Theorem 3.2.9** *For any $k \geq 0$ we have*

$$v_k^* \leq D \left[ \frac{\mathrm{vol}_n\, S_k(X)}{\mathrm{vol}_n\, Q} \right]^{\frac{1}{n}}.$$

*Proof* Let $\alpha = v_k^*/D \ (\leq 1)$. Since $Q \subseteq B_2(x^*, D)$ we have the following inclusion:

$$(1-\alpha)x^* + \alpha Q \subseteq (1-\alpha)x^* + \alpha B_2(x^*, D) \ = \ B_2(x^*, v_k^*).$$

Since $Q$ is convex, we conclude that

$$(1-\alpha)x^* + \alpha Q \equiv [(1-\alpha)x^* + \alpha Q] \bigcap Q \ \subseteq \ B_2(x^*, v_k^*) \bigcap Q \subseteq S_k(X).$$

Therefore $\mathrm{vol}_n\, S_k(X) \geq \mathrm{vol}_n\, [(1-\alpha)x^* + \alpha Q] = \alpha^n \mathrm{vol}_n\, Q$. □

Quite often, the set $Q$ is very complicated and it is difficult to work directly with the sets $S_k(X)$. Instead, we can update some simple *upper approximations* of these sets. The process of generating such approximations is described by the following *cutting plane* scheme.

**General cutting plane scheme**

```
0. Choose a bounded set $E_0 \supseteq Q$.
1. $k$th iteration ($k \geq 0$).
   (a) Choose $y_k \in E_k$
   (b) If $y_k \in Q$ then compute $f(y_k)$, $g(y_k)$. If $y_k \notin Q$, then
       compute $\bar{g}(y_k)$, which separates $y_k$ from $Q$.
   (c) Set
       $g_k = \begin{cases} g(y_k), & \text{if } y_k \in Q, \\ \bar{g}(y_k), & \text{if } y_k \notin Q. \end{cases}$
   (d) Choose $E_{k+1} \supseteq \{x \in E_k \mid \langle g_k, y_k - x \rangle \geq 0\}$.
```

(3.2.52)

Let us estimate the performance of this process. Consider the sequence $Y = \{y_k\}_{k=0}^{\infty}$, involved in this scheme. Denote by $X$ a subsequence of feasible points in the sequence $Y$: $X = Y \bigcap Q$. Let us introduce the counter

$$i(k) \;=\; \text{number of points } y_j,\ 0 \le j < k, \text{ such that } y_j \in Q.$$

Thus, if $i(k) > 0$, then $X \neq \emptyset$.

**Lemma 3.2.5** *For any $k \ge 0$, we have $S_{i(k)} \subseteq E_k$.*

*Proof* Indeed, if $i(0) = 0$, then $S_0 = Q \subseteq E_0$. Let us assume that $S_{i(k)} \subseteq E_k$ for some $k \ge 0$. Then, at the next iteration there are two possibilities.

(a) $i(k+1) = i(k)$. This happens if and only if $y_k \notin Q$. Then

$$\begin{aligned} E_{k+1} &\supseteq \{x \in E_k \mid \langle \bar{g}(y_k), y_k - x\rangle \ge 0\} \\ &\supseteq \{x \in S_{i(k+1)} \mid \langle \bar{g}(y_k), y_k - x\rangle \ge 0\} \;=\; S_{i(k+1)} \end{aligned}$$

since $S_{i(k+1)} \subseteq Q$ and $\bar{g}(y_k)$ separates $y_k$ from $Q$.

(b) $i(k+1) = i(k) + 1$. In this case $y_k \in Q$. Then

$$\begin{aligned} E_{k+1} &\supseteq \{x \in E_k \mid \langle g(y_k), y_k - x\rangle \ge 0\} \\ &\supseteq \{x \in S_{i(k)} \mid \langle g(y_k), y_k - x\rangle \ge 0\} \;=\; S_{i(k)+1} \end{aligned}$$

since $y_k = x_{i(k)}$. □

The above results immediately lead to the following important conclusion.

**Corollary 3.2.4**

1. *For any $k$ such that $i(k) > 0$, we have*

$$v^*_{i(k)}(X) \le D\left[\frac{\mathrm{vol}_n S_{i(k)}(X)}{\mathrm{vol}_n Q}\right]^{\frac{1}{n}} \le D\left[\frac{\mathrm{vol}_n E_k}{\mathrm{vol}_n Q}\right]^{\frac{1}{n}}.$$

2. *If* $\mathrm{vol}_n E_k < \mathrm{vol}_n Q$, *then $i(k) > 0$.*

*Proof* We have already proved the first statement. The second one follows from the inclusion $Q = S_0 = S_{i(k)} \subseteq E_k$, which is valid for all $k$ such that $i(k) = 0$. □

Thus, if we manage to ensure $\mathrm{vol}_n E_k \to 0$, then we obtain a convergent scheme. Moreover, the rate of decrease of the volume automatically defines the rate of convergence of the corresponding method. Clearly, we should try to decrease $\mathrm{vol}_n E_k$ as quickly as possible.

Historically, the first nonsmooth minimization method, implementing the idea of cutting planes, was the *Center of Gravity Method*. It is based on the following geometric idea.

Consider a bounded convex set $S \subset \mathbb{R}^n$, $\operatorname{int} S \neq \emptyset$. Define the *center of gravity* of this set as

$$cg(S) = \frac{1}{\operatorname{vol}_n S} \int\limits_S x dx.$$

It appears that any cutting plane passing through the center of gravity divides the set into two almost proportional pieces.

**Lemma 3.2.6** *Let $g$ be a direction in $\mathbb{R}^n$. Define*

$$S_+ = \{x \in S \mid \langle g, cg(S) - x \rangle \geq 0\}.$$

*Then*

$$\frac{\operatorname{vol}_n S_+}{\operatorname{vol}_n S} \leq 1 - \frac{1}{e}.$$

(We accept this result without proof.) □

This observation naturally leads to the following minimization scheme.

**Method of Centers of Gravity**

```
0. Set $S_0 = Q$.
1. $k$th iteration ($k \geq 0$).
    (a) Choose $x_k = cg(S_k)$ and compute $f(x_k)$, $g(x_k)$.
    (b) Set $S_{k+1} = \{x \in S_k \mid \langle g(x_k), x_k - x \rangle \geq 0\}$.
```

Let us estimate the rate of convergence of this method. Define

$$f_k^* = \min_{0 \leq j \leq k} f(x_j).$$

**Theorem 3.2.10** *If $f$ is Lipschitz continuous on $B_2(x^*, D)$ with constant $M$, then for any $k \geq 0$ we have*

$$f_k^* - f^* \leq MD \left(1 - \frac{1}{e}\right)^{\frac{k}{n}}.$$

*Proof* The statement follows from Lemma 3.2.2, Theorem 3.2.9 and Lemma 3.2.6. □

Comparing this result with the lower complexity bound of Theorem 3.2.8, we see that the method of centers of gravity is optimal in finite dimensions. Its rate of convergence does not depend on any individual characteristics of our problem like the condition number, etc. However, we should accept that this method is absolutely impractical, since the computation of the center of gravity in a high-dimensional space is a more difficult problem than the problem of Convex Optimization.

Let us look at another method, which uses the possibility of approximating the localization sets. This method is based on the following geometrical observation.

Let $H$ be a positive definite symmetric $n \times n$ matrix. Consider the *ellipsoid*

$$E(H, \bar{x}) = \{x \in \mathbb{R}^n \mid \langle H^{-1}(x - \bar{x}), x - \bar{x}\rangle \leq 1\}.$$

Let us choose a direction $g \in \mathbb{R}^n$, and consider a half of the above ellipsoid, defined by the corresponding hyperplane:

$$E_+ = \{x \in E(H, \bar{x}) \mid \langle g, \bar{x} - x\rangle \geq 0\}.$$

It turns out that this set belongs to another ellipsoid, whose volume is strictly smaller than the volume of $E(H, \bar{x})$.

**Lemma 3.2.7** *Define*

$$\bar{x}_+ = \bar{x} - \tfrac{1}{n+1} \cdot \tfrac{Hg}{\langle Hg, g\rangle^{1/2}},$$

$$H_+ = \tfrac{n^2}{n^2-1}\left(H - \tfrac{2}{n+1} \cdot \tfrac{Hgg^T H}{\langle Hg, g\rangle}\right).$$

*Then* $E_+ \subset E(H_+, \bar{x}_+)$ *and*

$$\text{vol}_n\, E(H_+, \bar{x}_+) \leq \left(1 - \tfrac{1}{(n+1)^2}\right)^{\frac{n}{2}} \text{vol}_n\, E(H, \bar{x}).$$

*Proof* Let $G = H^{-1}$ and $G_+ = H_+^{-1}$. It is clear that

$$G_+ = \tfrac{n^2-1}{n^2}\left(G + \tfrac{2}{n-1} \cdot \tfrac{gg^T}{\langle Hg, g\rangle}\right).$$

Without loss of generality we can assume that $\bar{x} = 0$ and $\langle Hg, g\rangle = 1$. Suppose $x \in E_+$. Note that $\bar{x}_+ = -\frac{1}{n+1} Hg$. Therefore,

$$\| x - \bar{x}_+ \|^2_{G_+} = \tfrac{n^2-1}{n^2} \left( \| x - \bar{x}_+ \|^2_G + \tfrac{2}{n-1} \langle g, x - \bar{x}_+\rangle^2 \right),$$

$$\| x - \bar{x}_+ \|^2_G = \| x \|^2_G + \tfrac{2}{n+1} \langle g, x\rangle + \tfrac{1}{(n+1)^2},$$

$$\langle g, x - \bar{x}_+\rangle^2 = \langle g, x\rangle^2 + \tfrac{2}{n+1} \langle g, x\rangle + \tfrac{1}{(n+1)^2}.$$

Putting all the terms together, we obtain

$$\| x - \bar{x}_+ \|^2_{G_+} = \tfrac{n^2-1}{n^2} \left( \| x \|^2_G + \tfrac{2}{n-1} \langle g, x\rangle^2 + \tfrac{2}{n-1} \langle g, x\rangle + \tfrac{1}{n^2-1} \right).$$

Note that $\langle g, x\rangle \leq 0$ and $\| x \|_G \leq 1$. Therefore

$$\langle g, x\rangle^2 + \langle g, x\rangle = \langle g, x\rangle (1 + \langle g, x\rangle) \leq 0.$$

Hence,

$$\| x - \bar{x}_+ \|^2_{G_+} \leq \tfrac{n^2-1}{n^2} \left( \| x \|^2_G + \tfrac{1}{n^2-1} \right) \leq 1.$$

Thus, we have proved that $E_+ \subset E(H_+, \bar{x}_+)$.

Let us estimate the volume of $E(H_+, \bar{x}_+)$.

$$\begin{aligned} \frac{\mathrm{vol}_n\, E(H_+, \bar{x}_+)}{\mathrm{vol}_n\, E(H, \bar{x})} &= \left[ \frac{\det H_+}{\det H} \right]^{1/2} = \left[ \left( \frac{n^2}{n^2-1} \right)^n \frac{n-1}{n+1} \right]^{1/2} \\ &= \left[ \frac{n^2}{n^2-1} \left( 1 - \frac{2}{n+1} \right)^{\frac{1}{n}} \right]^{\frac{n}{2}} \leq \left[ \frac{n^2}{n^2-1} \left( 1 - \frac{2}{n(n+1)} \right) \right]^{\frac{n}{2}} \\ &= \left[ \frac{n^2(n^2+n-2)}{n(n-1)(n+1)^2} \right]^{\frac{n}{2}} = \left[ 1 - \frac{1}{(n+1)^2} \right]^{\frac{n}{2}}. \qquad \square \end{aligned}$$

It turns out that the ellipsoid $E(H_+, \bar{x}_+)$ is the ellipsoid of *minimal* volume containing half of the initial ellipsoid $E_+$.

Our observations can be implemented in the following algorithmic scheme of the famous *Ellipsoid Method*.

**Ellipsoid Method**

**0.** Choose $y_0 \in \mathbb{R}^n$ and $R > 0$ such that $B_2(y_0, R) \supseteq Q$. Set $H_0 = R^2 \cdot I_n$.

**1.** ***k*th iteration ($k \geq 0$).**

$$g_k = \begin{cases} g(y_k), \text{ if } y_k \in Q, \\ \\ \bar{g}(y_k), \text{ if } y_k \notin Q, \end{cases} \tag{3.2.53}$$

$$y_{k+1} = y_k - \tfrac{1}{n+1} \cdot \tfrac{H_k g_k}{\langle H_k g_k, g_k \rangle^{1/2}},$$

$$H_{k+1} = \tfrac{n^2}{n^2-1} \left( H_k - \tfrac{2}{n+1} \cdot \tfrac{H_k g_k g_k^T H_k}{\langle H_k g_k, g_k \rangle} \right).$$

This method can be seen as a particular implementation of the general cutting plane scheme (3.2.52) by choosing

$$E_k = \{x \in \mathbb{R}^n \mid \langle H_k^{-1}(x - y_k), x - y_k \rangle \leq 1\}$$

with $y_k$ being the center of the ellipsoid.

Let us present an efficiency estimate for the Ellipsoid Method. Let $Y = \{y_k\}_{k=0}^{\infty}$, and let $X$ be a feasible subsequence of sequence $Y$:

$$X = Y \bigcap Q.$$

Define $f_k^* = \min\limits_{0 \leq j \leq k} f(x_j)$.

**Theorem 3.2.11** *Let the function $f$ be Lipschitz continuous on $B_2(x^*, R)$ with some constant $M$. Then for $i(k) > 0$, we have*

$$f_{i(k)}^* - f^* \leq MR \left(1 - \tfrac{1}{(n+1)^2}\right)^{\frac{k}{2}} \cdot \left[\tfrac{\mathrm{vol}_n B_2(x_0, R)}{\mathrm{vol}_n Q}\right]^{\frac{1}{n}}.$$

*Proof* The proof follows from Lemma 3.2.2, Corollary 3.2.4 and Lemma 3.2.7. □

We need additional assumptions to guarantee $X \neq \emptyset$. Assume that there exists some $\rho > 0$ and $\bar{x} \in Q$ such that

$$B_2(\bar{x}, \rho) \subseteq Q. \tag{3.2.54}$$

Then

$$\left[\frac{\mathrm{vol}_n E_k}{\mathrm{vol}_n Q}\right]^{\frac{1}{n}} \le \left(1 - \frac{1}{(n+1)^2}\right)^{\frac{k}{2}} \left[\frac{\mathrm{vol}_n B_2(x_0,R)}{\mathrm{vol}_n Q}\right]^{\frac{1}{n}} \le \frac{1}{\rho} e^{-\frac{k}{2(n+1)^2}} R.$$

In view of Corollary 3.2.4, this implies that $i(k) > 0$ for all

$$k > 2(n+1)^2 \ln \frac{R}{\rho}.$$

If $i(k) > 0$, then

$$f^*_{i(k)} - f^* \le \frac{1}{\rho} M R^2 \cdot e^{-\frac{k}{2(n+1)^2}}.$$

In order to ensure that (3.2.54) holds for a constrained minimization problem with functional constraints, it is enough to assume that all constraints are Lipschitz continuous and there is a feasible point at which all functional constraints are *strictly negative* (the Slater condition). We leave the details of the corresponding justification as an exercise for the reader.

Let us discuss now the total complexity of the Ellipsoid Method (3.2.53). Each iteration of this scheme is relatively cheap: it takes $O(n^2)$ arithmetic operations. On the other hand, in order to generate an $\epsilon$-solution of problem (3.2.51), satisfying assumption (3.2.54), this method needs

$$2(n+1)^2 \ln \frac{MR^2}{\rho\epsilon}$$

calls of the oracle. This efficiency estimate is not optimal (see Theorem 3.2.8), but it has linear dependence on $\ln \frac{1}{\epsilon}$, and polynomial dependence on the dimension and the logarithms of the class parameters $M$, $R$ and $\rho$. For problem classes, whose oracle also has a polynomial complexity, such algorithms are called (weakly) *polynomial*.

To conclude this section, note that there are several methods which work with localization sets in the form of the polytope:

$$E_k = \{x \in \mathbb{R}^n \mid \langle a_j, x\rangle \le b_j, \ j = 1 \dots m_k\}.$$

Let us mention the most important methods of this type:

- *Inscribed Ellipsoid Method*. The point $y_k$ in this scheme is chosen as follows:

$$y_k = \text{Center of the maximal ellipsoid } W_k : \ W_k \subset E_k.$$

- *Analytic Center Method*. In this method, the point $y_k$ is chosen as the minimum of the *analytic barrier*

$$F_k(x) = -\sum_{j=1}^{m_k} \ln(b_j - \langle a_j, x\rangle).$$

- *Volumetric Center Method*. This is also a barrier-type scheme. The point $y_k$ is chosen as the minimum of the *volumetric barrier*

$$V_k(x) = \ln \det \nabla^2 F_k(x),$$

where $F_k(\cdot)$ is the analytic barrier for the set $E_k$.

All these methods are polynomial with complexity bound

$$n \left( \ln \frac{1}{\epsilon} \right)^p,$$

where $p$ is either 1 or 2. However, the complexity of each iteration in these methods is much larger ($n^3 - n^4$ arithmetic operations). In Chap. 5, we will see that the test points $y_k$ for these schemes can be efficiently computed by *Interior-Point Methods*.

## 3.3 Methods with Complete Data

(Nonsmooth models of objective function; Kelley's method; The Level Method; Unconstrained minimization; Efficiency estimates; Problems with functional constraints.)

### *3.3.1 Nonsmooth Models of the Objective Function*

In the previous section, we looked at several methods for solving the following problem:

$$\min_{x \in Q} f(x), \tag{3.3.1}$$

where $f$ is a Lipschitz continuous convex function and $Q$ is a closed convex set. We have seen that the optimal method for problem (3.3.1) is the *Subgradient Method* (3.2.14), (3.2.16). Note that this conclusion is valid for the *whole* class of Lipschitz continuous functions. However, if we are going to minimize a particular function from this class, we can expect that it will not be as bad as in the worst case. We usually can hope that the actual performance of the minimization methods can be much better than the worst-case theoretical bound. Unfortunately, as far as the Subgradient Method is concerned, these expectations are too optimistic. The scheme of the Subgradient Method is very strict and in general it *cannot* converge faster than in theory. It can also be shown that the Ellipsoid Method (3.2.53) inherits this drawback of subgradient schemes. In practice it works more or less in accordance with its theoretical bound even when it is applied to a very simple function like $\| x \|^2$.

In this section, we will discuss algorithmic schemes which are more flexible than the Subgradient Method and Ellipsoid Method. These schemes are based on the notion of a *nonsmooth model* of a convex objective function.

**Definition 3.3.1** Let $X = \{x_k\}_{k=0}^{\infty}$ be a sequence of points in $Q$. Define

$$\hat{f}_k(X; x) = \max_{0 \le i \le k} [f(x_i) + \langle g(x_i), x - x_i \rangle],$$

where $g(x_i)$ are some subgradients of $f$ at $x_i$. The function $\hat{f}_k(X; \cdot)$ is called a *nonsmooth model* of the convex function $f$.

Note that $f_k(X; \cdot)$ is a piece-wise linear function. In view of inequality (3.1.23), we always have

$$f(x) \ \ge \ \hat{f}_k(X; x)$$

for all $x \in \mathbb{R}^n$. However, at all test points $x_i$, $0 \le i \le k$, we have

$$f(x_i) = \hat{f}_k(X; x_i), \quad g(x_i) \in \partial \hat{f}_k(X; x_i).$$

Moreover, the next model is always better than the previous one:

$$\hat{f}_{k+1}(X; x) \ \ge \ \hat{f}_k(X; x)$$

for all $x \in \mathbb{R}^n$.

## 3.3.2 Kelley's Method

The model $\hat{f}_k(X; \cdot)$ represents *complete information* on the function $f$ accumulated after $k$ calls of the oracle. Therefore, it seems natural to develop a minimization scheme, based on this object. Perhaps, the most natural method of this type is as follows.

**Kelley's Method**

```
0. Choose $x_0 \in Q$.
1. $k$th iteration ($k \ge 0$).
   Find $x_{k+1} \in \operatorname{Arg}\min_{x \in Q} \hat{f}_k(X; x)$.
```

(3.3.2)

Intuitively, this scheme looks very attractive. Even the presence of a complicated auxiliary problem is not too disturbing, since for polyhedral $Q$ it can be solved by linear optimization methods in finite time. However, it turns out that this method cannot be recommended for practical applications. The main reason for this is its instability. Note that the solution of the auxiliary problem in method (3.3.2) may be not unique. Moreover, the whole set $\operatorname{Arg}\min\limits_{x\in Q} \hat{f}_k(X;x)$ can be unstable with respect to an arbitrary small variation of data $\{f(x_i), g(x_i)\}$. This feature results in unstable practical behavior of the scheme. At the same time, it can be used to construct an example of a problem for which method (3.3.2) has a very disappointing *lower* complexity bound.

*Example 3.3.1* Consider the problem (3.3.1) with

$$f(y,x) = \max\{| \, y \, |, \| \, x \, \|^2\}, \quad y \in \mathbb{R}, \; x \in \mathbb{R}^n,$$

$$Q = \{z = (y,x) : \; y^2 + \| \, x \, \|^2 \leq 1\},$$

where the norm is standard Euclidean. Thus, the solution of this problem is $z^* = (y^*, x^*) = (0,0)$, and the optimal value $f^* = 0$. Denote by $Z_k^* = \operatorname{Arg}\min\limits_{z\in Q} \hat{f}_k(Z;z)$ the optimal set of model $\hat{f}_k(Z;z)$ and let $\hat{f}_k^* = \hat{f}_k(Z_k^*)$ be the optimal value of the model.

Let us choose $z_0 = (1,0)$. Then the initial model of the function $f$ is $\hat{f}_0(Z;z) = y$. Therefore, the first point, generated by Kelley's method, is $z_1 = (-1,0)$. Hence, the next model of the function $f$ is as follows:

$$\hat{f}_1(Z;z) = \max\{y, -y\} = | \, y \, | \, .$$

Clearly, $\hat{f}_1^* = 0$. Note that $\hat{f}_{k+1}^* \geq \hat{f}_k^*$. On the other hand,

$$\hat{f}_k^* \leq f(z^*) = 0.$$

Thus, for all consequent models with $k \geq 1$, we will have $\hat{f}_k^* = 0$ and $Z_k^* = (0, X_k^*)$, where

$$X_k^* = \{x \in B_2(0,1) : \| \, x_i \, \|^2 + \langle 2x_i, x - x_i \rangle \leq 0, \; i = 0 \dots k\}.$$

Let us estimate the efficiency of the cuts for the set $X_k^*$. Since $x_{k+1}$ can be an *arbitrary* point from $X_k^*$, at the first stage of the method we can choose $x_i$ with the unit norms: $\| \, x_i \, \| = 1$. Then the set $X_k^*$ is defined as follows:

$$X_k^* = \{x \in B_2(0,1) \mid \langle x_i, x \rangle \leq \frac{1}{2}, i = 0 \dots k\}.$$

We can do this if

$$S_2(0,1) \equiv \{x \in \mathbb{R}^n \mid \parallel x \parallel = 1\} \bigcap X_k^* \neq \emptyset.$$

As far as this is possible, we can have

$$f(z_i) \equiv f(0, x_i) = 1.$$

Let us estimate the possible length of this stage using the following fact.

Let $d$ be a direction in $\mathbb{R}^n$, $\parallel d \parallel = 1$. Consider a surface

$$S_d(\alpha) = \{x \in \mathbb{R}^n \mid \parallel x \parallel = 1,\ \langle d, x\rangle \geq \alpha\}, \quad \alpha \in [\frac{1}{2}, 1].$$

Then $v(\alpha) \equiv \mathrm{vol}_{n-1}(S(\alpha)) \leq v(0)\left[1 - \alpha^2\right]^{\frac{n-1}{2}}$.

At the first stage, each step cuts from the sphere $S_2(0,1)$ one of the segments $S_d(\frac{1}{2})$, at most. Therefore, we can continue the process for all $k \leq \left[\frac{2}{\sqrt{3}}\right]^{n-1}$. During these iterations we still have $f(z_i) = 1$.

Since at the first stage of the process the cuts are $\langle x_i, x\rangle \leq \frac{1}{2}$, for all $k$, $0 \leq k \leq N \equiv \left[\frac{2}{\sqrt{3}}\right]^{n-1}$, we have

$$B_2(0, \frac{1}{2}) \subset X_k^*.$$

This means that after $N$ iterations we can repeat our process with the ball $B_2(0, \frac{1}{2})$, etc. Note that $f(0, x) = \frac{1}{4}$ for all $x$ from $B_2(0, \frac{1}{2})$.

Thus, we prove the following *lower* bound for the Kelley's method (3.3.2):

$$f(x_k) - f^* \geq \left(\tfrac{1}{4}\right)^{k\left[\frac{\sqrt{3}}{2}\right]^{n-1}}.$$

This means that we cannot get an $\epsilon$-solution of our problem in fewer than

$$\tfrac{1}{2\ln 2}\left[\tfrac{2}{\sqrt{3}}\right]^{n-1} \ln \tfrac{1}{\epsilon}$$

calls of the oracle. It remains to compare this lower bound with the upper complexity bounds of other methods:

| | |
|---|---|
| **Ellipsoid method**: | $O\left(n^2 \ln \frac{1}{\epsilon}\right)$ |
| **Optimal methods**: | $O\left(n \ln \frac{1}{\epsilon}\right)$ |
| **Gradient method**: | $O\left(\frac{1}{\epsilon^2}\right)$ |

### *3.3.3 The Level Method*

Let us show that it is possible to work with a nonsmooth model of the objective function in a stable way. Define

$$\hat{f}_k^* = \min_{x \in Q} \hat{f}_k(X; x), \quad f_k^* = \min_{0 \le i \le k} f(x_i).$$

The first of these values is called the *minimal value* of the model, and the second one is the *record value* of the model. Clearly $\hat{f}_k^* \le f^* \le f_k^*$.

Let us choose some $\alpha \in (0, 1)$. Define

$$\ell_k(\alpha) = (1 - \alpha)\hat{f}_k^* + \alpha f_k^*.$$

Consider the level set

$$\mathscr{L}_k(\alpha) = \{x \in Q \mid \hat{f}_k(X; x) \le \ell_k(\alpha)\}.$$

Clearly, $\mathscr{L}_k(\alpha)$ is a closed convex set.

Note that the set $\mathscr{L}_k(\alpha)$ is certainly interesting for optimization schemes. Firstly, inside this set there is clearly no test point of the current model. Secondly, this set is stable with respect to a small perturbation of the data. Let us present a minimization method which deals directly with this level set.

**Level Method**

```
0. Choose a point $x_0 \in Q$, accuracy $\epsilon > 0$, and level coefficient $\alpha \in (0, 1)$.
1. $k$th iteration ($k \geq 0$).
    (a) Compute $\hat{f}_k^*$ and $f_k^*$.
    (b) If $f_k^* - \hat{f}_k^* \leq \epsilon$, then STOP.
    (c) Set $x_{k+1} = \pi_{\mathscr{L}_k(\alpha)}(x_k)$.
```

(3.3.3)

In this scheme, there are two potentially expensive operations. We need to compute an optimal value $\hat{f}_k^*$ of the current model. If $Q$ is a polytope, then this value can be obtained from the following linear programming problem:

$$\begin{aligned} &\min \quad t, \\ &\text{s.t. } f(x_i) + \langle g(x_i), x - x_i \rangle \leq t, \ i = 0 \ldots k, \\ &x \in Q. \end{aligned}$$

We also need to compute the Euclidean projection $\pi_{\mathscr{L}_k(\alpha)}(x_k)$. If $Q$ is a polytope, then this is a quadratic programming problem:

$$\begin{aligned} &\min \quad \| x - x_k \|^2, \\ &\text{s.t. } f(x_i) + \langle g(x_i), x - x_i \rangle \leq \ell_k(\alpha), \ i = 0 \ldots k, \\ &x \in Q. \end{aligned}$$

Both problems are solvable either by a standard simplex-type method, or by Interior-Point Methods (see Chap. 5).

Let us look at some properties of the Level Method. Recall that the optimal values of the model increase, and the record values decrease:

$$\hat{f}_k^* \leq \hat{f}_{k+1}^* \leq f^* \leq f_{k+1}^* \leq f_k^*.$$

Let $\Delta_k = [\hat{f}_k^*, f_k^*]$ and $\delta_k = f_k^* - \hat{f}_k^*$. We call $\delta_k$ the *gap* of the model $\hat{f}_k(X; x)$. Then

$$\Delta_{k+1} \subseteq \Delta_k, \quad \delta_{k+1} \leq \delta_k.$$

The next result is crucial for the analysis of the Level Method.

**Lemma 3.3.1** *Assume that for some $p \geq k$ the gap is still big enough:*

$$\delta_p \geq (1-\alpha)\delta_k.$$

*Then for all $i$, $k \leq i \leq p$, we have $\ell_i(\alpha) \geq \hat{f}_p^*$.*

*Proof* Note that for all such $i$, we have $\delta_p \geq (1-\alpha)\delta_k \geq (1-\alpha)\delta_i$. Therefore,

$$\ell_i(\alpha) = f_i^* - (1-\alpha)\delta_i \geq f_p^* - (1-\alpha)\delta_i = \hat{f}_p^* + \delta_p - (1-\alpha)\delta_i \geq \hat{f}_p^*. \qquad \square$$

Let us show that the steps of Level Method are large enough. Define

$$M_f = \max\{\| g \| \mid g \in \partial f(x), \ x \in Q\}.$$

**Lemma 3.3.2** *For the sequence of points $\{x_k\}$ generated by the Level Method, we have*

$$\| x_{k+1} - x_k \| \geq \tfrac{(1-\alpha)\delta_k}{M_f}.$$

*Proof* Indeed,

$$\begin{aligned} f(x_k) - (1-\alpha)\delta_k \geq f_k^* - (1-\alpha)\delta_k &= \ell_k(\alpha) \\ &\geq \hat{f}_k(x_{k+1}) \geq f(x_k) + \langle g(x_k), x_{k+1} - x_k\rangle \\ &\geq f(x_k) - M_f \| x_{k+1} - x_k \| . \end{aligned}$$

□

Finally, we need to show that the gap of the model is decreasing.

**Lemma 3.3.3** *Let the set $Q$ in problem (3.3.1) be bounded:* diam $Q \leq D$. *If for some $p \geq k$ we have $\delta_p \geq (1-\alpha)\delta_k$, then*

$$p + 1 - k \leq \tfrac{M_f^2 D^2}{(1-\alpha)^2 \delta_p^2}.$$

*Proof* Let $x_p^* \in \operatorname{Arg}\min\limits_{x \in Q} \hat{f}_p(X; x)$. In view of Lemma 3.3.1, we have

$$\hat{f}_i(X; x_p^*) \leq \hat{f}_p(X; x_p^*) = \hat{f}_p^* \leq \ell_i(\alpha)$$

for all $i, k \leq i \leq p$. Therefore, in view of Lemma 2.2.8 and Lemma 3.3.2, we get

$$\| x_{i+1} - x_p^* \|^2 \leq \| x_i - x_p^* \|^2 - \| x_{i+1} - x_i \|^2 \leq \| x_i - x_p^* \|^2 - \frac{(1-\alpha)^2 \delta_i^2}{M_f^2}$$

$$\leq \| x_i - x_p^* \|^2 - \frac{(1-\alpha)^2 \delta_p^2}{M_f^2}.$$

Summing up these inequalities in $i = k \dots p$, we get

$$(p + 1 - k) \frac{(1-\alpha)^2 \delta_p^2}{M_f^2} \leq \| x_k - x_p^* \|^2 \leq D^2. \qquad \square$$

Note that the number of indices in the segment $[k, p]$ is equal to $p + 1 - k$. Now we can prove the efficiency estimate of the Level Method.

**Theorem 3.3.1** *Let* diam $Q = D$. *Then Level Method terminates after*

$$N = \left\lfloor \frac{M_f^2 D^2}{\epsilon^2 \alpha (1-\alpha)^2 (2-\alpha)} \right\rfloor + 1$$

*iterations at most. The termination criterion of the method guarantees* $f_k^* - f^* \leq \epsilon$.

*Proof* Assume that $\delta_k \geq \epsilon$, $0 \leq k \leq N$. Let us represent the whole set of indices in *decreasing order* as a union of $m + 1$ groups,

$$\{N, \dots, 0\} = I(0) \bigcup I(1) \bigcup \cdots \bigcup I(m),$$

such that

$$I(j) = [p(j), k(j)], \quad p(j) \geq k(j), \quad j = 0 \dots m,$$

$$p(0) = N, \quad p(j + 1) = k(j) - 1, \quad k(m) = 0,$$

$$\delta_{k(j)} \leq \tfrac{1}{1-\alpha} \delta_{p(j)} < \delta_{k(j)+1} \equiv \delta_{p(j+1)}.$$

Clearly, for $j \geq 0$ we have

$$\delta_{p(j+1)} \geq \frac{\delta_{p(j)}}{1-\alpha} \geq \frac{\delta_{p(0)}}{(1-\alpha)^{j+1}} \geq \frac{\epsilon}{(1-\alpha)^{j+1}}.$$

In view of Lemma 3.3.3, $n(j) = p(j) + 1 - k(j)$ is bounded:

$$n(j) \leq \frac{M_f^2 D^2}{(1-\alpha)^2 \delta_{p(j)}^2} \leq \frac{M_f^2 D^2}{\epsilon^2 (1-\alpha)^2} (1 - \alpha)^{2j}.$$

Therefore,

$$N = \sum_{j=0}^{m} n(j) \le \frac{M_f^2 D^2}{\epsilon^2(1-\alpha)^2} \sum_{j=0}^{m} (1-\alpha)^{2j} \le \frac{M_f^2 D^2}{\epsilon^2(1-\alpha)^2(1-(1-\alpha)^2)}. \qquad \square$$

Let us discuss the above efficiency estimate. Note that we can obtain the optimal value of the level parameter $\alpha$ from the following maximization problem:

$$(1-\alpha)^2(1-(1-\alpha)^2) \quad \to \quad \max_{\alpha \in [0,1]} .$$

Its solution is $\alpha^* = \frac{1}{2+\sqrt{2}} \approx 0.2929$. Under this choice, we have the following efficiency bound of the Level Method:

$$N \le \frac{4}{\epsilon^2} M_f^2 D^2.$$

Comparing this result with Theorem 3.2.1, we see that Level Method is optimal *uniformly* in the dimension of the space of variables. Note that the analytical complexity bound of this method in *finite dimensions* is not known.

One of the advantages of this method is that the gap $\delta_k = f_k^* - \hat{f}_k^*$ provides us with an *exact* estimate of the current accuracy. Usually, this gap converges to zero much faster than in the worst case situation. For the majority of real-life optimization problems, the accuracy $\epsilon = 10^{-4} - 10^{-5}$ is obtained by the method after $3n$ to $4n$ iterations.

### 3.3.4 Constrained Minimization

Let us show how to use piece-wise linear models to solve constrained minimization problems. Consider the problem

$$\begin{gathered} \min_{x \in Q} \quad f(x), \\ \text{s.t. } f_j(x) \le 0, \;\; j = 1 \ldots m, \end{gathered} \tag{3.3.4}$$

where $Q$ is a bounded closed convex set, and functions $f(\cdot)$, $f_j(\cdot)$ are Lipschitz continuous on $Q$.

Let us rewrite this problem as a problem with a single functional constraint. Define $\bar{f}(x) = \max\limits_{1 \le j \le m} f_j(x)$. Then we obtain the equivalent problem

$$\begin{gathered} \min_{x \in Q} \quad f(x), \\ \text{s.t. } \bar{f}(x) \le 0. \end{gathered} \tag{3.3.5}$$

Note that the functions $f(\cdot)$ and $\bar{f}(\cdot)$ are convex and Lipschitz continuous. In this section, we will try to solve (3.3.5) using the models for both of them.

Let us define the corresponding models. Consider a sequence $X = \{x_k\}_{k=0}^{\infty}$. Define

$$\hat{f}_k(X;x) = \max_{0\le j\le k} [f(x_j) + \langle g(x_j), x - x_j\rangle] \le f(x),$$

$$\check{f}_k(X;x) = \max_{0\le j\le k} [\bar{f}(x_j) + \langle \bar{g}(x_j), x - x_j\rangle] \le \bar{f}(x),$$

where $g(x_j) \in \partial f(x_j)$ and $\bar{g}(x_j) \in \partial \bar{f}(x_j)$.

As in Sect. 2.3.4, our scheme is based on the *parametric* function

$$f(t;x) = \max\{f(x) - t, \bar{f}(x)\},$$

$$f^*(t) = \min_{x\in Q} f(t;x).$$

Recall that $f^*(t)$ is nonincreasing in $t$. Let $x^*$ be a solution to (3.3.5). Let $t^* = f(x^*)$. Then $t^*$ is the smallest root of thte function $f^*(t)$.

Using the models for the objective function and the constraint, we can introduce a model for the parametric function. Define

$$f_k(X;t,x) = \max\{\hat{f}_k(X;x) - t, \check{f}_k(X;x)\} \le f(t;x),$$

$$\hat{f}_k^*(X;t) = \min_{x\in Q} f_k(X;t,x) \le f^*(t).$$

Again, $\hat{f}_k^*(X;t)$ is nonincreasing in $t$. It is clear that its smallest root $t_k^*(X)$ does not exceed $t^*$.

We will need the following characterization of the root $t_k^*(X)$.

**Lemma 3.3.4**

$$t_k^*(X) = \min_{x\in Q}\{\hat{f}_k(X;x) \mid \ \check{f}_k(X;x) \le 0\}.$$

*Proof* Denote by $\hat{x}_k^*$ the solution of the minimization problem in the above equation and let $\hat{t}_k^* = \hat{f}_k(X;\hat{x}_k^*)$ be its optimal value. Then

$$\hat{f}_k^*(X;\hat{t}_k^*) \le \max\{\hat{f}_k(X;\hat{x}_k^*) - \hat{t}_k^*, \check{f}_k(X;\hat{x}_k^*)\} \le 0.$$

Thus, we always have $\hat{t}_k^* \ge t_k^*(X)$.

Assume that $\hat{t}_k^* > t_k^*(X)$. Then there exists a point $y$ such that

$$\hat{f}_k(X; y) - t_k^*(X) \le 0, \quad \check{f}_k(X; y) \le 0.$$

However, in this case $\hat{t}_k^* = \hat{f}_k(X; \hat{x}_k^*) \le \hat{f}_k(X; y) \le t_k^*(X) < \hat{t}_k^*$. This is a contradiction. □

In our analysis, we will also need the function

$$f_k^*(X; t) = \min_{0 \le j \le k} f_k(X; t, x_j),$$

the *record value* of our parametric model.

**Lemma 3.3.5** *Let $t_0 < t_1 \le t^*$. Assume that $\hat{f}_k^*(X; t_1) > 0$. Then $t_k^*(X) > t_1$ and*

$$\hat{f}_k^*(X; t_0) \ge \hat{f}_k^*(X; t_1) + \tfrac{t_1 - t_0}{t_k^*(X) - t_1} \hat{f}_k^*(X; t_1). \tag{3.3.6}$$

*Proof* Let $x_k^*(t) \in \operatorname{Arg\,min} f_k(X; t, x)$, $t_2 = t_k^*(X)$, $\alpha = \frac{t_1 - t_0}{t_2 - t_0} \in [0, 1]$. Then

$$t_1 = (1 - \alpha) t_0 + \alpha t_2$$

and inequality (3.3.6) is equivalent to the following:

$$\hat{f}_k^*(X; t_1) \le (1 - \alpha) \hat{f}_k^*(X; t_0) + \alpha \hat{f}_k^*(X; t_2) \tag{3.3.7}$$

(note that $\hat{f}_k^*(X; t_2) = 0$). Let $x_\alpha = (1 - \alpha) x_k^*(t_0) + \alpha x_k^*(t_2)$. Then we have

$$\hat{f}_k^*(X; t_1) \le \max\{\hat{f}_k(X; x_\alpha) - t_1; \check{f}_k(X; x_\alpha)\}$$

$$\le \max\{(1 - \alpha)(\hat{f}_k(X; x_k^*(t_0)) - t_0) + \alpha(\hat{f}_k(X; x_k^*(t_2)) - t_2);$$

$$(1 - \alpha) \check{f}_k(X; x_k^*(t_0)) + \alpha \check{f}_k(X; x_k^*(t_2))\}$$

$$\le (1 - \alpha) \max\{\hat{f}_k(X; x_k^*(t_0)) - t_0; \check{f}_k(X; x_k^*(t_0))\}$$

$$+ \alpha \max\{\hat{f}_k(X; x_k^*(t_2)) - t_2; \check{f}_k(X; x_k^*(t_2))\}$$

$$= (1 - \alpha) \hat{f}_k^*(X; t_0) + \alpha \hat{f}_k^*(X; t_2),$$

and we get (3.3.7). □

We also need the following statement (compare with Lemma 2.3.5).

**Lemma 3.3.6** *For any $\Delta \geq 0$, we have*

$$f^*(t) - \Delta \leq f^*(t+\Delta),$$

$$\hat{f}_k^*(X;t) - \Delta \leq \hat{f}_k^*(X;t+\Delta).$$

*Proof* Indeed, for $f^*(t)$ we have

$$\begin{aligned} f^*(t+\Delta) &= \min_{x\in Q}\,[\max\{f(x)-t;\ \bar{f}(x)+\Delta\}-\Delta] \\ &\geq \min_{x\in Q}\,[\max\{f(x)-t;\ \bar{f}(x)\}-\Delta] = f^*(t)-\Delta. \end{aligned}$$

The proof of the second inequality is similar. □

Now we are ready to present a constrained minimization scheme (compare with the constrained minimization scheme of Sect. 2.3.5).

**Constrained Level Method**

**0.** Choose $x_0 \in Q$, $t_0 < t^*$, $\varkappa \in (0, \frac{1}{2})$, and accuracy $\epsilon > 0$.

**1.** ***k*th iteration ($k \geq 0$).**

(a) Keep generating the sequence $X = \{x_j\}_{j=0}^{\infty}$ by the Level Method as applied to the function $f(t_k;x)$. If the internal termination criterion

$$\hat{f}_j^*(X;t_k) \geq (1-\varkappa) f_j^*(X;t_k)$$

holds, then stop the internal process and set $j(k) = j$.

**Global stop:** $f_j^*(X;t_k) \leq \epsilon$.

(b) Set $t_{k+1} = t_{j(k)}^*(X)$.

(3.3.8)

We are interested in the analytical complexity of this method. Therefore, the complexity of the computation of the root $t_j^*(X)$ and of the value $\hat{f}_j^*(X;t)$ is not important for us now. We need to estimate the rate of convergence of the *master process* and the complexity of Step 1(a).

Let us start from the master process.

**Lemma 3.3.7** *For all $k \geq 0$, we have*

$$f_{j(k)}^*(X;t_k) \leq \tfrac{t_0-t^*}{1-\varkappa}\left[\tfrac{1}{2(1-\varkappa)}\right]^k.$$

*Proof* Define

$$\sigma_k = \frac{f^*_{j(k)}(X;t_k)}{\sqrt{t_{k+1}-t_k}}, \quad \beta = \frac{1}{2(1-\varkappa)} \quad (< 1).$$

Since $t_{k+1} = t^*_{j(k)}(X)$, in view of Lemma 3.3.5, for all $k \geq 1$, we have

$$\sigma_{k-1} = \frac{1}{\sqrt{t_k - t_{k-1}}} f^*_{j(k-1)}(X; t_{k-1}) \geq \frac{1}{\sqrt{t_k - t_{k-1}}} \hat{f}^*_{j(k)}(X; t_{k-1})$$

$$\geq \frac{2}{\sqrt{t_{k+1}-t_k}} \hat{f}^*_{j(k)}(X; t_k) \geq \frac{2(1-\varkappa)}{\sqrt{t_{k+1}-t_k}} f^*_{j(k)}(X; t_k) = \frac{\sigma_k}{\beta}.$$

Thus, $\sigma_k \leq \beta \sigma_{k-1}$ and we obtain

$$f^*_{j(k)}(X; t_k) = \sigma_k \sqrt{t_{k+1} - t_k} \leq \beta^k \sigma_0 \sqrt{t_{k+1} - t_k}$$

$$= \beta^k f^*_{j(0)}(X; t_0) \sqrt{\frac{t_{k+1}-t_k}{t_1 - t_0}}.$$

Further, in view of Lemma 3.3.6, $t_1 - t_0 \geq \hat{f}^*_{j(0)}(X; t_0)$. Therefore,

$$f^*_{j(k)}(X; t_k) \leq \beta^k f^*_{j(0)}(X; t_0) \sqrt{\frac{t_{k+1}-t_k}{\hat{f}^*_{j(0)}(X;t_0)}} \leq \frac{\beta^k}{1-\varkappa} \sqrt{\hat{f}^*_{j(0)}(X; t_0)(t_{k+1} - t_k)}$$

$$\leq \frac{\beta^k}{1-\varkappa} \sqrt{f^*(t_0)(t_0 - t^*)}.$$

It remains to note that $f^*(t_0) \leq t_0 - t^*$ (see Lemma 3.3.6). □

Let the Global Stop condition in (3.3.8) be satisfied: $f^*_j(X; t_k) \leq \epsilon$. Then there exists a $j^*$ such that

$$f(t_k; x_{j^*}) = f^*_j(X; t_k) \leq \epsilon.$$

Therefore, we have

$$f(t_k; x_{j^*}) = \max\{f(x_{j^*}) - t_k;\ \bar{f}(x_{j^*})\} \leq \epsilon.$$

Since $t_k \leq t^*$, we conclude that

$$\begin{aligned} f(x_{j^*}) &\leq t^* + \epsilon, \\ \bar{f}(x_{j^*}) &\leq \epsilon. \end{aligned} \tag{3.3.9}$$

In view of Lemma 3.3.7, we can get (3.3.9) at most in

$$N(\epsilon) = \frac{1}{\ln[2(1-\varkappa)]} \ln \frac{t_0 - t^*}{(1-\varkappa)\epsilon}$$

*full* iterations of the master process. (The last iteration of the process is terminated by the Global Stop rule.) Note that in the above expression, $\varkappa$ is an absolute constant (for example, we can take $\varkappa = \frac{1}{4}$).

Let us estimate the complexity of the internal process. Define

$$M_f = \max\{\| \; g \; \| \mid \; g \in \partial f(x) \bigcup \partial \bar{f}(x), \; x \in Q\}.$$

We need to analyze two cases.

1. *Full step.* At this step, the internal process is terminated by the rule

$$\hat{f}^*_{j(k)}(X; t_k) \geq (1 - \varkappa) f^*_{j(k)}(X; t_k).$$

The corresponding inequality for the *gap* is as follows:

$$f^*_{j(k)}(X; t_k) - \hat{f}^*_{j(k)}(X; t_k) \leq \varkappa f^*_{j(k)}(X; t_k).$$

In view of Theorem 3.3.1, this happens at most after

$$\frac{M_f^2 D^2}{\varkappa^2 (f^*_{j(k)}(X;t_k))^2 \alpha(1-\alpha)^2(2-\alpha)}$$

iterations of the internal process. Since at the full step $f^*_{j(k)}(X; t_k)) \geq \epsilon$, we conclude that

$$j(k) - j(k-1) \leq \frac{M_f^2 D^2}{\varkappa^2 \epsilon^2 \alpha(1-\alpha)^2(2-\alpha)}$$

for any full iteration of the master process.

2. *Last step.* The internal process of this step was terminated by the Global Stop rule:

$$f^*_j(X; t_k) \leq \epsilon.$$

Since the normal stopping criterion did not work, we conclude that

$$f^*_{j-1}(X; t_k) - \hat{f}^*_{j-1}(X; t_k) \geq \varkappa f^*_{j-1}(X; t_k) \geq \varkappa\epsilon.$$

Therefore, in view of Theorem 3.3.1, the number of iterations at the last step does not exceed

$$\frac{M_f^2 D^2}{\varkappa^2 \epsilon^2 \alpha(1-\alpha)^2(2-\alpha)}.$$

Thus, we come to the following estimate of *total* complexity of the Constrained Level Method:

$$
\begin{aligned}
&(N(\epsilon)+1)\tfrac{M_f^2 D^2}{\varkappa^2\epsilon^2\alpha(1-\alpha)^2(2-\alpha)}\\
&= \tfrac{M_f^2 D^2}{\varkappa^2\epsilon^2\alpha(1-\alpha)^2(2-\alpha)}\left[1+\tfrac{1}{\ln[2(1-\varkappa)]}\ln\tfrac{t_0-t^*}{(1-\varkappa)\epsilon}\right]\\
&= \tfrac{M_f^2 D^2 \ln\frac{2(t_0-t^*)}{\epsilon}}{\epsilon^2\alpha(1-\alpha)^2(2-\alpha)\varkappa^2\ln[2(1-\varkappa)]}.
\end{aligned}
$$

A reasonable choice for the parameters of this scheme is $\alpha=\varkappa=\frac{1}{2+\sqrt{2}}$.

The principal term in the above complexity bound is of the order $O(\frac{1}{\epsilon^2}\ln\frac{2(t_0-t^*)}{\epsilon})$. Thus, the Constrained Level Method is suboptimal (see Theorem 3.2.1).

In this method, at each iteration of the master process we need to find the root $t^*_{j(k)}(X)$. In view of Lemma 3.3.4, this is equivalent to the following problem:

$$
\min_{x\in Q}\{\hat{f}_k(X;x) \mid \ \check{f}_k(X;x)\le 0\}.
$$

In other words, we need to solve the problem

$$
\begin{aligned}
&\min_{x\in Q} \quad t,\\
&\text{s.t. } f(x_j)+\langle g(x_j),x-x_j\rangle \le t,\ \ j=0\ldots k,\\
&\qquad \bar{f}(x_j)+\langle \bar{g}(x_j),x-x_j\rangle \le 0,\ j=0\ldots k.
\end{aligned}
$$

If $Q$ is a polytope, this problem can be solved by finite linear programming methods (simplex method). If $Q$ is more complicated, we can use Interior-Point Schemes (Chap. 5).

To conclude this section, let us note that we can use a better model for the functional constraints. Since

$$
\bar{f}(x)=\max_{1\le i\le m} f_i(x),
$$

it is possible to work with

$$
\check{f}_k(X;x)=\max_{0\le j\le k}\ \max_{1\le i\le m}[f_i(x_j)+\langle g_i(x_j),x-x_j\rangle],
$$

where $g_i(x_j) \in \partial f_i(x_j)$. In practice, this *complete* model significantly accelerates the convergence of the process. However, clearly each iteration becomes much more expensive.

As far as the practical behavior of this scheme is concerned, we note that usually the process is very fast. There are some technical problems related to the accumulation of many linear pieces in the model. However, in all practical implementations of the Level Method there exist some strategies for dropping the old inactive elements of the model.

# Chapter 4
# Second-Order Methods

In this chapter, we study Black-Box second-order methods. In the first two sections, these methods are based on cubic regularization of the second-order model of the objective function. With an appropriate proximal coefficient, this model becomes a global upper approximation of the objective function. At the same time, the global minimum of this approximation is computable in polynomial time even if the Hessian of the objective is not positive semidefinite. We study global and local convergence of the Cubic Newton Method in convex and non-convex cases. In the next section, we derive the lower complexity bounds and show that this method can be accelerated using the estimating sequences technique. In the last section, 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.

## 4.1 Cubic Regularization of Newton's Method

(Cubic regularization of quadratic approximation; General convergence results; Global rate of convergence for different problem classes; Implementation issues; Complexity results for strongly convex functions.)

### *4.1.1 Cubic Regularization of Quadratic Approximation*

In this section, we consider the simplest unconstrained minimization problem

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

Y. Nesterov, *Lectures on Convex Optimization*, Springer Optimization
and Its Applications 137, https://doi.org/10.1007/978-3-319-91578-4_4

with a twice continuously differentiable objective function. The standard second-order scheme for this problem, Newton's method, is as follows:

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k). \qquad (4.1.1)$$

We have already looked at this method in Sect. 1.2.

Despite its very natural motivation, this scheme has several hidden drawbacks. First of all, it may happen that at the current test point the Hessian is degenerate; in this case the method is not well-defined. Secondly, it may happen that this scheme diverges or converges to a saddle point or even to a point of local maximum. In order to overcome these difficulties, there are three standard recipes.

- *Levenberg–Marquardt regularization*. If $\nabla^2 f(x_k)$ is indefinite, let us regularize it with a unit matrix. Namely, use the matrix $G_k = \nabla^2 f(x_k) + \gamma I_n \succ 0$ in order to perform the step:

  $$x_{k+1} = x_k - G_k^{-1} \nabla f(x_k).$$

  This strategy is sometimes considered as a way of mixing Newton's method with the gradient method.
- *Line search.* Since we are interested in minimization, it is reasonable to introduce in method (4.1.1) a certain step size $h_k > 0$:

  $$x_{k+1} = x_k - h_k [\nabla^2 f(x_k)]^{-1} \nabla f(x_k).$$

  (This is a *damped* Newton method. Compare with the scheme (5.1.28).) This can help in generating a monotone sequence of function values: $f(x_{k+1}) \leq f(x_k)$.
- *Trust-region methods*. In accordance with this approach, at point a $x_k$ we have to define a neighborhood, where the second-order approximation of the objective function is reliable. This is a certain trust region $\Delta(x_k)$. For example, we can take

  $$\Delta(x_k) = \{x : \|x - x_k\| \leq \epsilon\}$$

  with some $\epsilon > 0$. Then the next point $x_{k+1}$ can be chosen as a solution to the following auxiliary problem:

  $$\min_{x \in \Delta(x_k)} \left[ \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle \right].$$

  Note that for $\Delta(x_k) \equiv \mathbb{R}^n$, this is exactly the standard Newton step.

Unfortunately, none of these approaches seems to be useful in addressing the global behavior of second-order schemes. In this section, we present a modification of Newton's method, which is constructed in a similar way as the *Gradient Mapping* (see Sect. 2.2.4).

Let $\mathscr{F} \subseteq \mathbb{R}^n$ be an open convex set. Consider a function $f$ which is twice differentiable on $\mathscr{F}$. Let $x_0 \in \mathscr{F}$ be a starting point of our iterative scheme. We assume that the set $\mathscr{F}$ is large enough: It contains at least the level set

$$\mathscr{L}(f(x_0)) \equiv \{x \in \mathbb{R}^n : \; f(x) \le f(x_0).\}$$

Moreover, in this section we always assume the following.

**Assumption 4.1.1** *The Hessian of the function $f$ is Lipschitz continuous on $\mathscr{F}$:*

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L\|x-y\|, \quad \forall x, y \in \mathscr{F}, \tag{4.1.2}$$

*with some constant $L > 0$. In this section, the norm is always standard Euclidean.*

For the reader's convenience, let us recall the following variant of Lemma 1.2.4.

**Lemma 4.1.1** *For any $x$ and $y$ from $\mathscr{F}$ we have*

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x)\| \overset{(1.2.13)}{\le} \frac{1}{2}L\|y-x\|^2, \tag{4.1.3}$$

$$|f(y) - f(x) - \langle \nabla f(x), y-x\rangle - \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle| \overset{(1.2.14)}{\le} \tfrac{L}{6}\|y-x\|^3. \tag{4.1.4}$$

Let $M$ be a positive parameter. Define a modified Newton step by minimizing a *cubic regularization* of the quadratic approximation of the function $f$:

$$\min_y \left[\langle \nabla f(x), y-x\rangle + \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle + \tfrac{M}{6}\|y-x\|^3\right]. \tag{4.1.5}$$

Denote by $T_M(x)$ and arbitrary point from the set of global minima of this minimization problem. We postpone the discussion of computational complexity of finding this point up to Sect. 4.1.4.1.

Note that point $T_M(x)$ satisfies the following first-order optimality condition:

$$\nabla f(x) + \nabla^2 f(x)(T_M(x)-x) + \tfrac{M}{2}\|T_M(x)-x\| \cdot (T_M(x)-x) \overset{(1.2.4)}{=} 0. \tag{4.1.6}$$

Let $r_M(x) = \|x - T_M(x)\|$. Multiplying (4.1.6) by $T_M(x)-x$, we get the equation

$$\langle \nabla f(x), T_M(x)-x\rangle + \langle \nabla^2 f(x)(T_M(x)-x), T_M(x)-x\rangle + \tfrac{M}{2}r_M^3(x) = 0. \tag{4.1.7}$$

In our analysis of the process (4.1.16), we need the following fact.

**Lemma 4.1.2** *For any $x \in \mathscr{F}$, we have*

$$\nabla^2 f(x) + \tfrac{M}{2} r_M(x) I_n \succeq 0. \tag{4.1.8}$$

This statement will be justified later in Sect. 4.1.4.1. Let us now present the main properties of the vector function $T_M(\cdot)$.

**Lemma 4.1.3** *For any $x \in \mathscr{L}(f(x_0)$, we have the following relation:*

$$\langle \nabla f(x), x - T_M(x) \rangle \geq 0. \tag{4.1.9}$$

*If $M > \frac{2}{3}L$ and $x \in$ int $\mathscr{F}$, then $T_M(x) \in \mathscr{L}(f(x)) \subset \mathscr{F}$.*

*Proof* Indeed, multiplying (4.1.8) by $x - T_M(x)$ twice, we get

$$\langle \nabla^2 f(x)(T_M(x) - x), T_M(x) - x \rangle + \tfrac{M}{2} r_M^3(x) \geq 0.$$

Therefore, (4.1.9) follows from (4.1.7).

Further, let $M > \frac{2}{3}L$. Assume that $T_M(x) \notin \mathscr{F}$. Then $r_M(x) > 0$. Consider the following points:

$$y(\alpha) = x + \alpha(T_M(x) - x), \quad \alpha \in [0, 1].$$

Since $y(0) \in \mathscr{F}$, the value

$$\bar{\alpha} : \; y(\bar{\alpha}) \in \partial \mathrm{cl}\; (\mathscr{F})$$

is well defined. In accordance with our assumption, $\bar{\alpha} \leq 1$ and $y(\alpha) \in \mathscr{F}$ for all $\alpha \in [0, \bar{\alpha})$. Therefore, using (4.1.4), relation (4.1.7), and inequality (4.1.9), we get

$$\begin{aligned}
f(y(\alpha)) &\leq f(x) + \langle \nabla f(x), y(\alpha) - x \rangle \\
&\quad + \tfrac{1}{2}\langle \nabla^2 f(x)(y(\alpha) - x), y(\alpha) - x \rangle + \tfrac{\alpha^3 L}{6} r_M^3(x) \\
&= f(x) + \langle \nabla f(x), y(\alpha) - x \rangle \\
&\quad + \tfrac{1}{2}\langle \nabla^2 f(x)(y(\alpha) - x), y(\alpha) - x \rangle + \tfrac{\alpha^3 M}{4} r_M^3(x) - \alpha^3 \delta \\
&= f(x) + (\alpha - \tfrac{\alpha^2}{2})\langle \nabla f(x), T_M(x) - x \rangle - \tfrac{\alpha^2(1-\alpha)}{4} M r_M^3(x) - \alpha^3 \delta \\
&\leq f(x) - \tfrac{\alpha^2(1-\alpha)}{4} M r_M^3(x) - \alpha^3 \delta,
\end{aligned}$$

where $\delta = \left(\frac{M}{4} - \frac{L}{6}\right) r_M^3(x) > 0$. Thus, $f(y(\bar{\alpha})) < f(x)$. Therefore $y(\bar{\alpha}) \in \mathscr{L}(f(x)) \subset \mathscr{F}$. This is a contradiction. Hence, $T_M(x) \in \mathscr{F}$. Using the same arguments, we prove that $f(T_M(x)) \le f(x)$. □

**Lemma 4.1.4** *If $T_M(x) \in \mathscr{F}$, then*

$$\|\nabla f(T_M(x))\| \le \frac{1}{2}(L+M) r_M^2(x). \tag{4.1.10}$$

*Proof* From Eq. (4.1.6), we get

$$\|\nabla f(x) + \nabla^2 f(x)(T_M(x) - x)\| = \frac{1}{2} M r_M^2(x).$$

On the other hand, in view of (4.1.3), we have

$$\|\nabla f(T_M(x)) - \nabla f(x) - \nabla^2 f(x)(T_M(x) - x)\| \le \frac{1}{2} L r_M^2(x).$$

Combining these two relations, we obtain inequality (4.1.10). □

Define

$$\bar{f}_M(x) = \min_y \left[ f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle + \tfrac{M}{6}\|y-x\|^3 \right].$$

**Lemma 4.1.5** *For any $x \in \mathscr{F}$, we have*

$$\bar{f}_M(x) \le \min_{y \in \mathscr{F}} \left[ f(y) + \tfrac{L+M}{6}\|y-x\|^3 \right], \tag{4.1.11}$$

$$f(x) - \bar{f}_M(x) \ge \tfrac{M}{12} r_M^3(x). \tag{4.1.12}$$

*Moreover, if $M \ge L$, then $T_M(x) \in \mathscr{F}$ and*

$$f(T_M(x)) \le \bar{f}_M(x). \tag{4.1.13}$$

*Proof* Indeed, using the lower bound in (4.1.4), for any $y \in \mathscr{F}$ we have

$$f(x) + \langle \nabla f(x), y-x\rangle + \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle \le f(y) + \tfrac{L}{6}\|y-x\|^3,$$

and inequality in (4.1.11) follows from the definition of $\bar{f}_M(x)$.

Further, in view of the definition of the point $T_M(x)$, relation (4.1.7), and inequality (4.1.9), we have

$$\begin{aligned} f(x) - \bar{f}_M(x) &= \langle \nabla f(x), x - T_M(x) \rangle \\ &\quad -\tfrac{1}{2}\langle \nabla^2 f(x)(T_M(x) - x), T_M(x) - x \rangle - \tfrac{M}{6} r_M^3(x) \\ &= \tfrac{1}{2}\langle \nabla f(x), x - T_M(x) \rangle + \tfrac{M}{12} r_M^3(x) \ \geq \ \tfrac{M}{12} r_M^3(x). \end{aligned}$$

Finally, if $M \geq L$, then $T_M(x) \in \mathscr{F}$ in view of Lemma 4.1.3. Therefore, we get inequality (4.1.13) from the upper bound in (4.1.4). $\square$

### 4.1.2 General Convergence Results

In this section, our main problem of interest is as follows:

$$\min_{x \in \mathbb{R}^n} f(x), \tag{4.1.14}$$

where the objective function $f(\cdot)$ satisfies Assumption 4.1.1. Recall that the necessary conditions for a point $x^*$ to be a local minimum of problem (4.1.14) are as follows (see Theorem 1.2.2):

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \succeq 0. \tag{4.1.15}$$

Therefore, for arbitrary $x \in \mathscr{F}$, we can introduce the following measure of local optimality:

$$\mu_M(x) = \max \left\{ \sqrt{\tfrac{2}{L+M} \| \nabla f(x) \|}, -\tfrac{2}{2L+M} \lambda_{\min}(\nabla^2 f(x)) \right\},$$

where $M$ is a positive parameter, and $\lambda_{\min}(\cdot)$ is the minimal eigenvalue of the corresponding matrix. It is clear that for any $x$ from $\mathscr{F}$ the measure $\mu_M(x)$ is non-negative and it vanishes only at the points satisfying conditions (4.1.15). The analytical form of this measure can be justified by the following result.

**Lemma 4.1.6** *For any $x \in \mathscr{F}$ we have $\mu_M(T_M(x)) \leq r_M(x)$.*

*Proof* The proof follows immediately from inequality (4.1.10) and relation (4.1.8) since

$$\nabla^2 f(T_M(x)) \succeq \nabla^2 f(x) - L r_M(x) I \succeq -(\tfrac{1}{2} M + L) r_M(x) I. \qquad \square$$

Let $L_0 \in (0, L]$ be a positive parameter. Consider the following regularized Newton method.

```
Cubic Regularization of Newton's Method

Initialization: Choose $x_0 \in \mathbb{R}^n$.

Iteration $k$, ($k \geq 0$):

    1. Find $M_k \in [L_0, 2L]$ such that $f(T_{M_k}(x_k)) \leq \bar{f}_{M_k}(x_k)$.

    2. Set $x_{k+1} = T_{M_k}(x_k)$.
```

(4.1.16)

Since $\bar{f}_M(x) \leq f(x)$, this process is monotone:

$$f(x_{k+1}) \leq f(x_k).$$

If the constant $L$ is known, in Step 1 of this scheme we can take $M_k \equiv L$. In the opposite case, it is possible to apply a simple search procedure; we will discuss its complexity later in Sect. 4.1.4.2.

Let us start from the following simple observation.

**Theorem 4.1.1** *Let the sequence* $\{x_i\}$ *be generated by method (4.1.16). Assume that the objective function* $f(\cdot)$ *is bounded below:*

$$f(x) \geq f^* \quad \forall x \in \mathscr{F}.$$

*Then* $\sum\limits_{i=0}^{\infty} r^3_{M_i}(x_i) \leq \frac{12}{L_0}(f(x_0) - f^*)$. *Hence,* $\lim\limits_{i \to \infty} \mu_L(x_i) = 0$ *and for any* $k \geq 1$ *we have*

$$\min_{1 \leq i \leq k} \mu_L(x_i) \leq \frac{8}{3} \cdot \left( \frac{3(f(x_0) - f^*)}{2k \cdot L_0} \right)^{1/3}. \tag{4.1.17}$$

*Proof* In view of inequality (4.1.12), we have

$$f(x_0) - f^* \geq \sum_{i=0}^{k-1} [f(x_i) - f(x_{i+1}) \geq \sum_{i=0}^{k-1} \frac{M_i}{12} r^3_{M_i}(x_i) \geq \frac{L_0}{12} \sum_{i=0}^{k-1} r^3_{M_i}(x_i).$$

It remains to use the statement of Lemma 4.1.6 and the upper bound on $M_k$ at Step 1 in (4.1.16):

$$r_{M_i}(x_i) \geq \mu_{M_i}(x_{i+1}) \;\geq\; \tfrac{3}{4}\mu_L(x_{i+1}). \qquad \square$$

Note that inequality (4.1.17) implies that

$$\min_{1\leq i\leq k} \|\nabla f(x_i)\| \leq O(k^{-2/3}).$$

We have seen that for a gradient scheme, the right-hand side in this inequality can be of the order $O\left(k^{-1/2}\right)$ (see inequality (1.2.24)).

Theorem 4.1.1 helps us to get convergence results in many different situations. We mention only one of them.

**Theorem 4.1.2** *Let the sequence $\{x_i\}$ be generated by method (4.1.16). Let us assume that for some $i \geq 0$ the set $\mathscr{L}(f(x_i))$ is bounded. Then there exists a limit*

$$\lim_{i\to\infty} f(x_i) = f^*.$$

*The set $X^*$ of limit points of this sequence is non-empty. Moreover, this is a connected set such that for any $x^* \in X^*$ we have*

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

*Proof* The proof of this theorem can be derived from Theorem 4.1.1 in a standard way. □

Let us describe now the behavior of the process (4.1.16) in a neighborhood of a non-degenerate stationary point, which is not a point of local minimum.

**Lemma 4.1.7** *Let $\bar{x} \in \mathscr{F}$ be a non-degenerate saddle point or a point of local maximum of the function $f(\cdot)$:*

$$\nabla f(\bar{x}) = 0, \quad \lambda_{\min}(\nabla^2 f(\bar{x})) < 0.$$

*Then there exist constants $\epsilon, \delta > 0$ such that whenever the point $x_i$ appears to be in the set $Q = \{x : \|x - \bar{x}\| \leq \epsilon,\ f(x) \geq f(\bar{x})\}$ (for instance, if $x_i = \bar{x}$), then the next point $x_{i+1}$ leaves the set $Q$:*

$$f(x_{i+1}) \leq f(\bar{x}) - \delta.$$

*Proof* Let us choose a direction $d$, $\|d\| = 1$, with negative curvature:

$$\langle \nabla^2 f(\bar{x})d, d\rangle \equiv -2\sigma < 0.$$

And let $\bar{\tau} > 0$ be small enough: $\bar{x} \pm \bar{\tau} d \in \mathscr{F}$. Define $\epsilon = \min\left\{\frac{\sigma}{2L}, \bar{\tau}\right\}$ and $\delta = \frac{\sigma}{6}\epsilon^2$. Then, in view of inequality (4.1.11), upper bound on $M_i$, and inequality (4.1.4), for $|\tau| \leq \bar{\tau}$ we get the following estimate

$$\begin{aligned} f(x_{i+1}) &\leq f(\bar{x} + \tau d) + \tfrac{L}{2}\|\bar{x} + \tau d - x_i\|^3 \\ &\leq f(\bar{x}) - \sigma\tau^2 + \tfrac{L}{6}|\tau|^3 + \tfrac{L}{2}\left[\epsilon^2 + 2\tau\langle d, \bar{x} - x_i\rangle + \tau^2\right]^{3/2}. \end{aligned}$$

Since we are free in the choice of the sign of $\tau$, we can guarantee that

$$f(x_{i+1}) \leq f(\bar{x}) - \sigma\tau^2 + \tfrac{L}{6}|\tau|^3 + \tfrac{L}{2}\left[\epsilon^2 + \tau^2\right]^{3/2}, \quad |\tau| \leq \bar{\tau}.$$

Let us choose $|\tau| = \epsilon \leq \bar{\tau}$. Then

$$f(x_{i+1}) \leq f(\bar{x}) - \sigma\tau^2 + \tfrac{5L}{3}|\tau|^3 \leq f(\bar{x}) - \sigma\tau^2 + \tfrac{5L}{3} \cdot \tfrac{\sigma}{2L} \cdot \tau^2 = f(\bar{x}) - \tfrac{1}{6}\sigma\tau^2.$$

Since the process (4.1.16) is monotone with respect to the objective function, it will never return to $Q$. □

Consider now the behavior of the regularized Newton scheme (4.1.16) in a neighborhood of a non-degenerate local minimum. It appears that in such a situation, condition $L_0 > 0$ is no longer necessary. Let us analyze a relaxed version of (4.1.16):

$$\boxed{x_{k+1} = T_{M_k}(x_k), \;\; k \geq 0} \tag{4.1.18}$$

where $M_k \in (0, 2L]$. Define

$$\delta_k = \frac{L\|\nabla f(x_k)\|}{\lambda^2_{\min}(\nabla^2 f(x_k))}.$$

**Theorem 4.1.3** *Let $\nabla^2 f(x_0) \succ 0$ and $\delta_0 \leq \frac{1}{4}$. Let the points $\{x_k\}$ be generated by method (4.1.18). Then:*

1. *For all $k \geq 0$ the values $\delta_k$ are well defined and converge quadratically to zero:*

$$\delta_{k+1} \leq \tfrac{3}{2}\left(\tfrac{\delta_k}{1-\delta_k}\right)^2 \leq \tfrac{8}{3}\delta_k^2 \leq \tfrac{2}{3}\delta_k, \quad k \geq 0. \tag{4.1.19}$$

2. *Minimal eigenvalues of all Hessians $\nabla^2 f(x_k)$ satisfy the following bounds:*

$$e^{-1}\lambda_{\min}(\nabla^2 f(x_0)) \leq \lambda_{\min}(\nabla^2 f(x_k)) \leq e^{3/4}\lambda_{\min}(\nabla^2 f(x_0)). \tag{4.1.20}$$

3. *The whole sequence $\{x_i\}$ converges quadratically to a point $x^*$, which is a non-degenerate local minimum of the function $f$. In particular, for any $k \geq 1$ we have*

$$\|\nabla f(x_k)\| \leq \lambda_{\min}^2(\nabla^2 f(x_0)) \, \tfrac{9e^{3/2}}{16L} \left(\tfrac{1}{2}\right)^{2^k}. \tag{4.1.21}$$

*Proof* Assume that $\nabla^2 f(x_k) \succ 0$ for some $k \geq 0$. Then the corresponding $\delta_k$ is well defined. Assume that $\delta_k \leq \frac{1}{4}$. From Eq. (4.1.6), we have

$$\begin{aligned} r_{M_k}(x_k) &= \|T_{M_k}(x_k) - x_k\| = \|(\nabla^2 f(x_k) + r_{M_k}(x_k)\frac{M_k}{2} I_n)^{-1} \nabla f(x_k)\| \\ &\leq \frac{\|\nabla f(x_k)\|}{\lambda_{\min}(\nabla^2 f(x_k))} \; = \; \frac{1}{L}\lambda_{\min}(\nabla^2 f(x_k))\delta_k. \end{aligned} \tag{4.1.22}$$

Note also that $\nabla^2 f(x_{k+1}) \overset{(4.1.2)}{\succeq} \nabla^2 f(x_k) - r_{M_k}(x_k) L \, I_n$. Therefore,

$$\begin{aligned} \lambda_{\min}(\nabla^2 f(x_{k+1})) &\geq \lambda_{\min}(\nabla^2 f(x_k)) - r_{M_k}(x_k) L \\ &\geq \lambda_{\min}(\nabla^2 f(x_k)) - \tfrac{L\|\nabla f(x_k)\|}{\lambda_{\min}(\nabla^2 f(x_k))} \\ &= (1-\delta_k)\lambda_{\min}(\nabla^2 f(x_k)). \end{aligned} \tag{4.1.23}$$

Thus, $\nabla^2 f(x_{k+1})$ is also positive definite. Moreover, using inequality (4.1.10) and the upper bound for $M_k$, we obtain

$$\begin{aligned} \delta_{k+1} &= \tfrac{L\|\nabla f(x_{k+1})\|}{\lambda_{\min}^2(\nabla^2 f(x_{k+1}))} \;\leq\; \tfrac{3L^2 r_{M_k}^2(x_k)}{2\lambda_{\min}^2(\nabla^2 f(x_{k+1}))} \;\leq\; \tfrac{3L^2\|\nabla f(x_k)\|^2}{2\lambda_{\min}^4(\nabla^2 f(x_k))(1-\delta_k)^2} \\ &= \tfrac{3}{2}\left(\tfrac{\delta_k}{1-\delta_k}\right)^2 \;\leq\; \tfrac{8}{3}\delta_k^2. \end{aligned}$$

Thus, $\delta_{k+1} \leq \frac{1}{4}$ and we prove (4.1.19) by induction. We also get $\delta_{k+1} \leq \frac{2}{3}\delta_k$, and, since $\delta_0 \leq \frac{1}{4}$, we come to the following bound:

$$\sum_{i=0}^{\infty} \delta_i \;\leq\; \tfrac{\delta_0}{1-\frac{2}{3}} \leq 1 - \delta_0. \tag{4.1.24}$$

Further,

$$\ln \tfrac{\lambda_{\min}(\nabla^2 f(x_k))}{\lambda_{\min}(\nabla^2 f(x_0))} \overset{(4.1.23)}{\geq} \sum_{i=0}^{\infty} \ln(1-\delta_i) \geq -\sum_{i=0}^{\infty} \tfrac{\delta_i}{1-\delta_i} \geq -\tfrac{1}{1-\delta_0}\sum_{i=0}^{\infty}\delta_i \geq -1.$$

In order to get an upper bound, note that $\nabla^2 f(x_{k+1}) \overset{(4.1.2)}{\preceq} \nabla^2 f(x_k) + r_{M_k}(x_k) L\, I_n$. Hence,

$$\lambda_{\min}(\nabla^2 f(x_{k+1})) \leq \lambda_{\min}(\nabla^2 f(x_k)) + r_{M_k}(x_k) L \overset{(4.1.22)}{\leq} (1+\delta_k)\lambda_{\min}(\nabla^2 f(x_k)).$$

Therefore

$$\ln \tfrac{\lambda_{\min}(\nabla^2 f(x_k))}{\lambda_{\min}(\nabla^2 f(x_0))} \leq \sum_{i=0}^{\infty} \ln(1+\delta_i) \leq \sum_{i=0}^{\infty} \delta_i \leq \tfrac{3}{4}.$$

It remains to prove Item 3 of the theorem. In view of inequalities (4.1.22) and (4.1.20), we have

$$r_{M_k}(x_k) \leq \tfrac{1}{L}\lambda_{\min}(\nabla^2 f(x_k))\delta_k \leq \tfrac{e^{3/4}}{L}\lambda_{\min}(\nabla^2 f(x_0))\delta_k.$$

Thus, in view of the bound (4.1.24), $\{x_i\}$ is a Cauchy sequence, which has a unique limit point $x^*$. Since the eigenvalues of $\nabla^2 f(x)$ are continuous functions of $x$, from the first inequality in (4.1.20) we conclude that $\nabla^2 f(x^*) \succ 0$.

Further, by inequality (4.1.19), we get the bound

$$\delta_{k+1} \leq \tfrac{\delta_k^2}{(1-\delta_0)^2} \leq \tfrac{16}{9}\delta_k^2.$$

Defining $\hat{\delta}_k = \frac{16}{9}\delta_k$, we get $\hat{\delta}_{k+1} \leq \hat{\delta}_k^2$. Thus, for any $k \geq 1$, we have

$$\delta_k = \tfrac{9}{16}\hat{\delta}_k \leq \tfrac{9}{16}\hat{\delta}_0^{2^k} < \tfrac{9}{16}\left(\tfrac{1}{2}\right)^{2^k}.$$

Using the upper bound in (4.1.20), we get the last upper bound (4.1.21). □

### 4.1.3 Global Efficiency Bounds on Specific Problem Classes

In the previous section, we have already seen that the modified Newton scheme can be supported by a global efficiency estimate (4.1.17) on a general class of non-convex problems. The main goal of this section is to show that by specifying some additional properties of non-convex functions, it is possible to get for this method much better performance guarantees. A nice feature of method (4.1.16) consists in its ability to automatically adjust its rate of convergence to the specific problem classes.

#### 4.1.3.1 Star-Convex Functions

Let us start from a definition.

**Definition 4.1.1** We call the function $f$ *star-convex* if its set of global minimums $X^*$ is not empty and for any $x^* \in X^*$ and any $x \in \mathbb{R}^n$ we have

$$f(\alpha x^* + (1-\alpha)x) \leq \alpha f(x^*) + (1-\alpha)f(x) \quad \forall x \in \mathscr{F}, \ \forall \alpha \in [0,1]. \tag{4.1.25}$$

A particular example of a star-convex function is a usual convex function. However, in general star-convex function need not to be convex, even in the scalar case. For instance, $f(x) = |x|(1-e^{-|x|})$, $x \in \mathbb{R}$, is star-convex, but not convex. Star-convex functions arise quite often in optimization problems related to sum of squares. For example the function $f(x,y) = x^2y^2 + x^2 + y^2$ with $(x,y) \in \mathbb{R}^2$ belongs to this class.

**Theorem 4.1.4** *Assume that the objective function in the problem (4.1.14) is star-convex, and the set $\mathscr{F}$ is bounded: diam $\mathscr{F} = D < \infty$. Let the sequence $\{x_k\}$ be generated by method (4.1.16).*

1. *If $f(x_0) - f^* \geq \frac{3}{2}LD^3$, then $f(x_1) - f^* \leq \frac{1}{2}LD^3$.*
2. *If $f(x_0) - f^* \leq \frac{3}{2}LD^3$, then the rate of convergence of process (4.1.16) is as follows:*

$$f(x_k) - f(x^*) \leq \frac{3LD^3}{2(1+\frac{1}{3}k)^2}, \quad k \geq 0. \tag{4.1.26}$$

*Proof* Indeed, in view of inequality (4.1.11) the upper bound on the parameters $M_k$, and definition (4.1.25), for any $k \geq 0$ we have:

$$\begin{aligned} f(x_{k+1}) - f(x^*) \leq \ & \min_y \ \Big[ f(y) - f(x^*) + \tfrac{L}{2}\|y - x_k\|^3 : \\ & \qquad y = \alpha x^* + (1-\alpha)x_k, \alpha \in [0,1] \Big] \\ \leq \ & \min_{\alpha \in [0,1]} \Big[ f(x_k) - f(x^*) \\ & \qquad -\alpha(f(x_k) - f(x^*)) + \tfrac{L}{2}\alpha^3\|x^* - x_k\|^3 \Big] \\ \leq \ & \min_{\alpha \in [0,1]} \Big[ f(x_k) - f(x^*) - \alpha(f(x_k) - f(x^*)) + \tfrac{L}{2}\alpha^3 D^3 \Big]. \end{aligned}$$

The minimum of the objective function in the last minimization problem in $\alpha \geq 0$ is achieved for

$$\alpha_k = \sqrt{\frac{2(f(x_k) - f(x^*))}{3LD^3}}.$$

If $\alpha_k \geq 1$, then the actual optimal value corresponds to $\alpha = 1$. In this case,

$$f(x_{k+1}) - f(x^*) \leq \tfrac{1}{2} L D^3.$$

Since the process (4.1.16) is monotone, this can happen only at the first iteration of the method.

Assume that $\alpha_k \leq 1$. Then

$$f(x_{k+1}) - f(x^*) \leq f(x_k) - f(x^*) - \left[\tfrac{2}{3}(f(x_k) - f(x^*))\right]^{3/2} \tfrac{1}{\sqrt{LD^3}}.$$

Or, using the notation $\alpha_k = \sqrt{\frac{2(f(x_k)-f(x^*))}{3LD^3}}$, this is $\alpha_{k+1}^2 \leq \alpha_k^2 - \frac{2}{3}\alpha_k^3 < \alpha_k^2$. Therefore,

$$\tfrac{1}{\alpha_{k+1}} - \tfrac{1}{\alpha_k} = \tfrac{\alpha_k - \alpha_{k+1}}{\alpha_k \alpha_{k+1}} = \tfrac{\alpha_k^2 - \alpha_{k+1}^2}{\alpha_k \alpha_{k+1}(\alpha_k + \alpha_{k+1})} \geq \tfrac{\alpha_k^2 - \alpha_{k+1}^2}{2\alpha_k^3} \geq \tfrac{1}{3}.$$

Thus, $\frac{1}{\alpha_k} \geq \frac{1}{\alpha_0} + \frac{k}{3} \geq 1 + \frac{k}{3}$, and (4.1.26) follows. □

Let us now introduce the notion of a generalized non-degenerate global minimum.

**Definition 4.1.2** We say that the optimal set $X^*$ of function $f(\cdot)$ is *globally non-degenerate* if there exists a constant $\mu > 0$ such that for any $x \in \mathscr{F}$ we have

$$f(x) - f^* \geq \tfrac{\mu}{2} \rho^2(x, X^*), \tag{4.1.27}$$

where $f^*$ is the global minimal value of the function $f(\cdot)$, and $\rho(x, X^*)$ is the Euclidean distance from $x$ to $X^*$.

Of course, this property holds for strongly convex functions (see (3.2.43); in this case $X^*$ is a singleton). However, it can also hold for some non-convex functions. As an example, we can look at the function

$$f(x) = (\|x\|^2 - 1)^2, \quad X^* = \{x : \|x\| = 1\} \subset \mathbb{R}^n.$$

Note also that if the set $X^*$ has a connected non-trivial component, then the Hessians of the objective function at these points are necessarily degenerate. However, as we will see, in this situation the modified Newton scheme still ensures a super-linear rate of convergence. Define

$$\bar{\omega} = \tfrac{1}{L^2} \left(\tfrac{\mu}{2}\right)^3.$$

**Theorem 4.1.5** *Let a function $f$ be star-convex. Assume that it also has a globally non-degenerate optimal set. Then the performance of the scheme (4.1.16) on this problem is as follows.*

1. *If $f(x_0) - f(x^*) \geq \frac{4}{9}\bar{\omega}$, then at the first phase of the process we get the following rate of convergence:*

$$f(x_k) - f(x^*) \leq \left[ (f(x_0) - f(x^*))^{1/4} - \tfrac{k}{6}\sqrt{\tfrac{2}{3}}\bar{\omega}^{1/4} \right]^4. \tag{4.1.28}$$

*This phase is terminated as soon as $f(x_{k_0}) - f(x^*) \leq \frac{4}{9}\bar{\omega}$ for some $k_0 \geq 0$.*

2. *For $k \geq k_0$ the sequence converges superlinearly:*

$$f(x_{k+1}) - f(x^*) \leq \tfrac{1}{2}(f(x_k) - f(x^*))\sqrt{\tfrac{f(x_k)-f(x^*)}{\bar{\omega}}}. \tag{4.1.29}$$

*Proof* Denote by $x_k^*$ the projection of the point $x_k$ onto the optimal set $X^*$. In view of inequality (4.1.11) the upper bound on the parameters $M_k$ and definitions (4.1.25), (4.1.27), for any $k \geq 0$ we have:

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\leq \min_{\alpha \in [0,1]} \Big[ f(x_k) - f(x^*) - \alpha(f(x_k) - f(x^*)) \\ &\qquad + \frac{L}{2}\alpha^3 \|x_k^* - x_k\|^3 \Big] \\ &\leq \min_{\alpha \in [0,1]} \Big[ f(x_k) - f(x^*) - \alpha(f(x_k) - f(x^*)) \\ &\qquad + \frac{L}{2}\alpha^3 \left( \frac{2}{\mu}(f(x_k) - f(x^*)) \right)^{3/2} \Big]. \end{aligned}$$

Defining $\Delta_k = (f(x_k) - f(x^*))/\bar{\omega}$, we get the inequality

$$\Delta_{k+1} \leq \min_{\alpha \in [0,1]} \left[ \Delta_k - \alpha\Delta_k + \frac{1}{2}\alpha^3 \Delta_k^{3/2} \right]. \tag{4.1.30}$$

Note that the first-order optimality condition for $\alpha \geq 0$ in this problem is

$$\alpha_k = \sqrt{\tfrac{2}{3}\Delta_k^{-1/2}}.$$

Therefore, if $\Delta_k \geq \frac{4}{9}$, we get

$$\Delta_{k+1} \leq \Delta_k - \left(\tfrac{2}{3}\right)^{3/2} \Delta_k^{3/4}.$$

Defining $u_k = \frac{9}{4}\Delta_k$, we get a simpler relation:

$$u_{k+1} \leq u_k - \tfrac{2}{3}u_k^{3/4},$$

which is applicable if $u_k \geq 1$. Since the right-hand side of this inequality is increasing for $u_k \geq \frac{1}{16}$, let us prove by induction that

$$u_k \leq \left[u_0^{1/4} - \tfrac{k}{6}\right]^4.$$

Indeed, inequality

$$\left[u_0^{1/4} - \tfrac{k+1}{6}\right]^4 \geq \left[u_0^{1/4} - \tfrac{k}{6}\right]^4 - \tfrac{2}{3}\left[u_0^{1/4} - \tfrac{k}{6}\right]^3$$

is clearly equivalent to

$$\tfrac{2}{3}\left[u_0^{1/4} - \tfrac{k}{6}\right]^3 \geq \left[u_0^{1/4} - \tfrac{k}{6}\right]^4 - \left[u_0^{1/4} - \tfrac{k+1}{6}\right]^4 = \tfrac{1}{6}\left\{ \left[u_0^{1/4} - \tfrac{k}{6}\right]^3 \right.$$

$$\left. + \left[u_0^{1/4} - \tfrac{k}{6}\right]^2 \left[u_0^{1/4} - \tfrac{k+1}{6}\right] + \left[u_0^{1/4} - \tfrac{k}{6}\right]\left[u_0^{1/4} - \tfrac{k+1}{6}\right]^2 + \left[u_0^{1/4} - \tfrac{k+1}{6}\right]^3 \right\},$$

which is obviously true.

Finally, if $u_k \leq 1$, then the optimal value for $\alpha$ in (4.1.30) is equal to one, and we get (4.1.29). □

#### 4.1.3.2 Gradient-Dominated Functions

Let us now look at another interesting class of nonconvex functions.

**Definition 4.1.3** A function $f(\cdot)$ is called *gradient dominated* of degree $p \in [1, 2]$ if it attains a global minimum at some point $x^*$ and for any $x \in \mathscr{F}$ we have

$$f(x) - f(x^*) \leq \tau_f \|\nabla f(x)\|^p, \tag{4.1.31}$$

where $\tau_f$ is a positive constant. The parameter $p$ is called the *degree* of domination.

We do not assume here that the global minimum of function $f$ is unique. Let us give several examples of gradient dominated functions.

*Example 4.1.1 (Convex Functions)* Let $f$ be convex on $\mathbb{R}^n$. Assume it achieves its minimum at point $x^*$. Then, for any $x \in \mathbb{R}^n$ with $\|x - x^*\| < R$, we have

$$f(x) - f(x^*) \overset{(2.1.2)}{\leq} \langle \nabla f(x), x - x^* \rangle \leq \|\nabla f(x)\| \cdot R.$$

Thus, the function $f$ is a gradient dominated function of degree one on the set $\mathscr{F} = \{x : \|x - x^*\| < R\}$ with $\tau_f = R$. □

*Example 4.1.2 (Strongly Convex Functions)* Let $f$ be differentiable and strongly convex on $\mathbb{R}^n$. This means that there exists a constant $\mu > 0$ such that

$$f(y) \overset{(2.1.20)}{\geq} f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2}\mu \| y - x \|^2, \tag{4.1.32}$$

for all $x, y \in \mathbb{R}^n$. Then, minimizing both sides of this inequality in $y$, we obtain,

$$f(x) - f(x^*) \leq \tfrac{1}{2\mu} \| \nabla f(x) \|^2 \quad \forall x \in \mathbb{R}^n.$$

Thus, $f$ is a gradient dominated function of degree two on the set $\mathscr{F} = \mathbb{R}^n$ with $\tau_f = \frac{1}{2\mu}$. □

*Example 4.1.3 (Sum of Squares)* Consider a system of non-linear equations:

$$g(x) = 0, \tag{4.1.33}$$

where $g(x) = (g_1(x), \ldots, g_m(x))^T : \mathbb{R}^n \to \mathbb{R}^m$ is a differentiable vector function. We assume that $m \leq n$ and that there exists a solution $x^*$ to (4.1.33). Let us assume in addition that the Jacobian

$$J^T(x) = (\nabla g_1(x), \ldots, \nabla g_m(x))$$

is uniformly non-degenerate on a certain convex set $\mathscr{F}$ containing $x^*$. This means that the value

$$\sigma \equiv \inf_{x \in \mathscr{F}} \lambda_{\min} \left( J(x) J^T(x) \right)$$

is positive. Consider the function

$$f(x) = \frac{1}{2} \sum_{i=1}^{m} g_i^2(x).$$

Clearly, $f(x^*) = 0$. Note that $\nabla f(x) = J^T(x) g(x)$. Therefore,

$$\| \nabla f(x) \|^2 = \langle \left( J(x) J^T(x) \right) g(x), g(x) \rangle \;\geq\; \sigma \| g(x) \|^2 \;=\; 2\sigma (f(x) - f(x^*)).$$

Thus, $f$ is a gradient dominated function on $\mathscr{F}$ of degree two with $\tau_f = \frac{1}{2\sigma}$. Note that, for $m < n$, the set of solutions to (4.1.33) *is not* a singleton and therefore the Hessians of the function $f$ are necessarily degenerate at the solutions. □

In order to study the complexity of minimization of the gradient dominated functions, we need one auxiliary result.

**Lemma 4.1.8** *At each step of method (4.1.16) we can guarantee the following decrease of the objective function:*

$$f(x_k) - f(x_{k+1}) \geq \frac{L_0 \cdot \|\nabla f(x_{k+1})\|^{3/2}}{3\sqrt{2} \cdot (L+L_0)^{3/2}}, \quad k \geq 0. \tag{4.1.34}$$

*Proof* In view of inequalities (4.1.12) and (4.1.10) we get

$$f(x_k) - f(x_{k+1}) \geq \frac{M_k}{12} r^3_{M_k}(x_k) \geq \frac{M_k}{12} \left( \frac{2\|\nabla f(x_{k+1})\|}{L+M_k} \right)^{3/2} = \frac{M_k \|\nabla f(x_{k+1})\|^{3/2}}{3\sqrt{2} \cdot (L+M_k)^{3/2}}.$$

It remains to note that the right-hand side of this inequality is increasing in $M_k \leq 2L$. Thus, we can replace $M_k$ by its lower bound $L_0$. □

Let us start from the analysis of gradient dominated functions of degree one. The following theorem shows that the process can be partitioned into two phases. The first phase (with large values of the objective function) is very short, while at the second phase we can guarantee the rate of convergence of the order $O(1/k^2)$.

**Theorem 4.1.6** *Let us use method (4.1.16) for minimizing a gradient dominated function $f$ of degree $p = 1$.*

1. *If the initial value of the objective function is large enough:*

$$f(x_0) - f(x^*) \geq \hat{\omega} \overset{\text{def}}{=} \frac{18}{L_0^2} \tau_f^3 \cdot (L + L_0)^3,$$

*then the process converges to the region $\mathscr{L}(\hat{\omega})$ superlinearly:*

$$\ln \left( \frac{1}{\hat{\omega}} (f(x_k) - f(x^*) \right) \leq \left( \frac{2}{3} \right)^k \ln \left( \frac{1}{\hat{\omega}} (f(x_0) - f(x^*) \right). \tag{4.1.35}$$

2. *If $f(x_0) - f(x^*) \leq \gamma^2 \hat{\omega}$ for some $\gamma > 1$, then we have the following estimate for the rate of convergence:*

$$f(x_k) - f(x^*) \leq \hat{\omega} \cdot \frac{\gamma^2 \left(2 + \frac{3}{2}\gamma\right)^2}{\left(2 + \left(k + \frac{3}{2}\right) \cdot \gamma\right)^2}, \quad k \geq 0. \tag{4.1.36}$$

*Proof* Using inequalities (4.1.34) and (4.1.31) with $p = 1$, we get

$$f(x_k) - f(x_{k+1}) \geq \frac{L_0 \cdot (f(x_{k+1}) - f(x^*))^{3/2}}{3\sqrt{2} \cdot (L+L_0)^{3/2} \cdot \tau_f^{3/2}} = \hat{\omega}^{-1/2} (f(x_{k+1}) - f(x^*))^{3/2}.$$

Defining $\delta_k = (f(x_k) - f(x^*))/\hat{\omega}$, we obtain

$$\delta_k - \delta_{k+1} \geq \delta_{k+1}^{3/2}. \tag{4.1.37}$$

Hence, $\ln \delta_k \geq \ln \delta_{k+1} + \ln(1+\delta_{k+1}^{1/2}) \geq \frac{3}{2}\ln \delta_{k+1}$. Thus, $\ln \delta_k \leq \left(\frac{2}{3}\right)^k \ln \delta_0$, and this is inequality (4.1.35).

Let us now prove inequality (4.1.36). Using inequality (4.1.37), we have

$$\frac{1}{\sqrt{\delta_{k+1}}} - \frac{1}{\sqrt{\delta_k}} \geq \frac{1}{\sqrt{\delta_{k+1}}} - \frac{1}{\sqrt{\delta_{k+1}+\delta_{k+1}^{3/2}}} = \frac{\sqrt{\delta_{k+1}+\delta_{k+1}^{3/2}}-\sqrt{\delta_{k+1}}}{\sqrt{\delta_{k+1}}\sqrt{\delta_{k+1}+\delta_{k+1}^{3/2}}} = \frac{\sqrt{1+\delta_{k+1}^{1/2}}-1}{\sqrt{\delta_{k+1}+\delta_{k+1}^{3/2}}}$$

$$= \frac{1}{\sqrt{1+\sqrt{\delta_{k+1}}}\cdot\left(1+\sqrt{1+\sqrt{\delta_{k+1}}}\right)} = \frac{1}{1+\sqrt{\delta_{k+1}}+\sqrt{1+\sqrt{\delta_{k+1}}}}$$

$$\geq \frac{1}{2+\frac{3}{2}\sqrt{\delta_{k+1}}} \geq \frac{1}{2+\frac{3}{2}\sqrt{\delta_0}}.$$

Thus, $\frac{1}{\sqrt{\delta_k}} \geq \frac{1}{\gamma} + \frac{k}{2+\frac{3}{2}\gamma}$, and this is (4.1.36). □

The reader should not be confused by the superlinear rate of convergence established by (4.1.35). It is valid only for the first stage of the process and describes a convergence to the set $\mathscr{L}(\hat{\omega})$. For example, the first stage of the process discussed in Theorem 4.1.4 is even shorter: it takes just one iteration.

Let us now look at the gradient dominated functions of degree two. Here we can also see two phases of the process.

**Theorem 4.1.7** *Let us apply method (4.1.16) for minimizing a gradient dominated function $f$ of degree $p = 2$.*

1. *If the initial value of the objective function is large enough:*

$$f(x_0) - f(x^*) \geq \tilde{\omega} \overset{\mathrm{def}}{=} \frac{L_0^4}{324(L+L_0)^6\,\tau_f^3}, \tag{4.1.38}$$

*then at its first phase the process converges as follows:*

$$f(x_k) - f(x^*) \leq (f(x_0) - f(x^*)) \cdot e^{-k\cdot\sigma}, \tag{4.1.39}$$

*where $\sigma = \frac{\tilde{\omega}^{1/4}}{\tilde{\omega}^{1/4}+(f(x_0)-f(x^*))^{1/4}}$. This phase ends at the first iteration $k_0$ for which (4.1.38) does not hold.*

2. *For $k \geq k_0$, the rate of convergence is super-linear:*

$$f(x_{k+1}) - f(x^*) \leq \tilde{\omega} \cdot \left(\frac{f(x_k)-f(x^*)}{\tilde{\omega}}\right)^{4/3}. \tag{4.1.40}$$

*Proof* Using inequalities (4.1.34) and (4.1.31) with $p = 2$, we get

$$f(x_k) - f(x_{k+1}) \geq \frac{L_0\cdot(f(x_{k+1})-f(x^*))^{3/4}}{3\sqrt{2}\cdot(L+L_0)^{3/2}\cdot\tau_f^{3/4}} = \tilde{\omega}^{1/4}(f(x_{k+1}) - f(x^*))^{3/4}.$$

Defining $\delta_k = (f(x_k) - f(x^*))/\tilde{\omega}$, we obtain

$$\delta_k \geq \delta_{k+1} + \delta_{k+1}^{3/4}. \tag{4.1.41}$$

Hence,

$$\tfrac{\delta_k}{\delta_{k+1}} \geq 1 + \delta_k^{-1/4} \geq 1 + \delta_0^{-1/4} = \tfrac{1}{1-\sigma} \geq e^{\sigma},$$

and we get (4.1.39). Finally, from (4.1.41) we have $\delta_{k+1} \leq \delta_k^{4/3}$, which is (4.1.40). □

Comparing the statement of Theorem 4.1.7 with other theorems of this section, we can see a significant difference. This is the first time when the initial residual $f(x_0) - f(x^*)$ enters the complexity estimate of the first phase of the process in a polynomial way. In all other cases, the dependence on this value is much weaker. However, we will observe a similar situation in Sect. 5.2, when we will address the complexity of minimizing self-concordant functions.

Note that it is possible to embed the gradient dominated functions of degree two into the class of gradient dominated functions of degree one. However, it is easy to check that this only makes the efficiency estimates established by Theorem 4.1.7 worse.

#### 4.1.3.3 Nonlinear Transformations of Convex Functions

Let $u(x) : \mathbb{R}^n \to \mathbb{R}^n$ be a non-degenerate vector function. Denote by $v(u)$ its inverse:

$$v(u) : \mathbb{R}^n \to \mathbb{R}^n, \quad v(u(x)) \equiv x.$$

Consider the following function:

$$f(x) = \phi(u(x)),$$

where $\phi(u)$ is a convex function with bounded level sets. Denote by $x^* \equiv v(u^*)$ its minimum. Let us fix some $x_0 \in \mathbb{R}^n$. Define

$$\sigma = \max_u \{\| v'(u) \| : \phi(u) \leq f(x_0)\},$$

$$D = \max_u \{\| u - u^* \| : \phi(u) \leq f(x_0)\}.$$

The following result is straightforward.

**Lemma 4.1.9** *For any $x, y \in \mathscr{L}(f(x_0))$ we have*

$$\|x - y\| \le \sigma \|u(x) - u(y)\|. \tag{4.1.42}$$

*Proof* Indeed, for $x, y \in \mathscr{L}(f(x_0))$, we have $\phi(u(x)) \le f(x_0)$ and $\phi(u(y)) \le f(x_0)$. Consider the trajectory $x(t) = v(tu(y) + (1-t)u(x))$, $t \in [0, 1]$. Then

$$y - x = \int\limits_0^1 x'(t)dt = \left( \int\limits_0^1 v'(tu(y) + (1-t)u(x))dt \right) \cdot (u(y) - u(x)),$$

and (4.1.42) follows. □

The following result is very similar to Theorem 4.1.4.

**Theorem 4.1.8** *Assume that the Hessian of the function $f$ is Lipschitz continuous on a convex set $\mathscr{F} \supset \mathscr{L}(f(x_0))$ with constant $L$ and let the sequence $\{x_k\}$ be generated by method (4.1.16).*

1. *If $f(x_0) - f^* \ge \frac{3}{2}L(\sigma D)^3$, then $f(x_1) - f^* \le \frac{1}{2}L(\sigma D)^3$.*
2. *If $f(x_0) - f^* \le \frac{3}{2}L(\sigma D)^3$, then the rate of convergence of the process (4.1.16) is as follows:*

$$f(x_k) - f(x^*) \le \frac{3L(\sigma D)^3}{2(1+\frac{1}{3}k)^2}, \quad k \ge 0. \tag{4.1.43}$$

*Proof* Indeed, in view of inequality (4.1.11), the upper bound on the parameters $M_k$, and definition (4.1.25), for any $k \ge 0$ we have:

$$\begin{aligned} f(x_{k+1}) - f(x^*) \le \min_y [\ & f(y) - f(x^*) + \tfrac{L}{2}\|y - x_k\|^3 : \\ & y = v(\alpha u^* + (1-\alpha)u(x_k)), \alpha \in [0,1]\ ]. \end{aligned}$$

By definition of the points $y$ in the above minimization problem and (4.1.42), we have

$$f(y) - f(x^*) = \phi(\alpha u^* + (1-\alpha)u(x_k)) - \phi(u^*) \le (1-\alpha)(f(x_k) - f(x^*)),$$

$$\|y - x_k\| \le \alpha\sigma \|u(x_k) - u^*\| \le \alpha\sigma D.$$

This means that the reasoning of Theorem 4.1.4 goes through replacing $D$ by $\sigma D$. □

Let us prove a statement on strongly convex $\phi$. Define $\check{\omega} = \frac{1}{L^2}\left(\frac{\mu}{2\sigma^2}\right)^3$.

**Theorem 4.1.9** *Let the function $\phi$ be strongly convex with convexity parameter $\mu > 0$. Then, under assumptions of Theorem 4.1.8, the performance of the scheme (4.1.16) is as follows.*

1. *If $f(x_0) - f(x^*) \geq \frac{4}{9}\breve{\omega}$, then in the first phase of the process we get the following rate of convergence:*

$$f(x_k) - f(x^*) \leq \left[ (f(x_0) - f(x^*))^{1/4} - \frac{k}{6}\sqrt{\frac{2}{3}}\breve{\omega}^{1/4} \right]^4. \tag{4.1.44}$$

*This phase is terminated as soon as $f(x_{k_0}) - f(x^*) \leq \frac{4}{9}\breve{\omega}$ for some $k_0 \geq 0$.*

2. *For $k \geq k_0$, the sequence converges superlinearly:*

$$f(x_{k+1}) - f(x^*) \leq \frac{1}{2}(f(x_k) - f(x^*))\sqrt{\frac{f(x_k)-f(x^*)}{\breve{\omega}}}. \tag{4.1.45}$$

*Proof* Indeed, in view of inequality (4.1.11), the upper bound on the parameters $M_k$, and definition (4.1.25), for any $k \geq 0$ we have:

$$f(x_{k+1}) - f(x^*) \leq \min_y \Big[ f(y) - f(x^*) + \frac{L}{2}\|y - x_k\|^3 :$$

$$y = v(\alpha u^* + (1-\alpha)u(x_k)), \alpha \in [0,1] \Big].$$

By definition of the points $y$ in the above minimization problem and (4.1.42), we have

$$f(y) - f(x^*) = \phi(\alpha u^* + (1-\alpha)u(x_k)) - \phi(u^*) \leq (1-\alpha)(f(x_k) - f(x^*)),$$

$$\|y - x_k\| \leq \alpha\sigma\|u(x_k) - u^*\| \overset{(2.1.21)}{\leq} \alpha\sigma\sqrt{\frac{2}{\mu}(f(x_0) - f(x^*))}.$$

This means that the reasoning of Theorem 4.1.5 goes through replacing $L$ by $\sigma^3 L$. □

Note that the functions described in this section are often used as test functions for non-convex optimization algorithms. The simplest way of defining a nondegenerate transformation $u(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ is as follows:

$$\begin{aligned} u^{(1)}(x) &= x^{(1)}, \\ u^{(2)}(x) &= x^{(2)} + \phi_1(x^{(1)}), \\ u^{(3)}(x) &= x^{(3)} + \phi_2(x^{(1)}, x^{(2)}), \\ \dots & \quad \dots \\ u^{(n)}(x) &= x^{(n)} + \phi_{n-1}(x^{(1)}, \dots, x^{(n-1)}), \end{aligned} \tag{4.1.46}$$

where $\phi_1, \cdots, \phi_{n-1}$ are *arbitrary* differentiable functions. It is clear that the Jacobian $u'(x)$ is an upper-triangular matrix with unit diagonal. Thus, this transformation is non-degenerate.

## 4.1.4 Implementation Issues

### 4.1.4.1 Minimizing the Cubic Regularization

In order to compute the mapping $T_M(x)$, we need to solve an auxiliary minimization problem (4.1.5), namely,

$$\min_{h\in\mathbb{R}^n}\left[v(h) \overset{\text{def}}{=} \langle g,h\rangle + \tfrac{1}{2}\langle Hh,h\rangle + \tfrac{M}{6}\|h\|^3\right]. \tag{4.1.47}$$

If the Hessian $H$ is indefinite, this problem is nonconvex. It can have many strict isolated minima, while we need to find a global one. Nevertheless, as we will show in this section, this problem is equivalent to a convex univariate optimization problem.

Note that the objective function of the optimization problem (4.1.47) can be represented in the following way:

$$v(h) = \min_{\tau\in\mathbb{R}}\left\{\tilde{v}(h,\tau) \overset{\text{def}}{=} \langle g,h\rangle + \tfrac{1}{2}\langle Hh,h\rangle + \tfrac{M}{6}|\tau|^{3/2} :\ \|h\|^2 \le \tau\right\}.$$

Thus, the point $T_M(x)$ can be found from the following problem

$$\min_{h\in\mathbb{R}^n,\tau\in\mathbb{R}}\left[\tilde{v}(h,\tau) :\ f(h,\tau) \overset{\text{def}}{=} \tfrac{1}{2}\|h\|^2 - \tfrac{1}{2}\tau \le 0\right].$$

Since this is already a *constrained* minimization problem, we can form for it a *Lagrangian dual problem* (see Sect. 1.3.3). Indeed, define the Lagrangian $\mathscr{L}(h,\tau,\lambda) = \tilde{v}(h,\tau) + \lambda[\frac{1}{2}\|h\|^2 - \frac{1}{2}\tau]$ with $h \in \mathbb{R}^n$ and $\tau,\lambda \in \mathbb{R}$. Then the dual function is

$$\psi(\lambda) = \inf_{h\in\mathbb{R}^n,\tau\in\mathbb{R}}\left\{\langle g,h\rangle + \tfrac{1}{2}\langle Hh,h\rangle + \tfrac{M}{6}|\tau|^{3/2} + \lambda[\tfrac{1}{2}\|h\|^2 - \tfrac{1}{2}\tau]\right\}.$$

The optimal value of $\tau$ can be found from the equation $\frac{M}{4}|\tau|^{1/2}\text{sign}(\tau) = \frac{1}{2}\lambda$. Therefore, $\tau(\lambda) = \frac{4\lambda|\lambda|}{M^2}$, and we have

$$\psi(\lambda) = \inf_{h\in\mathbb{R}^n}\left\{\langle g,h\rangle + \tfrac{1}{2}\langle (H+\lambda I_n)h,h\rangle - \tfrac{2}{3M^2}|\lambda|^3\right\},$$

$$\text{dom}\ \psi = \left\{\lambda\in\mathbb{R} :\ \inf_{h\in\mathbb{R}^n}[q_\lambda(h) \overset{\text{def}}{=} \langle g,h\rangle + \tfrac{1}{2}\langle (H+\lambda I_n)h,h\rangle] > -\infty\right\}.$$

Let us describe the structure of dom $\psi$. Without loss of generality, we can assume that $H$ is a diagonal matrix with values $\{H_i\}_{i=1}^n$ on the diagonal. Let $H_{\min} = \min\limits_{1 \le i \le n} H_i$.

If $\lambda > -H_{\min}$, then $\lambda \in$ dom $\psi$. If $\lambda < -H_{\min}$, then $\lambda \notin$ dom $\psi$. Thus, only the status of the point $\lambda = -H_{\min}$ can be different. Define

$$G^2 = \sum_{i \in I^*} (g^{(i)})^2, \quad I^* = \{i : H_i = H_{\min}\}.$$

There are three possibilities.

1. $G^2 > 0$. Then dom $\psi = \{\lambda \in \mathbb{R} : \lambda > -H_{\min}\}$. For any $\lambda$ in this domain we have

$$\psi(\lambda) = -\frac{1}{2}\frac{G^2}{H_{\min}+\lambda} - \frac{1}{2}\sum_{i \notin I^*} \frac{(g^{(i)})^2}{H_i+\lambda} - \frac{2}{3M^2}|\lambda|^3. \tag{4.1.48}$$

   At the same time, the optimal vector for the function $q_\lambda(\cdot)$ has the form

$$h(\lambda) = -(H + \lambda I_n)^{-1} g.$$

   This vector and value $\tau(\lambda)$ are uniquely defined and continuous on dom $\psi$. Hence, in view of Theorem 1.3.2, we have

$$\min_{h \in \mathbb{R}^n} v(h) = \max_{\lambda \in \text{dom } \psi \bigcap \mathbb{R}_+} \psi(\lambda). \tag{4.1.49}$$

2. $G^2 = 0$. Then dom $\psi = \{\lambda \in \mathbb{R} : \lambda \ge -H_{\min}\}$. In this case, for any $\lambda > -H_{\min}$, the optimal vector is uniquely defined as follows:

$$h^{(i)}(\lambda) = \begin{cases} \frac{g^{(i)}}{H_i+\lambda}, & \text{if } i \notin I^*, \\ 0, & \text{otherwise}, \end{cases} \quad i = 1, \dots, n. \tag{4.1.50}$$

   This vector is continuous on dom $\psi$. Therefore, if

$$\lambda^* \overset{\text{def}}{=} \arg \max_{\lambda \in \text{dom } \psi \bigcap \mathbb{R}_+} \psi(\lambda) > -H_{\min},$$

   then the conditions of Theorem 1.3.2 are satisfied. Hence, in this case relation (4.1.49) is also valid.
3. The only remaining case is $G^2 = 0$ and $\lambda^* = -H_{\min}$. This is possible only if $H_{\min} \le 0$ and the gradient is small enough (e.g. $g = 0$). In this situation, the rule (4.1.50) does not work and we need to form the solution of problem (4.1.47) using an eigenvector of matrix $H$, which corresponds to the eigenvalue $H_{\min}$.

Let us choose an arbitrary $k \in I^*$ and a small parameter $\delta > 0$. Define a new function

$$v_\delta(h) = v(h) + \delta h^{(k)}.$$

This function satisfies the condition of Item 1. Therefore, in view of (4.1.49) we have

$$\max_{h \in \mathbb{R}^n} v_\delta(h) = \max_{\lambda \in \mathrm{dom}\ \psi_\delta \bigcap \mathbb{R}_+} \psi_\delta(\lambda),$$

$$\psi_\delta(\lambda) = -\frac{1}{2}\frac{\delta^2}{H_{\min}+\lambda} - \frac{1}{2}\sum_{i \notin I^*} \frac{(g^{(i)})^2}{H_i+\lambda} - \frac{2}{3M^2}|\lambda|^3.$$

Since dom $\psi_\delta = (-H_{\min}, +\infty)$, the optimal point of the dual problem $\lambda_\delta^*$ can be found from the following equation:

$$\frac{\delta^2}{(H_{\min}+\lambda)^2} + \sum_{i \notin I^*} \frac{(g^{(i)})^2}{(H_i+\lambda)^2} = \frac{4\lambda^2}{M^2}. \tag{4.1.51}$$

Thus, the optimal vector for the primal problem is

$$h_*(\delta) = -(H + \lambda_\delta^* I_n)^{-1}(g + \delta e_k).$$

All components $h_*^{(i)}(\delta)$ with $i \neq k$ are continuous in $\delta$ (recall that $H$ is a diagonal matrix). For $i = k$, we have

$$h_*^{(k)}(\delta) = -\frac{\delta}{H_{\min}+\lambda_\delta^*} \overset{(4.1.51)}{=} -\left[\frac{4(\lambda_\delta^*)^2}{M^2} - \sum_{i \notin I^*} \frac{(g^{(i)})^2}{(H_i+\lambda_\delta^*)^2}\right]^{1/2}.$$

Thus, there exists a limit $h_* = \lim\limits_{\delta \to 0} h_*(\delta)$, defined as follows:

$$\begin{aligned} h_* &= \sum_{i \notin I^*} h_*^{(i)} e_i + h_*^{(k)} e_k, \quad h_*^{(i)} = -\frac{g^{(i)}}{H_i - H_{\min}}, \quad i \notin I^*, \\ h_*^{(k)} &= -\left[\frac{4H_{\min}^2}{M^2} - \sum_{i \notin I^*} \frac{(g^{(i)})^2}{(H_i - H_{\min})^2}\right]^{1/2}. \end{aligned} \tag{4.1.52}$$

It is easy to see that $h_*$ is a global optimum for problem (4.1.47). Indeed, for any $h \in \mathbb{R}^n$ we have

$$v_\delta(h) \geq v_\delta(h_*(\delta)) \ \geq \ v(h_*(\delta)) - \delta |h_*^{(k)}(\delta)|.$$

Taking in these inequalities the limit as $\delta \to 0$, we get $v(h) \geq v(h_*)$. □

Note that in both Items 1 and 2, the optimal solution of the dual problem $\lambda^*$ satisfies the first-order optimality condition

$$\psi'(\lambda^*) = -\frac{1}{2}\frac{G^2}{(H_{\min}+\lambda^*)^2} - \frac{1}{2}\sum_{i\notin I^*}\frac{(g^{(i)})^2}{(H_i+\lambda^*)^2} - \frac{2}{M^2}(\lambda^*)^2 \overset{(1.2.4)}{=} 0,$$

and the optimal global solution of primal problem (4.1.47) is $h_* = -(H+\lambda^* I_n)^{-1}g$. In other words, $\lambda^*$ satisfies the equation

$$\|(H+\lambda^* I_n)^{-1}g\| = \frac{2}{M}\lambda^*. \tag{4.1.53}$$

Thus, $r_M(x) = \|h_*\| = \frac{2}{M}\lambda^*$, and we conclude that $H + \frac{Mr_M(x)}{2}I_n \succeq 0$ (this is (4.1.8)). Note that in the case described in Item 3, we have $\|h_*\| = \frac{2|H_{\min}|}{M}$, Thus, we also have

$$H + \frac{Mr_M(x)}{2}I_n = H + |H_{\min}|I_n \quad \succeq 0.$$

Using the new variable $r$, we can rewrite equation (4.1.53) in the following form

$$r = \|\left(H + \frac{Mr}{2}I\right)^{-1} g\|, \tag{4.1.54}$$

with $r \geq \frac{2}{M}(-\lambda_{\min}(H))_+$. A technique for solving such equations is very well developed for the needs of Trust Region Methods. As compared with (4.1.54), the equations for Trust Region Schemes have a constant left-hand side. But of course, all possible difficulties with (4.1.54) are due to the non-linear convex right-hand side. In any case, before running a procedure for solving this equation, it is reasonable to transform the matrix $H$ into a tri-diagonal form using the Lanczos algorithm. In the general case, it takes $O(n^3)$ operations.

In order to illustrate possible difficulties arising in the dual problem, let us look at the following example.

*Example 4.1.4* Let $n = 2$ and

$$g = (-1,0)^T, \quad H_1 = 0, \quad H_2 = -1, \quad M = 1.$$

Thus, our primal problem is as follows:

$$\min_{h\in\mathbb{R}^2}\left\{\psi(h) \equiv -h^{(1)} - \frac{1}{2}\left(h^{(2)}\right)^2 + \frac{1}{6}\left[\sqrt{\left(h^{(1)}\right)^2 + \left(h^{(2)}\right)^2}\right]^3\right\}.$$

Following (4.1.6), we have to solve the system of two non-linear equations:

$$\frac{h^{(1)}}{2}\sqrt{\left(h^{(1)}\right)^2+\left(h^{(2)}\right)^2}=1,$$

$$\frac{h^{(2)}}{2}\sqrt{\left(h^{(1)}\right)^2+\left(h^{(2)}\right)^2}=h^{(2)}.$$

Thus, we have three candidate solutions:

$$h_1^*=(\sqrt{2},0)^T,\quad h_2^*=(1,\sqrt{3})^T,\quad h_3^*=(1,-\sqrt{3})^T.$$

By direct substitution, we can see that

$$\psi(h_1^*)=-\tfrac{2\sqrt{2}}{3}\ >\ -\tfrac{7}{6}=\psi(h_2^*)=\psi(h_3^*).$$

Thus, both $h_2^*$ and $h_3^*$ are our global solutions.

Let us look at the dual problem. Since $G^2=0$, we have the following objective:

$$\psi(\lambda)\overset{(4.1.48)}{=}-\tfrac{1}{2\lambda}-\tfrac{2}{3}\lambda^3.$$

We need to maximize this function subject to the constraint $\lambda\ge(-H_{\min})_+=1$. Since $\psi'(1)<0$, we conclude that $\lambda^*=1$. Thus, using representation (4.1.52), we get

$$h^*=-e_1\cdot\tfrac{-1}{0+1}+e_2\left[4H_{\min}^2-\tfrac{1}{(-H_{\min})^2}\right]^{1/2}\ =\ (1,\sqrt{3})^T.\qquad\square$$

To the best of our knowledge, a technique for finding the global minimum of problem (4.1.47) in the degenerate situation of Item 3 without computing an eigenvalue decomposition of the matrix $H$ is not known yet. Of course, we can always say that this degeneracy disappears with probability one after an arbitrary small random perturbation of the vector $g$.

#### 4.1.4.2 Line Search Strategies

Let us discuss the computational cost of Step 1 in method (4.1.16), which consists in finding $M_k\in[L_0,2L]$ satisfying the equation:

$$f(T_{M_k}(x_k))\le\bar{f}_{M_k}(x_k).$$

Note that for $M_k \geq L$ this inequality holds. Consider now the following *backtracking strategy*.

$$\begin{array}{l} \text{Find the first } i_k \geq 0 \text{ such that } f(T_{2^{i_k} M_k}(x)) \leq \bar{f}_{2^{i_k} M_k}(x_k). \\ \text{Define } x_{k+1} := T_{2^{i_k} M_k}(x_k) \text{ and } M_{k+1} := 2^{i_k} M_k. \end{array} \tag{4.1.55}$$

If we apply this procedure at each iteration of process (4.1.16), which starts from $M_0 \in [L_0, 2L]$, then we have the following advantages:

- $M_k \leq 2L$.
- The total amount of additional computations of mappings $T_{M_k}(\cdot)$ during $N$ iterations of process (4.1.16) is equal to

$$\sum_{k=0}^{N} i_k = \sum_{k=0}^{N} \log_2 \frac{M_{k+1}}{M_k} = \log_2 \frac{M_{N+1}}{M_0} \leq 1 + \log_2 \frac{L}{L_0}.$$

(Indeed, if $i_k = 0$, then we compute only one mapping $T_{M_k}(\cdot)$ at this iteration.) The right-hand side of the above bound does not depend on $N$, the number of iterations of the main process.

However, it may happen that rule (4.1.55) is too conservative. Indeed, we can only increase our estimates for the constant $L$ and never let them go down. This may force the method to take only short steps. A more optimistic strategy is as follows:

$$\begin{array}{l} \text{Find the first } i_k \geq 0 \text{ such that } f(T_{2^{i_k} M_k}(x_k)) \leq \bar{f}_{2^{i_k} M_k}(x_k). \\ \text{Define } x_{k+1} := T_{2^{i_k} M_k}(x_k) \text{ and } M_{k+1} := \max\left\{L_0, 2^{i_k - 1} M_k\right\}. \end{array} \tag{4.1.56}$$

Then the total amount of additional computations of mappings $T_{M_k}(\cdot)$ after $N$ iterations of the process (4.1.16) can be bounded as follows

$$\sum_{k=0}^{N} i_k \leq \sum_{k=0}^{N} \log_2 \frac{2M_{k+1}}{M_k} = N + 1 + \log_2 \frac{M_{N+1}}{M_0} \leq N + 2 + \log_2 \frac{L}{L_0}.$$

Thus, after $N$ iterations of this process, we never compute more than

$$2N + 3 + \log_2 \frac{2L}{L_0}$$

mappings $T_M(\cdot)$. This is a reasonable price to pay for the possibility of moving by long steps.

### 4.1.5 Global Complexity Bounds

Let us compare the complexity results presented in this section with some known facts on global efficiency bounds of other minimization schemes.

Assume that the function $f$ is strongly convex on $\mathbb{R}^n$ with convexity parameter $\mu > 0$ (see (4.1.32)). In this case, there exists its unique global minimum $x^*$, and condition (4.1.27) holds for all $x \in \mathbb{R}^n$ (see Theorem 2.1.8). Assume also that the Hessian of this function is Lipschitz continuous:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

For such functions, let us obtain the complexity bounds of method (4.1.16) using the results of Theorems 4.1.4 and 4.1.5.

Indeed, let us fix some $x_0 \in \mathbb{R}^n$. Denote by $D$ the *radius* of its level set:

$$D = \max_x \{\|x - x^*\| : \ f(x) \leq f(x_0)\}.$$

From the condition (4.1.27), we get

$$D \leq \left[\frac{2}{\mu}(f(x_0) - f(x^*))\right]^{1/2}.$$

We will see that it is natural to measure the quality of the starting point $x_0$ by the following characteristic:

$$\varkappa \equiv \varkappa(x_0) = \frac{LD}{\mu}.$$

Let us introduce three switching values

$$\omega_0 = \frac{\mu^3}{18L^2} \equiv \frac{4}{9}\bar{\omega}, \quad \omega_1 = \frac{3}{2}\mu D^2, \quad \omega_2 = \frac{3}{2}LD^3.$$

In view of Theorem 4.1.4, we can reach the level $f(x_0) - f(x^*) \leq \frac{1}{2}LD^3$ in one iteration. Therefore, without loss of generality we assume that

$$f(x_1) - f(x^*) \leq \omega_2.$$

Suppose we are interested in a very high accuracy of the solution. Note that the case $\varkappa \leq 1$ is very easy since the first iteration of method (4.1.16) comes very close to the region of super-linear convergence (see Item 2 of Theorem 4.1.5).

Consider the case $\varkappa \geq 1$. Then $\omega_0 \leq \omega_1 \leq \omega_2$. Let us estimate the duration of the following phases:

$$\text{Phase 1: } \omega_1 \leq f(x_i) \leq \omega_2,$$

$$\text{Phase 2: } \omega_0 \leq f(x_i) \leq \omega_1,$$

$$\text{Phase 3: } \quad \epsilon \leq f(x_i) \leq \omega_0.$$

In view of Theorem 4.1.4, the duration $k_1$ of the first phase is bounded as follows:

$$\omega_1 \leq \tfrac{3LD^3}{2(1+\frac{1}{3}k_1)^2}.$$

Thus, $k_1 \leq 3\sqrt{\varkappa}$. Further, in view of Item 1 of Theorem 4.1.5, we can bound the duration $k_2$ of the second phase:

$$\begin{aligned} \omega_0^{1/4} &\leq (f(x_{k_1+1}) - f(x^*))^{1/4} - \tfrac{k_2}{6}\omega_0^{1/4} \\ &\leq (\tfrac{1}{2}\mu D^2)^{1/4} - \tfrac{k_2}{6}\omega_0^{1/4}. \end{aligned}$$

This gives the following bound: $k_2 \leq 3^{3/4}2^{1/2}\sqrt{\varkappa} \leq 3.25\sqrt{\varkappa}$.

Finally, let $\delta_k = \frac{1}{4\omega_0}(f(x_k) - f(x^*))$. In view of inequality (4.1.29) we have:

$$\delta_{k+1} \leq \delta_k^{3/2}, \quad k \geq \bar{k} \equiv k_1 + k_2 + 1.$$

At the same time $f(x_{\bar{k}}) - f(x^*) \leq \omega_0$. Thus, $\delta_{\bar{k}} \leq \frac{1}{4}$, and the bound on the duration $k_3$ of the last phase can be found from the following inequality:

$$\left(\tfrac{3}{2}\right)^{k_3} \ln 4 \leq \ln \tfrac{4\omega_0}{\epsilon}.$$

This is $k_3 \leq \log_{\frac{3}{2}} \log_4 \frac{2\mu^3}{9\epsilon L^2}$. Putting all the bounds together, we obtain that the total number of steps $N$ in (4.1.16) is bounded as follows:

$$N \leq 6.25\sqrt{\tfrac{LD}{\mu}} + \log_{\frac{3}{2}}\left(\log_4 \tfrac{1}{\epsilon} + \log_4 \tfrac{2\mu^3}{9L^2}\right). \tag{4.1.57}$$

It is interesting that in estimate (4.1.57) the parameters of our problem interact with accuracy in an *additive* way. Recall that usually such an interaction is multiplicative. Let us estimate, for example, the complexity of our problem for the Fast Gradient Method (2.2.20) for strongly convex functions with Lipschitz continuous gradient. Denote by $\hat{L}$ the largest eigenvalue of the matrix $\nabla^2 f(x^*)$.

Then we can guarantee that

$$\mu I \preceq \nabla^2 f(x) \preceq (\hat{L} + LD)I \quad \forall x, \ \|x - x^*\| \le D.$$

Thus, the complexity bound for the optimal gradient method is of the order of

$$O\left(\sqrt{\tfrac{\hat{L}+LD}{\mu}} \ln \tfrac{(\hat{L}+LD)D^2}{\epsilon}\right)$$

iterations. For the Gradient Method (2.1.37) it is even worse:

$$O\left(\tfrac{\hat{L}+LD}{\mu} \ln \tfrac{(\hat{L}+LD)D^2}{\epsilon}\right).$$

Thus, we conclude that the global complexity bounds of the Cubic Newton Method (4.1.16) are considerably better than the estimates of the gradient schemes. At the same time, we should recall, of course, the difference in computational cost of each iteration.

Note that similar bounds can be obtained for other classes of non-convex problems. For example, for nonlinear transformations of convex functions (see Sect. 4.1.3.3), the complexity bound is as follows:

$$N \le 6.25\sqrt{\tfrac{\sigma}{\mu} LD} + \log_{\frac{3}{2}}\left(\log_4 \tfrac{1}{\epsilon} + \log_4 \tfrac{2\mu^3}{9\sigma^6 L^2}\right). \tag{4.1.58}$$

To conclude, note that in scheme (4.1.16) it is possible to find elements of the Levenberg–Marquardt approach (see relation (4.1.8)), or a trust-region idea (see the discussion in Sect. 4.1.4.1), or a line-search technique (see the rule of Step 1 in (4.1.16)). However, all these facts are *consequences* of the main idea of the scheme, consisting in computation of the next test point of the process as a global minimizer of cubic regularization of the second-order approximation, which globally overestimates the values of the objective function.

## 4.2 Accelerated Cubic Newton

(Primal and dual spaces; Uniformly convex functions; Regularization of Newton iteration; An Accelerated scheme Global non-degeneracy for second-order schemes; Minimizing strongly convex functions; False accelerations.)

### *4.2.1 Real Vector Spaces*

Starting from this section, we often work with more abstract real vector spaces. In the previous part of the book, we were dealing mainly with the simplest space

$\mathbb{R}^n$. However, very often we need to highlight the fundamental difference between the vectors of decision variables and the vectors of gradients. The simplest way of doing this is to just keep them in different spaces. For us, the space for variables will always be the *primal space*, and the space for gradients will be the *dual space*.

Let $\mathbb{E}$ be a finite-dimensional real vector space, and $\mathbb{E}^*$ be its dual space, comprised of linear functions on $\mathbb{E}$. Denote by $\langle s, x\rangle_{\mathbb{E}}$ the value of $s \in \mathbb{E}^*$ at a point $x \in \mathbb{E}$ (sometimes it is called the *scalar product* of $s$ and $x$). If there is no ambiguity of notation, the subscript of the scalar product is usually omitted. Since we always work in finite dimensions, we have $(\mathbb{E}^*)^* = \mathbb{E}$.

Consider, for example, a differentiable function $f$ with $\mathrm{dom}\, f = \mathbb{E}$. Then, by definition of the gradient, we have

$$f(x+h) = f(x) + \langle \nabla f(x), h\rangle + o(\|h\|), \quad x, h \in \mathbb{E}.$$

Thus, the gradient defines a *linear function* of $x$, and therefore $\nabla f(x) \in \mathbb{E}^*$. It is important to remember that the coordinate form of the gradient (1.2.3) makes sense only if $\mathbb{E} = \mathbb{E}^* = \mathbb{R}^n$. In order to convert $\mathbb{E}$ to $\mathbb{R}^n$, we need to fix a basis of this space. This operation can be done in many different ways, which significantly change the topology of functions and their characteristics. Therefore, it is often convenient to avoid this operation in explaining the principles of optimization schemes.

Further, for two spaces $\mathbb{E}_1$ and $\mathbb{E}_2^*$, we can consider a linear operator $A : \mathbb{E}_1 \to \mathbb{E}_2^*$. For this operator, we can define the *adjoint operator* $A^*$ as follows:

$$\langle Ax, y\rangle_{\mathbb{E}_2} \equiv \langle A^* y, x\rangle_{\mathbb{E}_1}, \quad \forall x \in \mathbb{E}_1,\ y \in \mathbb{E}_2.$$

Clearly, $A^*$ maps $\mathbb{E}_2$ to $\mathbb{E}_1^*$. In the case when $\mathbb{E}_1 = \mathbb{R}^n$, and $\mathbb{E}_2 = \mathbb{R}^m$ the operator $A$ can be represented by an $(m \times n)$-matrix. Then the matrix for $A^*$ is just its transpose: $A^* = A^T$.

In order to have a full picture, let us describe a standard procedure for converting $\mathbb{E}$ and $\mathbb{E}^*$ into $\mathbb{R}^n$. Let $n = \dim \mathbb{E}$. Let us choose a basis $B = (b_1, \dots, b_n)$ in $\mathbb{E}$. We can treat it as a linear operator $B : \mathbb{R}^n \to \mathbb{E}$ defined by the following rule:

$$x = B\bar{x} \overset{\mathrm{def}}{=} \sum_{i=1}^{n} b_i \bar{x}^{(i)}, \quad \bar{x} = (\bar{x}^{(1)}, \dots, \bar{x}^{(n)})^T \in \mathbb{R}^n.$$

Using this basis, we can define a linear operator $B^* : \mathbb{E}^* \to \mathbb{R}^n$ as follows:

$$\bar{s} = (\bar{s}^{(1)}, \dots, \bar{s}^{(n)})^T \ = \ B^* s \in R^n, \quad s \in \mathbb{E}^*,$$

which is equivalent to the following rules:

$$\bar{s}^{(i)} = \langle s, b_i\rangle, \quad i = 1, \dots, n.$$

Then, using the operator $(B^*)^{-1} : \mathbb{R}^n \to \mathbb{E}^*$, we can define the *dual basis* in $\mathbb{E}^*$. Indeed, $s = (B^*)^{-1}\bar{s} \in \mathbb{E}^*$ for $\bar{s} \in \mathbb{R}^n$. Therefore, the corresponding basis vectors in $\mathbb{E}^*$ are as follows:

$$\left((B^*)^{-1}e_1, \dots, (B^*)^{-1}e_n\right),$$

where $e_i$ are the unit coordinate vectors in $\mathbb{R}^n$, $i = 1.\dots, n$. Note that

$$\begin{aligned} \langle (B^*)^{-1}\bar{s}, b_i\rangle_{\mathbb{E}} &= \langle (B^*)^{-1}\bar{s}, Be_i\rangle_{\mathbb{E}} = \langle B^*(B^*)^{-1}\bar{s}, e_i\rangle_{\mathbb{R}^n} \\ &= \bar{s}^{(i)}, \quad i = 1, \dots, n. \end{aligned} \tag{4.2.1}$$

Hence, we get the following representation for the *scalar product* of two vectors $s \in \mathbb{E}^*$ and $x \in \mathbb{E}$:

$$\begin{aligned} \langle s, x\rangle_{\mathbb{E}} &= \langle (B^*)^{-1}\bar{s}, B\bar{x}\rangle_{\mathbb{E}} = \sum_{i=1}^{n} \bar{x}^{(i)} \langle (B^*)^{-1}\bar{s}, b_i\rangle \\ &\overset{(4.2.1)}{=} \sum_{i=1}^{m} \bar{x}^{(i)}\bar{s}^{(i)} = \bar{s}^T\bar{x} \equiv \langle \bar{s}, \bar{x}\rangle_{\mathbb{R}^n}. \end{aligned}$$

Further, the operator $B : \mathbb{E} \to \mathbb{E}^*$ is called *self-adjoint* if

$$\langle Bx, y\rangle \equiv \langle By, x\rangle, \quad \forall x, y \in \mathbb{E}.$$

For $\mathbb{E} = \mathbb{R}^n$ a self-adjoint operator is represented by a symmetric matrix. The most important examples of self-adjoint operators are given by *Hessians*. Indeed, by definition (see (1.2.7)), we have

$$\nabla f(x+h) = \nabla f(x) + \nabla^2 f(x)h + \mathbf{o}(\|h\|) \in \mathbb{E}^*, \quad x \in \mathbb{E}, \ h \in \mathbb{E}.$$

Thus, $\nabla^2 f(x)$ is a linear operator from $\mathbb{E}$ to $\mathbb{E}^*$. This interpretation confirms the validity of the Newton direction:

$$[\nabla^2 f(x)]^{-1}\nabla f(x) \in \mathbb{E}.$$

It is well known that for twice continuously differentiable functions the matrix representation of the Hessian is symmetric. This means that any Hessian is a self-adjoint operator.

Finally, a self-adjoint operator $B : \mathbb{E} \to \mathbb{E}^*$ is positive semidefinite if

$$\langle Bx, x\rangle \geq 0, \quad \forall x \in \mathbb{E},$$

notation $B \succeq 0$. If the above inequality is strict for all $x \neq 0$, we call the operator positive definite (notation $B \succ 0$). Positive definite operators are invertible.

Now we can define all necessary objects. Let us fix a positive definite self-adjoint operator $B : \mathbb{E} \to \mathbb{E}^*$. Define the primal norm for the space $\mathbb{E}$:

$$\| h \| = \langle Bh, h \rangle^{1/2}, \quad h \in \mathbb{E}. \tag{4.2.2}$$

Our above discussion suggests that the most natural candidates for such an operator $B$ are nondegenerate Hessians of convex functions. We will discuss this possibility in detail in Chap. 5.

The dual norm for $\mathbb{E}^*$ can be defined in the standard way:

$$\| s \|_* = \max_{x \in \mathbb{E}} \{ \langle s, x \rangle : \ \| x \| \leq 1 \} \overset{(3.1.64)}{=} \langle s, B^{-1} s \rangle^{1/2}, \quad s \in \mathbb{E}^*. \tag{4.2.3}$$

An immediate consequence of this definition is the *Cauchy–Schwarz inequality*

$$\langle s, x \rangle \overset{(4.2.3)}{\leq} \| s \|_* \cdot \| x \|, \quad x \in \mathbb{E}, \ s \in \mathbb{E}^*. \tag{4.2.4}$$

Finally, for a linear operator $A : \mathbb{E} \to \mathbb{E}^*$ we have

$$\| A \| = \max_{\| h \| \leq 1} \| Ah \|_*. \tag{4.2.5}$$

If the operator $A$ is self-adjoint, the same norm can be defined as

$$\| A \| = \max_{\| h \| \leq 1} | \langle Ah, h \rangle |. \tag{4.2.6}$$

Any $s \in \mathbb{E}^*$ generates a rank-one self-adjoint operator $ss^* : \mathbb{E} \to \mathbb{E}^*$ acting as follows

$$ss^* \cdot x \ = \ \langle s, x \rangle \cdot s, \quad x \in \mathbb{E}.$$

We extend the operator $A(s) \overset{\text{def}}{=} \frac{ss^*}{\| s \|_*}$ onto the origin in a continuous way: $A(0) = 0$.

In this section, we mainly consider functions with Lipschitz-continuous Hessian:

$$\| \nabla^2 f(x) - \nabla^2 f(y) \| \ \leq \ L_3 \| x - y \|, \quad x, y \in \mathbb{E}, \tag{4.2.7}$$

where $L_3 \overset{\text{def}}{=} L_3(f)$. Consequently, for all $x$ and $y$ from $\mathbb{E}$ we have

$$\| \nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x) \|_* \overset{(1.2.13)}{\leq} \tfrac{1}{2} L_3 \| y - x \|^2. \tag{4.2.8}$$

Moreover, for the quadratic model

$$f_2(x; y) \overset{\text{def}}{=} f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle$$

we can bound the residual:

$$|f(y) - f_2(x; y)| \overset{(1.2.14)}{\leq} \frac{L_3}{6}\|y - x\|^3, \quad x, y \in \mathbb{E}. \tag{4.2.9}$$

### 4.2.2 Uniformly Convex Functions

In this section, we will often use the *cubic power function*

$$d_3(x) = \tfrac{1}{3}\|x - x_0\|^3, \quad \nabla d_3(x) \;=\; \|x - x_0\| \cdot B(x - x_0), \quad x \in \mathbb{E}.$$

This is the simplest example of the *uniformly convex* function. In order to understand their properties, we need to develop some theory.

Let the function $d(\cdot)$ be differentiable on a closed convex set $Q$. We call it *uniformly convex on $Q$ of degree $p \geq 2$* if there exists a constant $\sigma_p = \sigma_p(d) > 0$ such that[1]

$$d(y) \geq d(x) + \langle \nabla d(x), y - x\rangle + \tfrac{1}{p}\sigma_p\|y - x\|^p, \quad \forall x, y \in Q. \tag{4.2.10}$$

The constant $\sigma_p$ is called the *parameter of uniform convexity* of this function. By adding such a function to an arbitrary convex function, we get a uniformly convex function of the same degree and with the same value of parameter. Recall that degree $p = 2$ corresponds to *strongly convex* functions (see (2.1.20)). In our old notation, the parameter $\mu$ of strong convexity for the function $f$ corresponds to $\sigma_2(f)$.

Note that any uniformly convex function grows faster than any linear function. Therefore, its level sets are always bounded. This implies that any minimization problem with uniformly convex objective is always solvable provided that its feasible set is nonempty. Moreover, its solution is always unique.

Adding two copies of inequality (4.2.10) with $x$ and $y$ interchanged, we get

$$\langle \nabla d(x) - \nabla d(y), x - y\rangle \geq \tfrac{2}{p}\sigma_p\|x - y\|^p, \quad \forall x, y \in Q. \tag{4.2.11}$$

It appears that this condition is sufficient for uniform convexity (however, for $p > 2$ the convexity parameter is changing).

**Lemma 4.2.1** *Assume that for some $p \geq 2$, $\sigma > 0$, and all $x, y \in Q$ the following inequality holds:*

$$\langle \nabla d(x) - \nabla d(y), x - y\rangle \geq \sigma\|x - y\|^p, \quad x, y \in Q. \tag{4.2.12}$$

*Then the function $d$ is uniformly convex on $Q$ with degree $p$ and parameter $\sigma$.*

[1] It could be a good exercise for the reader to prove that there are no uniformly convex functions with degree $p \in (0, 2)$.

*Proof* Indeed,

$$\begin{aligned} d(y) - d(x) - \langle \nabla d(x), y - x \rangle &= \int_0^1 \langle d(x + \tau(y-x)) - \nabla d(x), y - x \rangle d\tau \\ &= \int_0^1 \frac{1}{\tau} \langle d(x + \tau(y-x)) - \nabla d(x), \tau(y - x) \rangle d\tau \\ &\overset{(4.2.12)}{\geq} \int_0^1 \sigma \tau^{p-1} \| y - x \|^p d\tau \;=\; \frac{1}{p} \sigma \| y - x \|^p. \qquad \square \end{aligned}$$

**Lemma 4.2.2** *Let $d$ be uniformly convex on $Q$ of degree $p \geq 2$. Then for all $x, y \in Q$ we have*

$$d(y) - d(x) - \langle \nabla d(x), y - x \rangle \leq \tfrac{p-1}{p} \left( \tfrac{1}{\sigma_p} \right)^{\frac{1}{p-1}} \| \nabla d(y) - \nabla d(x) \|_*^{\frac{p}{p-1}}. \tag{4.2.13}$$

*Proof* Assume that $d$ attains its global minimum on $\mathbb{E}$ at some point $x^* \in Q$. Then

$$\begin{aligned} d(x^*) \;&=\; \min_{y \in Q} d(y) \overset{(4.2.10)}{\geq} \min_{x \in Q} \left[ d(x) + \langle \nabla d(x), y - x \rangle + \tfrac{1}{p} \sigma_p \| y - x \|^p \right] \\ &\geq\; \min_{x \in \mathbb{E}} \left[ d(x) + \langle \nabla d(x), y - x \rangle + \tfrac{1}{p} \sigma_p \| y - x \|^p \right] \\ &\overset{(4.2.3)}{=} d(x) - \tfrac{p-1}{p} \left( \tfrac{1}{\sigma_p} \right)^{\frac{1}{p-1}} \| \nabla d(x) \|_*^{\frac{p}{p-1}}. \end{aligned}$$

Let us fix $x \in Q$ and consider the convex function $\phi(y) = d(y) - \langle \nabla d(x), y \rangle$. It is uniformly convex of degree $p$ and parameter $\sigma_p$. Moreover, it attains its minimum at $y = x \in Q$. Hence, applying the above inequality to $\phi(y)$, we get (4.2.13). □

Let us give an important example of a uniformly convex function. By fixing an arbitrary $x_0 \in \mathbb{E}$, we define the function $d_p(x) = \frac{1}{p} \| x - x_0 \|^p$, where the norm is Euclidean (see (4.2.2)). Then

$$\nabla d_p(x) \;=\; \| x - x_0 \|^{p-2} \cdot B(x - x_0), \quad x \in \mathbb{E}.$$

**Lemma 4.2.3** *For any $x$ and $y$ from $\mathbb{E}$ we have*

$$\langle \nabla d_p(x) - \nabla d_p(y), x - y \rangle \geq \left( \tfrac{1}{2} \right)^{p-2} \| x - y \|^p, \tag{4.2.14}$$

$$d_p(x) - d_p(y) - \langle \nabla d_p(y), x - y \rangle \geq \tfrac{1}{p} \left( \tfrac{1}{2} \right)^{p-2} \| x - y \|^p. \tag{4.2.15}$$

*Proof* Without loss of generality, let us assume that $x_0 = 0$. Then

$$\begin{aligned}\langle \nabla d_p(x) - \nabla d_p(y), x - y\rangle &= \langle \|x\|^{p-2} \cdot Bx - \|y\|^{p-2} \cdot By, x - y\rangle \\ &= \|x\|^p + \|y\|^p - \langle Bx, y\rangle(\|x\|^{p-2} + \|y\|^{p-2}).\end{aligned}$$

To prove (4.2.14), we need to show that the right-hand side of the latter equality is greater than or equal to

$$\left(\frac{1}{2}\right)^{p-2} \|x - y\|^p = \left(\frac{1}{2}\right)^{p-2} \left[\|x\|^2 + \|y\|^2 - 2\langle Bx, y\rangle\right]^{p/2}.$$

Without loss of generality we can assume that $x \neq 0$ and $y \neq 0$. Then, defining

$$\tau = \tfrac{\|y\|}{\|x\|}, \quad \alpha = \tfrac{\langle Bx, y\rangle}{\|x\| \cdot \|y\|} \in [-1, 1],$$

we obtain the statement to be proved:

$$1 + \tau^p \geq \alpha\tau(1 + \tau^{p-2}) + \left(\frac{1}{2}\right)^{p-2} [1 + \tau^2 - 2\alpha\tau]^{p/2}, \quad \tau \geq 0, \quad |\alpha| \leq 1. \tag{4.2.16}$$

Since the right-hand side of this inequality is convex in $\alpha$, in view of Corollary 3.1.2, we need to justify two marginal inequalities:

$$\begin{aligned}\alpha = 1: &\quad 1 + \tau^p \geq \tau(1 + \tau^{p-2}) + \left(\tfrac{1}{2}\right)^{p-2} |1 - \tau|^p, \\ \alpha = -1: &\quad 1 + \tau^p \geq -\tau(1 + \tau^{p-2}) + \left(\tfrac{1}{2}\right)^{p-2} (1 + \tau)^p\end{aligned} \tag{4.2.17}$$

for all $\tau \geq 0$.

The second inequality in (4.2.17) can be derived from the lower bound for the ratio

$$\tfrac{1+\tau^p+\tau(1+\tau^{p-2})}{(1+\tau)^p} = \tfrac{1+\tau^{p-1}}{(1+\tau)^{p-1}}, \quad \tau \geq 0.$$

Indeed, its minimum is attained at $\tau = 1$, and this proves the second line in (4.2.17). To prove the first line, note that it is valid for $\tau = 1$. If $\tau \geq 0$ and $\tau \neq 1$, then we need to estimate from below the ratio

$$\tfrac{1+\tau^p-\tau(1+\tau^{p-2})}{|1-\tau|^p} = \tfrac{(1-\tau)(1-\tau^{p-1})}{|1-\tau|^p} = \tfrac{1+\tau+\cdots+\tau^{p-2}}{|1-\tau|^{p-2}}.$$

Since the absolute value of any coefficient of the polynomial $(1-\tau)^{p-2}$ does not exceed $2^{p-2}$, the first line in inequality (4.2.17) is also justified. This proves (4.2.14), and, to prove (4.2.15), we can now use Lemma 4.2.1. □

The main property of uniformly convex functions is the following *growth condition*.

**Theorem 4.2.1** *Let $d$ be uniformly convex on $Q$ of degree $p \ge 2$ with positive constant $\sigma_p$. Let $x^* = \arg\min\limits_{x\in Q} d(x)$. Then for all $x \in Q$ we have*

$$d(x) \ge d(x^*) + \tfrac{1}{p}\sigma_p \|x-x^*\|^p. \tag{4.2.18}$$

*Proof* Indeed, in view of the first-order optimality condition (2.2.39), we have

$$\langle \nabla d(x^*), x - x^* \rangle \ge 0, \quad x \in Q.$$

Therefore, (4.2.18) follows from (4.2.10). □

Thus, by (4.2.14) and Lemma 4.2.1 we conclude that $\sigma_3(d_3) = \frac{1}{2}$. On the other hand we can prove the following important fact.

**Lemma 4.2.4** *For any $x, y \in \mathbb{E}$ we have*

$$\|\nabla^2 d_3(x) - \nabla^2 d_3(y)\| \le 2\,\|x-y\|. \tag{4.2.19}$$

*Proof* For any $x \in \mathbb{E}$, we have $\nabla^2 d_3(x) = \|x\|B + \frac{1}{\|x\|}Bxx^*B$. Clearly, for all $x \in \mathbb{E}$ we have

$$\|\nabla^2 d_3(x)\| \overset{(4.2.4)}{\le} 2\|x\|. \tag{4.2.20}$$

Let us fix two points $x, y \in \mathbb{E}$ and an arbitrary direction $h \in \mathbb{E}$. Define $x(\tau) = x + \tau(y-x)$ and

$$\phi(\tau) = \langle \nabla^2 d_3(x(\tau))h, h\rangle \;=\; \|x(\tau)\| \cdot \|h\|^2 + \tfrac{1}{\|x(\tau)\|}\langle Bx(\tau), h\rangle^2, \quad \tau \in [0,1].$$

Assume first that $0 \notin [x,y]$. Then $\phi(\tau)$ is continuously differentiable on $[0,1]$ and

$$\begin{aligned}\phi'(\tau) &= \tfrac{\langle Bx(\tau), y-x\rangle}{\|x(\tau)\|} \cdot \|h\|^2 + \tfrac{2\langle Bx(\tau), h\rangle}{\|x(\tau)\|}\langle Bh, y-x\rangle - \tfrac{\langle Bx(\tau),h\rangle^2}{\|x(\tau)\|^3}\langle Bx(\tau), y-x\rangle \\ &= \tfrac{\langle Bx(\tau), y-x\rangle}{\|x(\tau)\|} \cdot \underbrace{\left(\|h\|^2 - \frac{\langle Bx(\tau), h\rangle^2}{\|x(\tau)\|^2}\right)}_{\ge 0 \text{ by } (4.2.4)} + \tfrac{2\langle Bx(\tau),h\rangle}{\|x(\tau)\|}\langle Bh, y-x\rangle.\end{aligned}$$

Let $\alpha = \frac{\langle Bx(\tau),h\rangle}{\|x(\tau)\|\cdot\|h\|} \in [-1,1]$. Then

$$|\phi'(\tau)| \le \|y-x\| \cdot \|h\|^2 \cdot (1-\alpha^2 + 2|\alpha|) \;\le\; 2\,\|y-x\| \cdot \|h\|^2.$$

Hence,

$$|\langle(\nabla^2 d_3(y) - \nabla^2 d_3(x))h, h\rangle| \; = \; |\phi(1) - \phi(0)| \; \leq \; 2\,\|y - x\| \cdot \|h\|^2,$$

and we get (4.2.19) from (4.2.6).

The remaining case $0 \in [x, y]$ is trivial since then $\|x - y\| = \|x\| + \|y\|$ and we can apply (4.2.20). □

In the sequel, we often use Lipschitz constants for different derivatives. For $p \geq 2$, denote by $L_p(f)$ the Lipschitz constant for the $(p-1)$-st derivative of the function $f$:

$$\|\nabla^{(p-1)} f(x) - \nabla^{(p-1)} f(y)\| \leq L_p(f)\|x - y\|, \quad x, y \in \mathrm{dom}\ f. \tag{4.2.21}$$

In this notation, $L_2(f)$ is the Lipschitz constant for the gradient of the function $f$. At the same time, by Lemma 4.2.4, we conclude that $L_3(d_3) = 2$.

We often establish the complexity of different problem classes in terms of *condition numbers* of variable degree:

$$\gamma_p(f) \overset{\mathrm{def}}{=} \frac{\sigma_p(f)}{L_p(f)}, \quad p \geq 2. \tag{4.2.22}$$

It is clear, for example, that for $d_2(x) = \frac{1}{2}\|x - x_0\|^2$ we have $\gamma_2(d_2) = 1$. On the other hand, we have seen that $\gamma_3(d_3) = \frac{1}{4}$.

### 4.2.3 Cubic Regularization of Newton Iteration

Consider the following minimization problem:

$$\min_{x \in \mathbb{E}} \; f(x), \tag{4.2.23}$$

where $\mathbb{E}$ is a finite-dimension real vector space, and $f$ is a twice differentiable *convex* function with Lipschitz-continuous Hessian. As was shown in Sect. 4.1, the global rate of convergence of the Cubic Newton Method (CNM) on this problem class is of the order $O(\frac{1}{k^2})$, where $k$ is the iteration counter (see Theorem 4.1.4). However, note that CNM is a *local one-step* second-order method. From the complexity theory of smooth Convex Optimization, it is known that the rate of convergence of the local one-step first-order method (this is just the *Gradient* Method, see Theorem 2.1.14) can be improved from $O(\frac{1}{k})$ to $O(\frac{1}{k^2})$ by applying a *multi-step* strategy (see, for example, Theorem 2.2.3). In this section we show that a similar trick also works with CNM. As a result, we get a new method, which converges on the specified problem class as $O(\frac{1}{k^3})$.

Let us recall the most important properties of cubic regularization of Newton's method, taking into account the convexity of the objective function.

As suggested in Sect. 4.1, we introduce the following mapping:

$$T_M(x) \overset{\text{def}}{=} \operatorname{Arg}\min_{y \in \mathbb{E}} \left[ \hat{f}_M(x;y) \overset{\text{def}}{=} f_2(x;y) + \tfrac{M}{6}\|y-x\|^3 \right]. \tag{4.2.24}$$

Note that $T = T_M(x)$ is a unique solution of the following equation

$$\nabla f(x) + \nabla^2 f(x)(T-x) + \tfrac{1}{2}M \cdot \|T-x\| \cdot B(T-x) = 0. \tag{4.2.25}$$

Define $r_M(x) = \|x - T_M(x)\|$. Then,

$$\begin{aligned} \|\nabla f(T)\|_* &\overset{(4.2.25)}{=} \|\nabla f(T) - \nabla f(x) - \nabla^2 f(x)(T-x) - \frac{M}{2} r_M(x) B(T-x)\|_* \\ &\overset{(4.2.8)}{\leq} \frac{L_3+M}{2} r_M^2(x). \end{aligned} \tag{4.2.26}$$

Further, multiplying (4.2.25) by $T-x$, we obtain

$$\langle \nabla f(x), x-T \rangle \;=\; \langle \nabla^2 f(x)(T-x), T-x \rangle + \frac{1}{2} M r_M^3(x). \tag{4.2.27}$$

Let us assume that $M \geq L_3$. Then, in view of (4.2.9), we have

$$\begin{aligned} f(x) - f(T) &\geq f(x) - \hat{f}_M(x;T) \\ &= \langle \nabla f(x), x-T \rangle - \frac{1}{2}\langle \nabla^2 f(x)(T-x), T-x \rangle - \frac{M}{6} r_M^3(x) \\ &= \frac{1}{2}\langle \nabla^2 f(x)(T-x), T-x \rangle + \frac{M}{3} r_M^3(x). \end{aligned} \tag{4.2.28}$$

In particular, since $f$ is convex,

$$f(x) - f(T) \overset{(4.2.28)}{\geq} \tfrac{M}{3} r_M^3(x) \overset{(4.2.26)}{\geq} \tfrac{M}{3}\left( \tfrac{2}{L_3+M} \|\nabla f(T)\|_* \right)^{3/2}. \tag{4.2.29}$$

Sometimes we need to interpret this step from a global perspective:

$$\begin{aligned} f(T) &\overset{(M \geq L_3)}{\leq} \min_y \left[ f_2(x;y) + \tfrac{M}{6}\|y-x\|^3 \right] \\ &\overset{(4.2.9)}{\leq} \min_y \left[ f(y) + \tfrac{L_3+M}{6}\|y-x\|^3 \right]. \end{aligned} \tag{4.2.30}$$

Finally, let us prove the following result.

**Lemma 4.2.5** *If $M \geq 2L_3$, then*

$$\langle \nabla f(T), x - T\rangle \geq \sqrt{\tfrac{2}{L_3+M}} \cdot \|\nabla f(T)\|_*^{3/2}. \tag{4.2.31}$$

*Proof* Let $T = T_M(x)$ and $r = r_M(x)$. Then

$$\begin{aligned} \tfrac{1}{4}L_3^2 r^4 &= \left(\tfrac{L_3}{2}\|T-x\|^2\right)^2 \overset{(4.2.8)}{\geq} \|\nabla f(T) - \nabla f(x) - \nabla^2 f(x)(T-x)\|_*^2 \\ &\overset{(4.2.25)}{=} \|\nabla f(T) + \tfrac{1}{2}M \cdot r \cdot B(T-x)\|_*^2 \\ &= \|\nabla f(T)\|_*^2 + Mr\langle \nabla f(T), T - x\rangle + \tfrac{1}{4}M^2 r^4. \end{aligned}$$

Hence,

$$\langle \nabla f(T), x - T\rangle \geq \tfrac{1}{Mr}\|\nabla f(T)\|_*^2 + \tfrac{1}{4M}(M^2 - L_3^2)r^3. \tag{4.2.32}$$

In view of the conditions of the lemma, we can estimate the derivative in $r$ of the right-hand side of inequality (4.2.32):

$$-\tfrac{1}{Mr^2}\|\nabla f(T)\|_*^2 + \tfrac{3r^2}{4M}(M^2 - L_3^2) \geq -\tfrac{1}{Mr^2}\|\nabla f(T)\|_*^2 + \left(\tfrac{L_3+M}{2}\right)^2 \tfrac{r^2}{M} \overset{(4.2.26)}{\geq} 0.$$

Thus, its minimum is attained at the boundary point $r = \left[\frac{2}{L_3+M}\|\nabla f(T)\|_*\right]^{1/2}$ of the feasible ray (4.2.26). Substituting this value into (4.2.32), we obtain (4.2.31). □

To conclude this section, let us estimate the rate of convergence of CNM as applied to our main problem (4.2.23). We assume that there exists a solution of this problem $x^*$, and the Lipschitz constant $L_3$ for the Hessian of objective function is known. Thus, we just iterate

$$x_{k+1} = T_{L_3}(x_k), \quad k = 0, 1, \dots . \tag{4.2.33}$$

**Theorem 4.2.2** *Assume that the level sets of problem (4.2.23) are bounded:*

$$\|x - x^*\| \leq D \quad \forall x : \; f(x) \leq f(x_0). \tag{4.2.34}$$

*If the sequence $\{x_k\}_{k=1}^\infty$ is generated by method (4.2.33), then*

$$f(x_k) - f(x^*) \leq \tfrac{9L_3 D^3}{(k+4)^2}, \quad k \geq 1. \tag{4.2.35}$$

*Proof* In view of (4.2.28), $f(x_{k+1}) \leq f(x_k)$ for all $k \geq 0$. Thus, $\|x_k - x^*\| \leq D$, $k \geq 0$. Further, in view of (4.2.30), we have

$$f(x_1) \leq f(x^*) + \tfrac{L_3}{3}D^3. \tag{4.2.36}$$

Consider now an arbitrary $k \geq 1$. Let $x_k(\tau) = x^* + (1-\tau)(x_k - x^*)$. In view of inequality (4.2.30), for any $\tau \in [0,1]$ we have

$$f(x_{k+1}) \leq f(x_k(\tau)) + \tau^3 \tfrac{L_3}{3}\|x_k - x^*\|^3 \;\leq\; f(x_k) - \tau(f(x_k) - f(x^*)) + \tau^3 \tfrac{L_3 D^3}{3}.$$

The minimum of the right-hand side of this inequality in $\tau$ is attained for

$$\tau = \sqrt{\tfrac{f(x_k)-f(x^*)}{L_3 D^3}} \;\leq\; \sqrt{\tfrac{f(x_1)-f(x^*)}{L_3 D^3}} \overset{(4.2.36)}{<} 1.$$

Thus, for any $k \geq 1$, we have

$$f(x_{k+1}) \leq f(x_k(\tau)) - \tfrac{2}{3}\cdot \tfrac{(f(x_k)-f(x^*))^{3/2}}{\sqrt{L_3 D^3}}. \tag{4.2.37}$$

Let $\delta_k = f(x_k) - f(x^*)$. Then

$$\begin{aligned}\tfrac{1}{\sqrt{\delta_{k+1}}} - \tfrac{1}{\sqrt{\delta_k}} = \tfrac{\delta_k - \delta_{k+1}}{\sqrt{\delta_k\delta_{k+1}}(\sqrt{\delta_k}+\sqrt{\delta_{k+1}})} &\overset{(4.2.37)}{\geq} \tfrac{2}{3\sqrt{L_3D^3}} \cdot \tfrac{\delta_k}{\sqrt{\delta_{k+1}}(\sqrt{\delta_k}+\sqrt{\delta_{k+1}})}\\ &\geq \tfrac{1}{3\sqrt{L_3 D^3}}.\end{aligned}$$

Thus, for any $k \geq 1$, we have

$$\tfrac{1}{\sqrt{\delta_k}} \geq \tfrac{1}{\sqrt{\delta_1}} + \tfrac{k-1}{3\sqrt{L_3D^3}} \overset{(4.2.36)}{\geq} \tfrac{1}{\sqrt{L_3 D^3}} \cdot \left(\sqrt{3} + \tfrac{k-1}{3}\right) \geq \tfrac{k+4}{3\sqrt{L_3D^3}}. \qquad \square$$

### 4.2.4 An Accelerated Scheme

In order to accelerate method (4.2.33), we apply a variant of the *estimating sequences technique*, which we presented in Sect. 2.2.1 as a tool for accelerating the usual Gradient Method. In our situation, this idea can be applied to CNM in the following way.

To solve the problem (4.2.23), we recursively update the following sequences.

- The sequence of estimating functions

$$\psi_k(x) = \ell_k(x) + \tfrac{C}{6}\|x - x_0\|^3, \quad k = 1, 2, \ldots, \tag{4.2.38}$$

where $\ell_k(x)$ are linear functions in $x \in \mathbb{E}$, and $C$ is a positive parameter.
- The minimizing sequence $\{x_k\}_{k=1}^{\infty}$.
- The sequence of scaling parameters $\{A_k\}_{k=1}^{\infty}$:

$$A_{k+1} \overset{\text{def}}{=} A_k + a_k, \quad k = 1, 2, \ldots.$$

For these objects, we are going to maintain the following relations:

$$\left.\begin{array}{l} \mathscr{R}_k^1 : A_k f(x_k) \leq \psi_k^* \;\equiv\; \min\limits_{x\in\mathbb{E}} \psi_k(x), \\[2ex] \mathscr{R}_k^2 : \quad \psi_k(x) \leq A_k f(x) + \frac{2L_3+C}{6}\|x-x_0\|^3, \;\; \forall x \in \mathbb{E} \end{array}\right\}, \quad k \geq 1. \tag{4.2.39}$$

Let us ensure that relations (4.2.39) hold for $k = 1$. We choose

$$x_1 \;=\; T_{L_3}(x_0), \quad \ell_1(x) \;\equiv\; f(x_1), \; x \in \mathbb{E}, \quad A_1 \;=\; 1. \tag{4.2.40}$$

Then $\psi_1^* = f(x_1)$, so $\mathscr{R}_1^1$ holds. On the other hand, in view of definition (4.2.38), we get

$$\begin{array}{rcl} \psi_1(x) & = & f(x_1) + \frac{C}{6}\|x-x_0\|^3 \\[2ex] & \overset{(4.2.30)}{\leq} & \min\limits_{y\in\mathbb{E}} \left[ f(y) + \frac{2L_3}{6}\|y-x_0\|^3 \right] + \frac{C}{6}\|x-x_0\|^3, \end{array}$$

and $\mathscr{R}_1^2$ follows.

Assume now that relations (4.2.39) hold for some $k \geq 1$. Let

$$v_k \;=\; \arg\min_{x\in\mathbb{E}} \; \psi_k(x).$$

Let us choose some $a_k > 0$ and $M \geq 2L_3$. Define[2]

$$\begin{array}{l} \alpha_k = \frac{a_k}{A_k+a_k}, \quad y_k \;=\; (1-\alpha_k)x_k + \alpha_k v_k, \quad x_{k+1} \;=\; T_M(y_k), \\[2ex] \psi_{k+1}(x) = \psi_k(x) + a_k[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle]. \end{array} \tag{4.2.41}$$

In view of $\mathscr{R}_k^2$, for any $x \in \mathbb{E}$ we have

$$\begin{array}{rl} \psi_{k+1}(x) \leq A_k f(x) + \frac{2L_3+C}{6}\|x-x_0\|^3 + a_k[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle] \\[2ex] \overset{(2.1.2)}{\leq} (A_k + a_k) f(x) + \frac{2L_3+C}{6}\|x-x_0\|^3, \end{array}$$

and this is $\mathscr{R}_{k+1}^2$. Let us show now that, for the appropriate choices of $a_k$, $C$ and $M$, relation $\mathscr{R}_{k+1}^1$ is also valid.

[2] This is the main difference with the technique presented in Sect. 2.2.1: we update the estimating function by a linearization computed at the *new point* $x_{k+1}$.

Indeed, in view of $\mathscr{R}_k^1$ and Lemma 4.2.3 with $p = 3$, for any $x \in \mathbb{E}$, we have

$$\begin{aligned} \psi_k(x) \equiv \ell_k(x) + \tfrac{C}{2} d_3(x) \;&\geq\; \psi_k^* + \tfrac{C}{2} \cdot \tfrac{1}{6} \|x - v_k\|^3 \\ &\geq A_k f(x_k) + \tfrac{C}{2} \cdot \tfrac{1}{6} \|x - v_k\|^3. \end{aligned} \tag{4.2.42}$$

Therefore,

$$\begin{aligned} \psi_{k+1}^* \;&=\; \min_{x \in \mathbb{E}} \{\psi_k(x) + a_k[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle]\} \\ &\overset{(4.2.42)}{\geq} \min_{x \in \mathbb{E}} \Big\{A_k f(x_k) + \tfrac{C}{12}\|x - v_k\|^3 \\ &\qquad + a_k[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle]\} \\ &\overset{(2.1.2)}{\geq} \min_{x \in \mathbb{E}} \{(A_k + a_k) f(x_{k+1}) + A_k \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle \\ &\qquad + a_k \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \tfrac{C}{12}\|x - v_k\|^3]\} \\ &\overset{(4.2.41)}{=} \min_{x \in \mathbb{E}} \{A_{k+1} f(x_{k+1}) + \langle \nabla f(x_{k+1}), A_{k+1} y_k - a_k v_k - A_k x_{k+1}\rangle \\ &\qquad + a_k \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \tfrac{C}{12}\|x - v_k\|^3]\} \\ &= \min_{x \in \mathbb{E}} \{A_{k+1} f(x_{k+1}) + A_{k+1} \langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle \\ &\qquad + a_k \langle \nabla f(x_{k+1}), x - v_k\rangle + \tfrac{C}{12}\|x - v_k\|^3]\}. \end{aligned}$$

Further, if we choose $M \geq 2L_3$, then by (4.2.31) we have

$$\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle \;\geq\; \sqrt{\tfrac{2}{L_3+M}} \cdot \|\nabla f(x_{k+1})\|_*^{3/2}.$$

Hence, our choice of parameters must ensure the following inequality:

$$A_{k+1}\sqrt{\tfrac{2}{L_3+M}} \cdot \|\nabla f(x_{k+1})\|_*^{3/2} + a_k \langle \nabla f(x_{k+1}), x - v_k\rangle + \tfrac{C}{12}\|x - v_k\|^3 \geq 0,$$

for all $x \in \mathbb{E}$. Minimizing this expression in $x \in \mathbb{E}$, we come to the following condition:

$$A_{k+1}\sqrt{\tfrac{2}{L_3+M}} \geq \tfrac{4}{3\sqrt{C}}\, a_k^{3/2}. \tag{4.2.43}$$

For $k \geq 1$, let us choose

$$\begin{aligned} A_k &= \tfrac{k(k+1)(k+2)}{6}, \\ a_k &= A_{k+1} - A_k = \tfrac{(k+1)(k+2)(k+3)}{6} - \tfrac{k(k+1)(k+2)}{6} \\ &= \tfrac{(k+1)(k+2)}{2}. \end{aligned} \tag{4.2.44}$$

Since

$$a_k^{-3/2} A_{k+1} = \tfrac{2^{3/2}(k+1)(k+2)(k+3)}{6[(k+1)(k+2)]^{3/2}} = \tfrac{2^{1/2}(k+3)}{3[(k+1)(k+2)]^{1/2}} \geq \tfrac{2}{3},$$

inequality (4.2.43) leads to the following condition on the parameters:

$$\tfrac{1}{L_3+M} \geq \tfrac{2}{C}.$$

Hence, we can choose

$$M = 2L_3, \quad C = 2(L_3 + M) = 6L_3. \tag{4.2.45}$$

In this case $2L_3 + C = 8L_3$.

Now we are ready to put all the pieces together.

**Accelerated Cubic Regularization of Newton's Method**

**Initialization:** Choose $x_0 \in \mathbb{E}$. Set $M = 2L_3$ and $C = 6L_3$.

Compute $x_1 = T_{L_3}(x_0)$ and define $\psi_1(x) = f(x_1) + \frac{C}{6}\|x - x_0\|^3$.

**Iteration** $k$**,**$(k \geq 1)$**:**

**1.** Compute $v_k = \arg\min\limits_{x \in \mathbb{E}} \psi_k(x)$ and choose $y_k = \frac{k}{k+3} x_k + \frac{3}{k+3} v_k$.

**2.** Compute $x_{k+1} = T_M(y_k)$ and update

$$\psi_{k+1}(x) = \psi_k(x) + \tfrac{(k+1)(k+2)}{2} \cdot [f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle].$$

(4.2.46)

The above discussion proves the following theorem.

**Theorem 4.2.3** *If the sequence $\{x_k\}_{k=1}^{\infty}$ is generated by method (4.2.46) as applied to problem (4.2.23), then for any $k \geq 1$ we have:*

$$f(x_k) - f(x^*) \leq \frac{8\, L_3\, \|x_0 - x^*\|^3}{k(k+1)(k+2)}, \tag{4.2.47}$$

*where $x^*$ is an optimal solution to the problem.*

*Proof* Indeed, we have shown that

$$A_k f(x_k) \overset{\mathscr{R}_k^1}{\leq} \psi_k^* \overset{\mathscr{R}_k^2}{\leq} A_k f(x^*) + \frac{2L_3 + C}{6} \|x_0 - x^*\|^3.$$

Thus, (4.2.47) follows from (4.2.44) and (4.2.45). □

Note that the point $v_k$ can be found in (4.2.46) by a closed-form expression. Consider

$$s_k = \nabla \ell_k(x).$$

Since the function $\ell_k(x)$ is linear, this vector does not depend on $x$. Therefore,

$$v_k = x_0 - \sqrt{\frac{2}{C\|s_k\|_*}} \cdot B^{-1} s_k.$$

### 4.2.5 Global Non-degeneracy for Second-Order Schemes

Traditionally, in Numerical Analysis the term *non-degenerate* is applied to certain classes of efficiently solvable problems. For unconstrained optimization, non-degeneracy of the objective function is usually characterized by a uniform lower bound $\tau(f)$ on the angle between the gradient at point $x$ and the direction pointing towards the optimal solution:

$$\alpha(x) \overset{\text{def}}{=} \frac{\langle \nabla f(x), x - x^* \rangle}{\|\nabla f(x)\|_* \cdot \|x - x^*\|} \;\geq\; \tau(f) \;>\; 0, \quad x \in \mathbb{E}. \tag{4.2.48}$$

This condition has a nice geometric interpretation. Moreover, there exists a large class of smooth convex functions possessing this property. This is the class of strongly convex functions with Lipschitz-continuous gradient.

**Lemma 4.2.6** $\tau(f) \;\geq\; \frac{2\sqrt{\gamma_2(f)}}{1+\gamma_2(f)} \;>\; \sqrt{\gamma_2(f)}$.

*Proof* Indeed, in view of inequality (2.1.32), we have

$$\begin{aligned} \langle \nabla f(x), x - x^* \rangle &\geq \frac{1}{\sigma_2 + L_2} \|\nabla f(x)\|_*^2 + \frac{\sigma_2 L_2}{\sigma_2 + L_2} \|x - x^*\|^2 \\ &\geq \frac{2\sqrt{\sigma_2 L_2}}{\sigma_2 + L_2} \cdot \|\nabla f(x)\|_* \cdot \|x - x^*\|, \end{aligned}$$

and this proves the required inequality. □

Note that the efficiency bounds of the first-order schemes for the class of smooth strongly convex functions can be completely characterized in terms of the condition number $\gamma_2$. Indeed, on one hand, the lower complexity bound for finding an $\epsilon$-solution for problems from this problem class is proven to be

$$O\left(\frac{1}{\sqrt{\gamma_2}} \ln \frac{\sigma_2 D^2}{\epsilon}\right) \tag{4.2.49}$$

calls of the oracle, where the constant $D$ bounds the distance between the initial point and the optimal solution (see Theorem 2.1.13). On the other hand, the simple numerical scheme (2.2.20) exhibits the required rate of convergence (see Theorem 2.2.3).

What can be said about the complexity of the above problem class for the second-order schemes? Surprisingly enough, in this situation it is difficult to find any favorable consequences of the condition (4.2.48). We will discuss the complexity bounds for this problem class in detail later in Sect. 4.2.6. Now let us present a new non-degeneracy condition, which replaces (4.2.48) for the second-order methods.

Assume that $\gamma_3(f) = \frac{\sigma_3(f)}{L_3(f)} > 0$. In this case,

$$f(x) - f(x^*) \overset{(4.2.13)}{\leq} \frac{2}{3\sqrt{\sigma_3}} \cdot \|\nabla f(x)\|_*^{3/2}. \tag{4.2.50}$$

Therefore, for method (4.2.33) we have

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\overset{(4.2.29)}{\geq} \frac{1}{3\sqrt{L_3}} \|\nabla f(x_{k+1})\|_*^{3/2} \\ &\overset{(4.2.50)}{\geq} \frac{1}{2}\sqrt{\gamma_3(f)} \cdot (f(x_{k+1}) - f(x^*)). \end{aligned} \tag{4.2.51}$$

Hence, for any $k \geq 1$ we have

$$\begin{aligned} f(x_k) - f(x^*) &\overset{(4.2.51)}{\leq} \frac{f(x_1) - f^*}{\left(1 + \frac{1}{2}\sqrt{\gamma_3(f)}\right)^{k-1}} \\ &\overset{(4.2.30)}{\leq} e^{-\frac{\sqrt{\gamma_3(f)} \cdot (k-1)}{2 + \sqrt{\gamma_3(f)}}} \cdot \frac{L_3}{3} \|x_0 - x^*\|^3. \end{aligned} \tag{4.2.52}$$

Thus, the complexity of minimizing a function with positive condition number $\gamma_3(f)$ by method (4.2.33) is of the order of

$$O\left(\frac{1}{\sqrt{\gamma_3(f)}} \ln \frac{L_3 D^3}{\epsilon}\right) \tag{4.2.53}$$

calls of the oracle. The structure of this estimate is similar to that of (4.2.49). Hence, it is natural to say that such functions possess *global second-order non-degeneracy*.

Let us demonstrate that the accelerated variant of Newton's method (4.2.46) can be used to improve the complexity estimate (4.2.53). Denote by $\mathscr{A}_k(x_0)$ the point $x_k$ generated by method (4.2.46) with starting point $x_0$. Consider the following process.

$$\begin{array}{l} \textbf{1.}\ \text{Define } m = \left\lceil \left(\frac{24e}{\gamma_3(f)}\right)^{1/3} \right\rceil, \text{ and set } y_0 = x_0. \\[2ex] \textbf{2.}\ \text{For } k \geq 0, \text{ iterate } y_{k+1} = \mathscr{A}_m(y_k). \end{array} \tag{4.2.54}$$

The performance of this scheme can be derived from the following lemma.

**Lemma 4.2.7** *For any $k \geq 0$ we have*

$$\begin{array}{rcl} \|y_{k+1} - x^*\|^3 & \leq & \frac{1}{e}\|y_k - x^*\|^3, \\[2ex] f(y_{k+1}) - f(x^*) & \leq & \frac{1}{e}(f(y_k) - f(x^*)). \end{array} \tag{4.2.55}$$

*Proof* Indeed, since $m \geq \left(\frac{24e}{\gamma_3(f)}\right)^{1/3}$, we have

$$\begin{array}{rcl} \frac{1}{3}\sigma_3\|y_{k+1} - x^*\|^3 & \overset{(4.2.10)}{\leq} & f(y_{k+1}) - f(x^*) \\[2ex] & \overset{(4.2.47)}{\leq} & \frac{8L_3\|y_k - x^*\|^3}{m(m+1)(m+2)} \leq \frac{1}{3e}\sigma_3\|y_k - x^*\|^3 \\[2ex] & \overset{(4.2.10)}{\leq} & \frac{1}{e}(f(y_k) - f(x^*)). \end{array}$$ □

Thus,

$$f(T_{L_3}(y_k)) - f(x^*) \overset{(4.2.30)}{\leq} \frac{L_3}{3}\|y_k - x^*\|^3 \overset{(4.2.30)}{\leq} \frac{L_3}{3}\|y_0 - x^*\|^3 \cdot e^{-k},$$

and we conclude that an $\epsilon$-solution to our problem can be found by (4.2.54) in

$$O\left(\frac{1}{[\gamma_3(f)]^{1/3}} \ln\left[\frac{L_3}{\epsilon}\|x_0 - x^*\|^3\right]\right) \tag{4.2.56}$$

iterations. Lower complexity bounds for this problem class have not yet been developed. So, we cannot say how far these results are from the best possible ones.

### 4.2.6 Minimizing Strongly Convex Functions

Let us look now at the complexity of problem (4.2.23) with

$$\sigma_2(f) > 0, \quad L_3(f) < \infty. \tag{4.2.57}$$

The main advantage of such functions consists in quadratic convergence of Newton's method (4.2.33) in a certain neighborhood of the optimal solution. Indeed, for $T = T_{L_3}(x)$ we have

$$\begin{aligned} f(x) - f(T) &\overset{(4.2.28)}{\geq} \frac{1}{2}\langle \nabla^2 f(T)(T-x), T-x\rangle \geq \frac{\sigma_2}{2} \cdot r_{L_3}^2(x) \\ &\overset{(4.2.26)}{\geq} \frac{\sigma_2}{2L_3} \cdot \|\nabla f(T)\|_* \overset{(4.2.13)}{\geq} \frac{\sigma_2}{2L_3} \cdot \left[2\sigma_2(f(T) - f(x^*))\right]^{1/2}. \end{aligned} \tag{4.2.58}$$

Hence,

$$f(T) - f(x^*) \overset{(4.2.58)}{\leq} \tfrac{2L_3^2}{\sigma_2^3}(f(x) - f(T))^2 \leq \tfrac{2L_3^2}{\sigma_2^3}(f(x) - f(x^*))^2. \tag{4.2.59}$$

Therefore, the region of quadratic convergence of method (4.2.33) can be defined as

$$\mathscr{Q}_f = \left\{ x \in \mathbb{E} : \ f(x) - f(x^*) \leq \tfrac{\sigma_2^3}{2L_3^2} \right\}. \tag{4.2.60}$$

Alternatively, the region of quadratic convergence can be described in terms of the norm of the gradient. Indeed,

$$\begin{aligned} &\tfrac{\sigma_2}{2} \cdot r_{L_3}^2(x) \leq \tfrac{1}{2}\langle \nabla^2 f(T)(T-x), T-x\rangle \\ &\overset{(4.2.28)}{\leq} f(x) - f(T) \leq \|\nabla f(x)\|_* \cdot r_{L_3}(x). \end{aligned}$$

Thus,

$$\|\nabla f(x)\|_* \geq \tfrac{\sigma_2}{2} \cdot r_{L_3}(x) \overset{(4.2.26)}{\geq} \tfrac{\sigma_2}{2}\left[\tfrac{1}{L_3}\|\nabla f(T)\|_*\right]^{1/2}.$$

Consequently,

$$\|\nabla f(T)\|_* \leq \tfrac{4L_3}{\sigma_2^2}\|\nabla f(x)\|_*^2, \tag{4.2.61}$$

and the region of quadratic convergence can be defined as

$$\mathscr{Q}_g = \left\{ x \in \mathbb{E} : \ \|\nabla f(x)\|_* \leq \tfrac{\sigma_2^2}{4L_3} \right\}. \tag{4.2.62}$$

Thus, the global complexity of problem (4.2.23), (4.2.57) is mainly related to the number of iterations required to come from $x_0$ to the region $\mathcal{Q}_f$ (or, to $\mathcal{Q}_g$). For method (4.2.33), this value can be estimated from above by

$$O\left(\sqrt{\frac{L_3(f)D}{\sigma_2(f)}}\right), \tag{4.2.63}$$

where $D$ is defined by (4.2.34) (see Sect. 4.1). Let us show that, using the accelerated scheme (4.2.46), it is possible to improve this complexity bound.

Assume that we know an upper bound for the distance to the solution:

$$\|x_0 - x^*\| \;\leq\; R \quad (\leq D).$$

Consider the following process.

$$\begin{array}{l} \textbf{1. } \text{Set } y_0 = T_{L_3}(x_0), \text{ and define } m_0 = \left\lceil \frac{64 L_3(f) R}{\sigma_2(f)} \right\rceil^{1/3}. \\[2ex] \textbf{2. While } \|\nabla f(T_{L_3}(y_k))\|_* \geq \frac{\sigma_2^2}{4L_3} \textbf{ do } \{y_{k+1} = \mathcal{A}_{m_k}(y_k),\;\; m_{k+1} = \frac{1}{2^{1/3}}\, m_k\}. \end{array} \tag{4.2.64}$$

**Theorem 4.2.4** *The process (4.2.64) terminates at most after*

$$\frac{1}{\ln 4} \ln\left(\frac{8}{3}\cdot\left(\frac{L_3(f)R}{\sigma_2(f)}\right)^3\right) \tag{4.2.65}$$

*stages. The total number of Newton steps in all stages does not exceed* $4m_0$.

*Proof* Let $R_k = R\cdot\left(\frac{1}{2}\right)^k$. It is clear that

$$m_k \geq 4\left(\frac{L_3(f)R_k}{\sigma_2(f)}\right)^{1/3}, \quad k \geq 0. \tag{4.2.66}$$

For $k \geq 0$, let us prove by induction that

$$\|y_k - x^*\| \leq R_k. \tag{4.2.67}$$

Assume that for some $k \geq 0$ this statement is valid (it is true for $k = 0$). Then,

$$\begin{aligned} \frac{\sigma_2}{2}\|y_{k+1} - x^*\|^2 &\overset{(2.1.21)}{\leq} f(y_{k+1}) - f(x^*) \overset{(4.2.47)}{\leq} \frac{8L_3R_k^3}{m_k(m_k+1)(m_k+2)} \\ &\overset{(4.2.66)}{\leq} \frac{8}{64}\sigma_2 R_k^2 \;=\; \frac{1}{8}\sigma_2 R_k^2 \;=\; \frac{1}{2}\sigma_2 R_{k+1}^2. \end{aligned}$$

Thus, (4.2.67) is valid for all $k \geq 0$. On the other hand,

$$f(y_{k+1}) - f(x^*) \overset{(4.2.47)}{\leq} \tfrac{8L_3\|y_k - x^*\|^3}{m_k(m_k+1)(m_k+2)} \overset{(4.2.67)}{\leq} \tfrac{8L_3\|y_k - x^*\|^2 R_k}{m_k(m_k+1)(m_k+2)}$$

$$\overset{(4.2.66)}{\leq} \tfrac{1}{8}\sigma_2\|y_k - x^*\|^2 \overset{(2.1.21)}{\leq} \tfrac{1}{4}(f(y_k) - f(x^*)).$$

Hence

$$\tfrac{\sigma_2}{2L_3}\|\nabla f(T_{L_3}(y_k))\|_* \overset{(4.2.58)}{\leq} f(y_k) - f(T_{L_3}(y_k)) \leq f(y_k) - f(x^*)$$

$$\leq \left(\tfrac{1}{4}\right)^k (f(y_0) - f(x^*)) \overset{(4.2.30)}{\leq} \left(\tfrac{1}{4}\right)^k \tfrac{L_3}{3} R^3,$$

and (4.2.65) follows from (4.2.62). Finally, the total number of Newton steps does not exceed

$$\sum_{k=0}^{\infty} m_k = m_0 \sum_{k=0}^{\infty} \tfrac{1}{2^{k/3}} = \tfrac{m_0}{2^{1/3}-1} < 4m_0. \qquad \square$$

### 4.2.7 False Acceleration

Note that the properties of the class of smooth strongly convex functions (4.2.57) leave some space for erroneous conclusions related to the rate of convergence of the optimization methods at the first stage of the process, aiming to enter the region of quadratic convergence. Let us demonstrate this with a particular example.

Consider a modified version $\mathscr{M}'$ of method (4.2.46). The only modification is introduced in Step 2. Now it is as follows:

**2'**. Compute $\hat{y}_k = T_M(y_k)$ and update

$$\psi_{k+1}(x) = \psi_k(x) + \tfrac{(k+1)(k+2)}{2} \cdot [f(\hat{y}_k) + \langle \nabla f(\hat{y}_k), x - \hat{y}_k \rangle].$$

Choose $\hat{x}_k : \ f(\hat{x}_k) = \min\{f(x_k), f(\hat{y}_k)\}$. Set $x_{k+1} = T_M(\hat{x}_k)$.

(4.2.68)

Note that for $\mathscr{M}'$ the statement of Theorem 4.2.3 is valid. Moreover, the process now becomes monotone, and, using the same reasoning as in (4.2.58) with $M = 2L_3$, we obtain

$$f(x_k) - f(x_{k+1}) \geq f(\hat{x}_k) - f(x_{k+1}) \geq \tfrac{\sqrt{2}\,\sigma_2^{3/2}}{3L_3} \cdot [f(x_{k+1}) - f(x^*)]^{1/2}. \tag{4.2.69}$$

Further, let us fix the number of steps $N$. Define $\hat{k} = \frac{2}{3}N$. Then, in view of (4.2.47), we can guarantee that

$$f(x_{\hat{k}}) - f(x^*) \le \left(\frac{3}{2}\right)^3 \frac{8L_3R^3}{N^3} = 3^3\frac{L_3R^3}{N^3}. \tag{4.2.70}$$

On the other hand

$$\begin{aligned} f(x_{\hat{k}}) - f(x^*) &\ge f(x_{\hat{k}}) - f(x_{N+1}) \\ &\overset{(4.2.69)}{\ge} \frac{1}{3}N \cdot \frac{\sqrt{2}\,\sigma_2^{3/2}}{3L_3} \cdot [f(x_{N+1}) - f(x^*)]^{1/2}. \end{aligned} \tag{4.2.71}$$

Combining (4.2.70) and (4.2.71) we obtain

$$f(x_{N+1}) - f(x^*) \le \frac{3^{10}\cdot L_3^4 \cdot R^6}{2\sigma_2^3} \cdot N^{-8}. \tag{4.2.72}$$

As compared with the rate of convergence (4.2.47), the proposed modification looks amazingly efficient. However, that is just an illusion. Indeed, in view of (4.2.60), in order to enter the region of quadratic convergence of Newton's method, we need to make the right-hand-side of inequality (4.2.72) smaller than $\frac{\sigma_2^3}{2L_3^2}$. For that we need

$$O\left(\left[\frac{L_3R}{\sigma_2}\right]^{3/4}\right) \tag{4.2.73}$$

iterations of $\mathscr{M}'$. This is much worse than the complexity estimate (4.2.63) of the basic scheme (4.2.33) even without acceleration (4.2.46).

Another clarification comes from an estimate for the number of steps, which is necessary for $\mathscr{M}'$ to halve the distance to the minimum. From (4.2.72) we see that it needs $O\left(\left[\frac{L_3R}{\sigma_2}\right]^{1/2}\right)$ iterations, which is worse than the corresponding estimate for the method (4.2.46).

### 4.2.8 Decreasing the Norm of the Gradient

Let us check now our ability to generate points with small norm of the gradient using second-order methods (compare with Sect. 2.2.2). We first look at the simplest method (4.2.33).

Denote by $T$ the total number of iterations of this scheme. For the sake of simplicity, let us assume that $T = 3m + 2$ for some integer $m \ge 0$. Let us divide all iterations of the method into two parts. For the first part of length $2m$ we have

$$f(x_{2m}) - f^* \overset{(4.2.35)}{\leq} \frac{9L_3D^3}{4(m+2)^2},$$

where $L_3 = L_3(f)$. For the second part of length $m+2$, we have

$$f(x_{2m}) - f(x_T) = \sum_{k=0}^{m+1} (f(x_{2m+k}) - f(x_{2m+k+1})) \overset{(4.2.29)}{\geq} \frac{m+2}{3L_3^{1/2}}(g_T^*)^{3/2},$$

where $g_T^* = \min\limits_{1 \leq k \leq T} \|\nabla f(x_k)\|_*$. Thus,

$$g_T^* \leq \left(\frac{27L_3^{3/2}D^3}{4(m+2)^3}\right)^{2/3} = \frac{3^4L_3D^3}{2^{4/3}(T+4)^2}. \tag{4.2.74}$$

Let us look now at the monotone version of the accelerated Cubic Newton Method (4.2.46), (4.2.68). Let $R_0 = \|x_0 - x^*\|$. Let $T = 4m$ for some integer $m \geq 1$. Then, for the first $3m$ iterations of this method we have

$$f(x_{3m}) - f^* \overset{(4.2.47)}{\leq} \frac{8L_3R_0^3}{3m(3m+1)(3m+2)}.$$

For the second part of length $m$, we have

$$f(x_{3m}) - f(x_T) = \sum_{k=0}^{m-1} (f(x_{3m+k}) - f(x_{3m+k+1})) \overset{(4.2.29)}{\geq} \frac{m}{3L_3^{1/2}}(g_T^*)^{3/2}.$$

Thus,

$$g_T^* \leq \left(\frac{8L_3^{3/2}R_0^3}{m^2(3m+1)(3m+2)}\right)^{2/3} < \frac{2^8L_3R_0^2}{T^{8/3}}. \tag{4.2.75}$$

Finally, let us check what can be achieved with the regularization technique. As in Sect. 2.2.2, we fix a regularization parameter $\delta > 0$ and introduce the following function:

$$f_\delta(x) = f(x) + \tfrac{1}{3}\delta\|x - x_0\|^3.$$

Let $D = \max\limits_{x \in \mathbb{E}}\{\|x - x_0\| : \ f(x) \leq f(x_0)\}$. Since $f_\delta(x) \geq f(x)$ for all $x \in \mathbb{E}$, inequality $f_\delta(x) \leq f(x_0)$ implies $\|x - x_0\| \leq D$.

In view of Lemmas 4.2.3 and 4.2.4, we have

$$\sigma_3(f_\delta) = \tfrac{1}{2}\delta, \quad L_3(f_\delta) = L_3 + 2\delta.$$

Thus, $\gamma_3(f_\delta) = \frac{\delta}{2L_3+4\delta}$.

Let $x_\delta^* = \arg\min\limits_{x \in \mathbb{E}} f_\delta(x)$ and let $m = \left\lceil \left( 24e \left( 4 + \frac{2L_3}{\delta} \right) \right)^{1/3} \right\rceil$. In view of Lemma 4.2.7, restarting strategy (4.2.54) ensures the following rate of convergence:

$$f_\delta(y_{k+1}) - f_\delta(x_\delta^*) \;\leq\; \tfrac{1}{e}(f_\delta(y_k) - f_\delta(x_\delta^*)),$$

where $y_0 = T_{L_3}(x_0)$. Thus, $f_\delta(y_k) - f_\delta(x_\delta^*) \overset{(4.1.11)}{\leq} \frac{1}{3e^k} L_3(f) D^3$.

Define $y_k^* = T_{L_3(f_\delta)}(y_k)$. Then $f_\delta(y_k^+) \leq f_\delta(y_k) \leq f(x_0)$. Hence, $\| y_k^+ - x_0 \| \leq D$ and we have

$$\begin{aligned} \|\nabla f(y_k^+)\|_* \;&\leq\; \|\nabla f_\delta(y_k^+)\|_* + \delta D^2 \\ &\overset{(4.2.29)}{\leq} \left[ 3L_3^{1/2}(f_\delta) \cdot \left( f_\delta(y_k) - f_\delta(x_\delta^*) \right) \right]^{2/3} + \delta D^2 \\ &\leq\; \tfrac{1}{e^{2k/3}} L_3 D^2 \sqrt{1 + \tfrac{2\delta}{L_3}} + \delta D^2. \end{aligned}$$

Let us choose now $\delta = \frac{\epsilon}{2D^2}$. Define $\varkappa = \frac{L_3 D^2}{\epsilon}$. Then, to ensure $\|\nabla f(y_k^+)\|_* \leq \epsilon$, we need to perform

$$k \geq \tfrac{3}{2} \ln \left( 2\sqrt{\varkappa^2 + \varkappa} \right)$$

iterations of the restarting strategy (4.2.54). Each cycle of this strategy needs $\left\lceil 2\,(12e(1+\varkappa))^{1/3} \right\rceil$ iterations of the Accelerated Cubic Newton Method (4.2.46). Thus, we get a bound which is asymptotically better than the simple estimate (4.2.75). However, it seems that for all practical values of the accuracy, the method (4.2.46), (4.2.68) has better performance guarantees.

### *4.2.9 Complexity of Non-degenerate Problems*

**1.** From the complexity results presented in the previous sections, we can derive a class of problems which are *easy* for the second-order schemes:

$$\sigma_2(f) \;>\; 0, \quad \sigma_3(f) \;>\; 0, \quad L_3(f) \;<\; \infty. \tag{4.2.76}$$

For such functions, the second-order methods exhibit a global linear rate of convergence and a local quadratic convergence. In accordance with (4.2.56) and (4.2.60), we need

$$O\left( \left[ \tfrac{L_3(f)}{\sigma_3(f)} \right]^{1/3} \ln \left[ \tfrac{L_3(f)}{\sigma_2(f)} \| x_0 - x^* \| \right] \right) \tag{4.2.77}$$

iterations of (4.2.46) to enter the region of quadratic convergence.

Note that the class (4.2.76) is non-trivial. It contains, for example, all functions

$$\xi_{\alpha,\beta}(x) = \alpha d_2(x) + \beta d_3(x), \quad \alpha, \beta > 0,$$

with parameters

$$\sigma_2(\xi_{\alpha,\beta}) = \alpha, \quad \sigma_3(\xi_{\alpha,\beta}) = \frac{1}{2}\beta, \quad L_3(\xi_{\alpha,\beta}) = 2\beta.$$

Moreover, any convex function with Lipschitz-continuous Hessian can be *regularized* by adding an auxiliary function $\xi_{\alpha,\beta}$.

**2.** For one important class of convex problems, namely, for problems with

$$\sigma_2(f) > 0, \quad L_2(f) < \infty, \quad L_3(f) < \infty, \tag{4.2.78}$$

we have actually failed to clarify the situation. The standard theory of optimal *first-order* methods (see Sect. 2.2) can bound the number of iterations which are required to enter the region of quadratic convergence (4.2.60), as follows:

$$O\left(\left[\frac{L_2(f)}{\sigma_2(f)}\right]^{1/2} \ln\left[\frac{L_2(f)L_3^2(f)}{\sigma_2^3(f)}\|x_0 - x^*\|^2\right]\right). \tag{4.2.79}$$

Note that in this estimate the role of the second-order scheme is quite weak: it is used only to establish the bounds of the termination stage. Of course, as is shown in Sect. 4.2.6, we could also use it at the first stage. However, in this case the size of the optimal solution $x^*$ enters *polynomially* the estimate for the number of iterations. Thus, the following question is still open:

> Can we get any advantage from the second-order schemes being used at the initial stage of minimization process as applied to a function from the problem class (4.2.78)?

We will come back to the complexity of problem class (4.2.78) again in Sect. 5.2, when we will discuss our possibilities in minimizing self-concordant functions.

## 4.3 Optimal Second-Order Methods

### *4.3.1 Lower Complexity Bounds*

Let us derive lower complexity bounds for the second-order methods as applied to the problem

$$f^* = \min_{x \in \mathbb{R}^n} f(x), \tag{4.3.1}$$

where the Hessian of the objective function is Lipschitz continuous. We assume that this problem is solvable and $x^*$ is its optimal solution.

For the sake of simplicity, as we did in Sect. 2.1.2 (see Assumption 2.1.4), let us first fix the natural rules for generating the test points. It can be easily checked that the second-order methods usually compute the next test point as follows:

$$x_{k+1} = x_k - h_k[\alpha_k I_n + (1-\alpha_k)\nabla^2 f(x_k)]^{-1}\nabla f(x_k),$$

where $h_k > 0$ is a step-size parameter, and the coefficient $\alpha_k \in [0, 1]$ depends on a particular optimization scheme. In the case $\alpha_k = 1$, we get the usual Gradient Method. The case $\alpha_k = 0$ corresponds to the standard Newton direction. Finally, the Cubic Regularization strategy (4.2.24) and the majority of Trust Region Methods compute these values from some equation (see, for example, (4.2.25)). Therefore, the following assumption looks quite reasonable.

**Assumption 4.3.1** *All iterative second-order schemes generate a sequence of test points* $\{x_k\}_{k\geq 0}$ *such that*

$$x_{k+1} \in x_0 + Lin\left\{\mathscr{G}_f(x_0), \ldots, \mathscr{G}_f(x_k)\right\}, \quad k \geq 0, \tag{4.3.2}$$

*where* $\mathscr{G}_f(x) = cl\ \left(Conv\ \left\{[\alpha I_n + (1-\alpha)\nabla^2 f(x)]^{-1}\nabla f(x),\ \alpha \in [0,1)\right\}\right)$.

Note that the set $\mathscr{G}_f(x)$ also contains $\nabla f(x)$. Therefore, the rules for computing the point $v_k$ in the accelerated method (4.2.46) also satisfy condition (4.3.2).

For $2 \leq k \leq n$, consider the following parametric family of functions:

$$f_k(x) = \tfrac{1}{3}\left\{\sum_{i=1}^{k-1} |x^{(i)} - x^{(i+1)}|^3 + \sum_{i=k}^{n} |x^{(i)}|^3\right\} - x^{(1)}, \quad x \in \mathbb{R}^n. \tag{4.3.3}$$

This is a uniformly convex function, and its unique minimum can be found from the following system of equations:

$$\begin{aligned} (x^{(1)} - x^{(2)})|x^{(1)} - x^{(2)}| &= 1, \\ (x^{(i)} - x^{(i-1)})|x^{(i)} - x^{(i-1)}| + (x^{(i)} - x^{(i+1)})|x^{(i)} - x^{(i+1)}| &= 0, \quad 2 \leq i \leq k-1, \\ (x^{(k)} - x^{(k-1)})|x^{(k)} - x^{(k-1)}| + x^{(k)}|x^{(k)}| &= 0, \\ x^{(i)}|x^{(i)}| &= 0, \quad k+1 \leq i \leq n. \end{aligned}$$

Clearly, the only solution of this system is given by vector $x_*$ with coordinates

$$x_*^{(i)} = (k-i+1)_+, \quad i = 1, \ldots, n, \tag{4.3.4}$$

where $(\tau)_+ = \max\{\tau, 0\}$. For our methods, we always take $x_0 = 0$. Therefore, we have the following characteristics of our problem (4.3.1) with $f = f_k$:

$$\begin{aligned} f_k^* &= -\tfrac{2}{3}k, \\ R_k^2 &= \|x_0 - x_*\|_{(2)}^2 \;=\; \sum_{i=1}^{k} i^2 \;<\; \tfrac{(k+1)^3}{3}. \end{aligned} \tag{4.3.5}$$

It remains to estimate the Lipschitz constant of the Hessian of the function $f_k$ with respect to the standard Euclidean norm.

Let us look first at the Hessian of the following function

$$\rho_3(u) = \tfrac{1}{3}\sum_{i=1}^{n} |u^{(i)}|^3, \quad u \in \mathbb{R}^n.$$

For a direction $h \in \mathbb{R}^n$, we have $\langle \nabla^2 \rho_3(u)h, h\rangle = 2\sum\limits_{i=1}^{n} |u^{(i)}|\,(h^{(i)})^2$. Therefore, for $u, v \in \mathbb{R}^n$ we get

$$\left|\langle(\nabla^2\rho_3(u) - \nabla^2\rho_3(v))h, h\rangle\right| = 2\left|\sum_{i=1}^{n}(|u^{(i)}| - |v^{(i)}|)(h^{(i)})^2\right| \;\le\; 2\|u - v\|_{(\infty)}\|h\|_{(2)}^2.$$

Note that function $f_k(\cdot)$ can be represented as follows:

$$f_k(x) = \rho_3(B_k x) - x^{(1)}, \quad B_k = \begin{pmatrix} A_k & 0 \\ 0 & I_{n-k} \end{pmatrix} \in \mathbb{R}^{n\times n},$$

where the upper bi-diagonal matrix $A_k \in \mathbb{R}^{k\times k}$ has the following structure:

$$A_k = \begin{pmatrix} 1 & -1 & 0 \dots & 0 \\ 0 & 1 & -1 \dots & 0 \\ & & \dots & 0 \\ & & \dots & -1 \\ 0 \dots & \dots & 0 & 1 \end{pmatrix}.$$

Therefore, for any point $x$, displacement $d$, and direction $h$ in $\mathbb{R}^n$ we have

$$\begin{aligned} \left|\langle(\nabla^2 f_k(x+d) - \nabla^2 f_k(x))h, h\rangle\right| &= \left|\langle(\nabla^2\rho_3(B_k(x+d)) - \nabla^2\rho_3(B_k x))B_k h, B_k h\rangle\right| \\ &\le 2\|B_k d\|_{(\infty)}\|B_k h\|_{(2)}^2. \end{aligned}$$

Note that for any $h \in \mathbb{R}^n$ we have

$$\begin{aligned}\| B_k d \|_{(\infty)} &\le \max_{1 \le i \le n-1} \{ |d^{(i)}| + |d^{(i+1)}| \} \;\le\; \max_{1 \le i \le n-1} \sqrt{2[(d^{(i)})^2 + (d^{(i+1)})^2]} \\ &\le 2^{1/2} \| d \|_{(2)},\end{aligned}$$

$$\| B_k h \|^2_{(2)} = \sum_{i=1}^{k-1} (h^{(i)} - h^{(i+1)})^2 + \sum_{i=k}^{n} (h^{(i)})^2 \;\le\; 4 \| h \|^2_{(2)}.$$

Thus, we conclude that

$$\| \nabla^2 f_k(x+d) - \nabla^2 f_k(x) \| \le 8\sqrt{2} \| d \|_{(2)},$$

and we can take the Lipschitz constant for the Hessian of this function $L = 2^{7/2}$.

In order to understand the behavior of numerical schemes satisfying condition (4.3.2), as applied to minimization of some function $f_t$ with $t$ big enough, we need to introduce the following subspaces (compare with Sect. 2.1.2):

$$\mathbb{R}^{k,n} = \{ x \in \mathbb{R}^n : \; x^{(i)} = 0 \text{ for } i > k \}, \quad 1 \le k \le n-1,$$

$$\mathbb{S}^{k,n} = \{ H \in \mathbb{R}^{n \times n} : \; H = H^T, \; H^{(i,j)} = 0 \text{ if } i \ne j \text{ and } (i > k \text{ or } j > k) \}.$$

Let us write down the first and the second derivatives of the function $f_t$ along direction $h \in \mathbb{R}^n$ (see (4.3.3):

$$\begin{aligned}\langle \nabla f_t(x), h \rangle &= \sum_{i=1}^{t-1} |x^{(i)} - x^{(i+1)}| (x^{(i)} - x^{(i+1)})(h^{(i)} - h^{(i+1)}) \\ &\quad + \sum_{i=t}^{n} |x^{(i)}| x^{(i)} h^{(i)} - h^{(1)},\end{aligned}$$

$$\langle \nabla^2 f_t(x) h, h \rangle = 2 \sum_{i=1}^{t-1} |x^{(i)} - x^{(i+1)}| (h^{(i)} - h^{(i+1)})^2 + 2 \sum_{i=t}^{n} |x^{(i)}| (h^{(i)})^2. \tag{4.3.6}$$

From this structure, we derive the following important conclusions.

**Lemma 4.3.1** *If $x \in \mathbb{R}^{i,n}$ and $i < k$, then $\nabla f_t(x) \in \mathbb{R}^{i+1,n}$ and $\nabla^2 f_t(x) \in \mathbb{S}^{i+1,n}$.* □

**Corollary 4.3.1** *Let $x_i \in \mathbb{R}^{i,n}$, $i = 0, \ldots, k$, and suppose the point $x_{k+1}$ satisfies condition (4.3.2) with $f(\cdot) = f_t(\cdot)$, where $k+1 \le t \le n$. Then $x_{k+1} \in \mathbb{R}^{k+1,n}$.*

*Proof* Indeed, in view of Lemma 4.3.1, we have

$$\nabla f_t(x_i) \in \mathbb{R}^{i+1,n} \subset \mathbb{R}^{k+1,n}, \quad \nabla^2 f_t(x_i) \in \mathbb{S}^{i+1,n} \subset \mathbb{S}^{k+1,n}, \quad i = 0, \ldots, k.$$

Therefore,

$$[\alpha I_n + (1-\alpha)\nabla^2 f_t(x_i)]^{-1}\nabla f_t(x_i) \in \mathbb{R}^{k+1,n}$$

for all $\alpha \in [0,1)$ and $i = 0,\dots,k$. □

Our last observation is as follows.

**Lemma 4.3.2** *For any $p \geq 0$ and $x \in \mathbb{R}^{k,n}$, we have $f_{k+p}(x) = f_k(x)$.* □

Now we can prove the lower complexity bound for the second-order methods.

**Theorem 4.3.1** *Let the Hessian of the objective function $f$ in problem (4.3.1) be Lipschitz continuous with constant $L_f$. Assume that the rules of a second-order method $\mathscr{M}$ satisfy condition (4.3.2), and for any starting point $x_0$ with $\|x_0 - x^*\|_{(2)} \leq \rho_0$ we can guarantee that*

$$\min_{0\leq i\leq k} f(x_i) - f(x^*) \leq \tfrac{L_f\rho_0^3}{C_{\mathscr{M}}(k)}, \tag{4.3.7}$$

*where $k$ is the number of generated test points. Then for $k = 3m+2$ with integer $m$, $0 \leq m \leq \frac{n}{4} - 1$, we have*

$$C_{\mathscr{M}}(k) \leq 36(k+1)^{3.5}. \tag{4.3.8}$$

*Proof* Let $k = 3m+2$ for some integer $m \geq 0$. Define $t = 4m+3$. Then

$$k+1 = 3(m+1), \quad t+1 \; = \; 4(m+1).$$

Let us apply method $\mathscr{M}$ for minimizing the function $f_t(\cdot)$ starting from the point $x_0 = 0$. Note that $\nabla f_t(x_0) = -e_1 \in \mathbb{R}^{1,n}$ and $\nabla^2 f_t(x_0) = 0$. Therefore, $x_1 \overset{(4.3.2)}{\in} \mathbb{R}^{1,n}$, and by induction, using Corollary 4.3.1, we get $x_k \overset{(4.3.2)}{\in} \mathbb{R}^{k,n}$, $0 \leq k \leq t$. Hence, by Lemma 4.3.2, we have

$$\begin{aligned}\tfrac{2}{3}(m+1) &\overset{(4.3.5)}{=} f_k^* - f_t^* \; \leq \; \min_{0\leq i\leq k} f_t(x_i) - f_t^* \overset{(4.3.7)}{\leq} \tfrac{L_f\rho_0^3}{C_{\mathscr{M}}(k)}\\ &\overset{(4.3.5)}{\leq} \tfrac{2^{7/2}}{C_{\mathscr{M}}(k)}\left(\tfrac{(t+1)^3}{3}\right)^{3/2}.\end{aligned}$$

Thus,

$$\begin{aligned}C_{\mathscr{M}}(k) &\leq \frac{2^{5/2}(t+1)^{9/2}}{(m+1)3^{1/2}} = \frac{2^{5/2}3^{1/2}}{k+1}\left(\frac{4}{3}(k+1)\right)^{9/2} = \frac{2^{23/2}}{3^4}(k+1)^{3.5}\\ &< 36(k+1)^{3.5}. \quad □\end{aligned}$$

As we can see, the lower bound (4.3.8) is a little bit better than the rate of convergence (4.2.47) of the Accelerated Cubic Regularization (4.2.46). In the next section, we will discuss the possibility of reaching this lower bound.

### *4.3.2 A Conceptual Optimal Scheme*

As in Sect. 4.2.3, let us fix a self-adjoint positive definite operator $B : \mathbb{E} \to \mathbb{E}^*$ and define primal and dual Euclidean norms

$$\|x\| = \langle Bx, x\rangle^{1/2}, \quad \|g\|_* = \langle g, B^{-1}g\rangle^{1/2}, \quad x \in \mathbb{E}, \; g \in \mathbb{E}^*.$$

Consider the problem of unconstrained optimization

$$\min_{x\in\mathbb{E}} f(x), \tag{4.3.9}$$

where the Hessian of the function $f$ satisfies the Lipschitz condition

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \le M_f \|x - y\|, \quad \forall x, y \in \mathbb{E}. \tag{4.3.10}$$

Our main iteration will be the *Cubic Newton Step*

$$\begin{aligned} T_M(x) = \arg\min_{T\in\mathbb{E}} \Big\{ & \langle \nabla f(x), T - x\rangle + \tfrac{1}{2}\langle \nabla^2 f(x)(T - x), T - x\rangle \\ & + \tfrac{M}{6}\|T - x\|^3 \Big\}. \end{aligned} \tag{4.3.11}$$

Let $r_M(x) = \|T_M(x) - x\|$. Then the point $T = T_M(x)$ is characterized by the following first-order optimality condition:

$$\nabla f(x) + \nabla^2 f(x)(T - x) + \tfrac{1}{2} M r_M(x)\, B(T - x) = 0. \tag{4.3.12}$$

**Lemma 4.3.3** *For any $x \in \mathbb{E}$ we have*

$$\langle \nabla f(T_M(x)), x - T_M(x)\rangle \ge \frac{1}{M r_M(x)}\|\nabla f(T_M(x))\|_*^2 + \frac{M^2 - M_f^2}{4M} r_M^3(x). \tag{4.3.13}$$

*Moreover, if $M \ge \frac{1}{\sigma} M_f$ for some $\sigma \in (0, 1]$, then*

$$\langle \nabla f(T_M(x)), x - T_M(x)\rangle \ge \frac{1}{M r_M(x)}\|\nabla f(T_M(x))\|_*^2 + \frac{1-\sigma^2}{4} M r_M^3(x). \tag{4.3.14}$$

*Proof* Let $T = T_M(x)$. Then

$$\begin{aligned} \frac{M_f^2 r_M^4(x)}{4} &\overset{(4.3.10)}{\geq} \|\nabla f(T) - \nabla f(x) - \nabla^2 f(x)(T-x)\|_*^2 \\ &\overset{(4.3.12)}{=} \|\nabla f(T) + \tfrac{1}{2} M r_M(x) B(T-x)\|_*^2 \\ &= \|\nabla f(T)\|_*^2 + M r_M(x) \langle \nabla f(T), T - x \rangle + \frac{M^2 r_M^4(x)}{4}. \end{aligned}$$

This is (4.3.13). Inequality (4.3.14) follows from (4.3.13) since $M_f \leq \sigma M$. □

Let us consider now the following conceptual version of the Optimal Cubic Newton Method.

**Optimal Cubic Newton Method (Conceptual Version)**

**Initialization.** Choose $x_0 \in \mathbb{E}$, $\sigma \in (0, 1)$. Define $\psi_0(x) = \frac{1}{2}\|x - x_0\|^2$.

Set $A_0 = 0$ and $M = \frac{1}{\sigma} M_f$.

***k*th iteration** $(k \geq 0)$.

(a) Compute $v_t = \arg\min\limits_{x \in \mathbb{E}} \psi_k(x)$.

(b) Choose $\rho_k > 0$ and find $a_{k+1} > 0$ from equation $a_{k+1}^2 = \frac{2(A_k + a_{k+1})}{M\rho_k}$.

(c) Set $A_{k+1} = A_k + a_{k+1}$, $\tau_k = \frac{a_{k+1}}{A_{k+1}}$, $y_k = (1 - \tau_k)x_k + \tau_k v_k$.

(d) Compute $x_{k+1} = T_M(y_k)$ and define

$\psi_{k+1}(x) = \psi_k(x) + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle]$.

(4.3.15)

Step (b) of method (4.3.15) is not completely specified since the definition of the parameter $\rho_k$ is missing. This is the reason why we call this method conceptual. Let us present some guidelines for its choice.

**Lemma 4.3.4** *Assume that parameters $\rho_k$ in method (4.3.15) satisfy condition*

$$r_M(y_k) \leq \rho_k. \tag{4.3.16}$$

*Then for any* $k \geq 0$ *we have*

$$A_k f(x_k) + B_k \leq \psi_k^* \overset{\text{def}}{=} \min_{x \in \mathbb{E}} \psi_k(x), \tag{4.3.17}$$

*where* $B_k = \frac{1-\sigma^2}{4} M \sum\limits_{i=0}^{k-1} A_{i+1} r_M^3(y_i)$.

*Proof* Let us prove (4.3.17) by induction. For $t = 0$ it is trivial. Assume that inequality (4.3.17) is valid for some $k \geq 0$. Then for any $x \in \mathbb{E}$ we have

$$\begin{aligned}
\psi_{k+1}(x) &\geq \psi_k^* + \tfrac{1}{2}\|x - v_k\|^2 + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle] \\
&\overset{(4.3.17)}{\geq} A_k f(x_k) + B_k + \tfrac{1}{2}\|x - v_k\|^2 \\
&\quad + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle] \\
&\geq A_{k+1} f(x_{k+1}) + B_k + \tfrac{1}{2}\|x - v_k\|^2 \\
&\quad + \langle \nabla f(x_{k+1}), A_k(x_k - x_{k+1}) + a_{k+1}(x - x_{k+1})\rangle \\
&= A_{k+1} f(x_{k+1}) + B_k + \tfrac{1}{2}\|x - v_k\|^2 \\
&\quad + \langle \nabla f(x_{k+1}), a_{k+1}(x - v_k) + A_{k+1}(y_k - x_{k+1})\rangle.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\psi_{k+1}^* &\geq A_{k+1} f(x_{k+1}) + B_k - \tfrac{1}{2} a_{k+1}^2 \|\nabla f(x_{k+1})\|_*^2 \\
&\quad + A_{k+1}\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle \\
&\overset{(4.3.14)}{\geq} A_{k+1} f(x_{k+1}) + B_k - \tfrac{A_{k+1}}{M\rho_k} \|\nabla f(x_{k+1})\|_*^2 \\
&\quad + A_{k+1}\left(\tfrac{1}{M r_M(y_k)} \|\nabla f(x_{k+1})\|_*^2 + \tfrac{1-\sigma^2}{4} M r_M^3(y_k)\right) \\
&\overset{(4.3.16)}{\geq} A_{k+1} f(x_{k+1}) + B_k + \tfrac{1-\sigma^2}{4} M A_{k+1} r_M^3(y_k).
\end{aligned}$$

□

In order to ensure a fast growth of the coefficients $A_k$, we need to introduce more conditions for the parameters $\rho_k$.

**Lemma 4.3.5** *Let us choose $\gamma \geq 1$. Assume that parameters $\rho_k$ in method (4.3.15) satisfy condition*

$$r_M(y_k) \leq \rho_k \ \leq \ \gamma r_M(y_k). \tag{4.3.18}$$

*Then for any $k \geq 1$ we have*

$$A_k \geq \frac{1}{4}\left(\frac{1}{\gamma}\right)^{3/2} \frac{\sqrt{1-\sigma^2}}{M\|x_0-x^*\|} \left(\frac{2k+1}{3}\right)^{3.5}. \tag{4.3.19}$$

*Proof* First of all, let us relate the rate of growth of coefficients $A_k$ to the values $r_M(y_k)$. Note that

$$A_{k+1}^{1/2} - A_k^{1/2} = \frac{a_{k+1}}{A_{k+1}^{1/2}+A_k^{1/2}} \ = \ \frac{1}{A_{k+1}^{1/2}+A_k^{1/2}}\sqrt{\frac{2A_{k+1}}{M\rho_k}} \ \geq \ \sqrt{\frac{1}{2M\rho_k}}.$$

Thus,

$$A_k \geq \frac{1}{2M}\left(\sum_{i=0}^{k-1} \frac{1}{\rho_i^{1/2}}\right)^2 \overset{(4.3.18)}{\geq} \ \frac{1}{2M\gamma}\left(\sum_{i=0}^{k-1} \frac{1}{r_M^{1/2}(y_i)}\right)^2. \tag{4.3.20}$$

On the other hand, we have $A_k f(x_k) + B_k \overset{(4.3.17)}{\leq} A_k f(x^*) + \frac{1}{2}\|x_0 - x^*\|^2$. Therefore,

$$B_k \ \equiv \ \frac{1-\sigma^2}{4} M \sum_{i=0}^{k-1} A_{i+1} r_M^3(y_i) \leq \frac{1}{2}\|x_0 - x^*\|^2.$$

Let us estimate from below the value $\sum\limits_{i=0}^{k-1} \frac{1}{r_M^{1/2}(y_i)}$ subject to the above constraint. Defining $\xi_i \ = r_M^{1/2}(y_i)$ and $D \ = \ \frac{2}{(1-\sigma^2)M}\|x_0 - x^*\|^2$, we come to the following minimization problem:

$$\xi^* \ = \ \min_{\xi \in \mathbb{R}^k} \left\{ \sum_{i=0}^{k-1} \frac{1}{\xi_i} : \quad \sum_{i=0}^{k-1} A_{i+1}\xi_i^6 \leq D \right\}.$$

Introducing a Lagrange multiplier $\lambda$ for the inequality constraint, we get the following optimality conditions:

$$\frac{1}{\xi_i^2} = \lambda A_{i+1}\xi_i^5, \quad i = 0, \ldots, k-1.$$

Thus, $\xi_i = \left(\frac{1}{\lambda A_{i+1}}\right)^{1/7}$. Since the constraint is active,

$$D = \sum_{i=0}^{k-1} A_{i+1} \left(\frac{1}{\lambda A_{i+1}}\right)^{6/7} = \frac{1}{\lambda^{6/7}} \sum_{i=0}^{k-1} A_{i+1}^{1/7}.$$

Therefore, $\xi^* = \sum\limits_{i=0}^{k-1} (\lambda A_{i+1})^{1/7} = \frac{1}{D^{1/6}} \left(\sum\limits_{i=0}^{k-1} A_{i+1}^{1/7}\right)^{7/6}$. Coming back to our initial notation, we get

$$\sum_{i=0}^{k-1} \frac{1}{r_M^{1/2}(y_i)} \geq \left(\frac{(1-\sigma^2)M}{2\|x_0-x^*\|^2}\right)^{1/6} \left(\sum_{i=0}^{k-1} A_{i+1}^{1/7}\right)^{7/6}.$$

In view of inequality (4.3.20), we come to the following relation:

$$A_k \geq \frac{1}{2\gamma} \left(\frac{1-\sigma^2}{2M^2\|x_0-x^*\|^2}\right)^{1/3} \left(\sum_{i=1}^{k} A_i^{1/7}\right)^{7/3}, \quad k \geq 1. \tag{4.3.21}$$

Denote the coefficient in the right-hand side of inequality (4.3.21) by $\theta$ and let $C_k = \left(\sum\limits_{i=1}^{k} A_i^{1/7}\right)^{2/3}$. Then (4.3.21) can be rewritten as

$$C_{k+1}^{3/2} - C_k^{3/2} \geq \theta^{1/7} C_{k+1}^{1/2}.$$

This means that $C_1 \geq \theta^{1/7}$ and

$$\begin{aligned}
\theta^{1/7} C_{k+1}^{1/2} &\leq (C_{k+1}^{1/2} - C_k^{1/2})(C_{k+1}^{1/2}(C_{k+1}^{1/2} + C_k^{1/2}) + C_k) \\
&\leq (C_{k+1}^{1/2} - C_k^{1/2})(C_{k+1}^{1/2}(C_{k+1}^{1/2} + C_k^{1/2}) + \tfrac{1}{2} C_{k+1}^{1/2}(C_{k+1}^{1/2} + C_k^{1/2})) \\
&= \tfrac{3}{2} C_{k+1}^{1/2}(C_{k+1} - C_k).
\end{aligned}$$

Thus, $C_k \geq \theta^{1/7}(1 + \frac{2}{3}(k-1))$, $k \geq 1$. Finally, we obtain

$$\begin{aligned}
A_k &\overset{(4.3.21)}{\geq} \theta (C_k^{3/2})^{7/3} \geq \theta \left(\theta^{1/7} \cdot \frac{2k+1}{3}\right)^{7/2} = \theta^{3/2} \left(\frac{2k+1}{3}\right)^{7/2} \\
&= \left(\frac{1}{2\gamma}\left(\frac{1-\sigma^2}{2M^2\|x_0-x^*\|^2}\right)^{1/3}\right)^{3/2} \left(\frac{2k+1}{3}\right)^{3.5} \\
&= \frac{1}{4}\left(\frac{1}{\gamma}\right)^{3/2} \frac{\sqrt{1-\sigma^2}}{M\|x_0-x^*\|} \left(\frac{2k+1}{3}\right)^{3.5}. \qquad \square
\end{aligned}$$

Now we are ready to justify the rate of convergence of method (4.3.15).

**Theorem 4.3.2** *Let us choose $\sigma \in (0, 1)$ and $\gamma \geq 1$. Suppose that the parameters $\rho_k$ in method (4.3.15) satisfy condition (4.3.18). If method (4.3.15) is applied with $M = \frac{1}{\sigma} M_f$, then for any $k \geq 1$ we have*

$$f(x_k) - f(x^*) \leq \frac{2\gamma^{3/2} M_f \|x_0 - x^*\|^3}{\sigma\sqrt{1-\sigma^2}} \left(\frac{3}{2k+1}\right)^{3.5}. \tag{4.3.22}$$

*Proof* Indeed, in view of inequality (4.3.17), we have

$$f(x_k) - f(x^*) \leq \frac{1}{2A_k}\|x_0 - x^*\|^2.$$

It remains to use the lower bound (4.3.19). □

The best value of $\sigma$ in the right-hand side of inequality (4.3.22) is $\sigma = \frac{1}{\sqrt{2}}$. In this case,

$$f(x_k) - f(x^*) \leq 4\gamma^{3/2} M_f \|x_0 - x^*\|^3 \left(\frac{3}{2k+1}\right)^{3.5}, \quad k \geq 1. \tag{4.3.23}$$

### 4.3.3 Complexity of the Search Procedure

In the previous section, we presented a conceptual second-order scheme (4.3.15), which reaches the best possible rate of convergence (4.3.8). In contrast to the Accelerated Cubic Newton Method (4.2.46), its estimating sequence $\{\psi_k\}$ starts from the squared Euclidean norm. Another difference consists in presenting the coefficient $\rho_k$ in the equation defining the scaling coefficient $a_{k+1}$ (see Step b)). In order to make this method function in accordance to its rate of convergence (4.3.22), we need to ensure that

$$\rho_k \approx r_M(y_k). \tag{4.3.24}$$

Note that the right-hand side of this equality is a continuous function of $\rho_k$. In this method, if $\rho_k = 0$, then $a_{k+1} = +\infty$ and $y_k = v_k$. In this case, the left-hand side of inequality (4.3.24) is smaller than its right-hand side. If $\rho_k \to \infty$, then $a_{k+1} \to 0$ and $y_k \to x_k$. Thus, there is always a root of equation (4.3.24).

However, the problem is that any search procedure in $\rho_k$ is very expensive. It needs to call the oracle many times. At present it is difficult to point out any favorable property of function $y_k = y_k(\rho_k)$ which could help.

At the same time, from the practical point of view, the gain from this acceleration of the rate of convergence is very small. Indeed, method (4.2.46) ensures $O(\frac{1}{\epsilon^{1/3}})$ complexity of finding an $\epsilon$-solution of problem (4.3.9). The number of iterations of method (4.3.15) is of the order $O(\frac{1}{\epsilon^{2/7}})$. Thus, the gain in the number of iterations of the "optimal" method is bounded by a factor proportional to $\left(\frac{1}{\epsilon}\right)^{\frac{1}{21}}$. For the values of $\epsilon$ used in practical applications, namely the range $10^{-4} \dots 10^{-12}$, this is just an absolute constant (since $\left(10^{12}\right)^{\frac{1}{21}} < 4$). Therefore, this factor, decreasing the total number of iterations, cannot compensate a significant increase in the analytical computational complexity of each iteration. That is the main reason why we drop the cumbersome analysis of the complexity of the corresponding search procedure in this book.

To conclude, from the practical point of view, method (4.2.46) is now the fastest second-order scheme. At the same time, the problem of finding the optimal second-order method with cheap iteration remains an open and challenging question in Optimization Theory.

## 4.4 The Modified Gauss–Newton Method

(Quadratic regularization; The modified Gauss–Newton process; Global rate of convergence; Comparative analysis; Implementation issues.)

### *4.4.1 Quadratic Regularization of the Gauss–Newton Iterate*

The problem of solving a system of nonlinear equations is one of the most fundamental problems in Numerical Analysis. The standard approach consists in replacing the initial problem

$$\text{Find } x \in \mathbb{E}: \ f_i(x) = 0, \quad i = 1, \dots, m, \tag{4.4.1}$$

by a minimization problem

$$\min_{x \in \mathbb{E}} \left[ f(x) \overset{\text{def}}{=} \phi(f_1(x), \dots, f_m(x)) \right], \tag{4.4.2}$$

where function $\phi(u)$ is non-negative and vanishes only at the origin. The most recommended choice for this *merit function* $\phi(u)$ is the standard squared Euclidean norm:

$$\phi(u) = \|u\|_{(2)}^2 \equiv \sum_{i=1}^{m} \left(u^{(i)}\right)^2, \tag{4.4.3}$$

where squaring the norm has the advantage of keeping the objective function in (4.4.2) smooth enough. Of course, the new problem (4.4.2), (4.4.3) can be solved by the standard second-order minimization schemes. However, it is possible to reduce the order of the required derivatives by applying the so-called *Gauss–Newton* approach. In this case, the search direction is defined as a solution of the following auxiliary problem:

$$\min_{h\in\mathbb{E}} \{\phi\,(f_1(x) + \langle\nabla f_1(x), h\rangle, \dots, f_m(x) + \langle\nabla f_m(x), h\rangle) : \; x+h \in D(x)\},$$

where $D(x)$ is a properly chosen neighborhood of the point $x$. Under some non-degeneracy assumptions, for this strategy it is possible to establish local quadratic convergence.

Despite its elegance, the above approach deserves some criticism. Indeed, the transformation of problem (4.4.1) into problem (4.4.2) is done in a quite straightforward way. For example, if the initial system of equations is linear, then such a transformation squares the condition number of the problem. Besides increasing numerical instability, for large problems this leads to *squaring* the number of iterations, which is necessary to get an $\epsilon$-solution of the original problem.

In this section, we consider another approach. At first glance, it looks very similar to the standard one: We replace our initial problem by a minimization problem (4.4.2). However, our merit function is *non-smooth*.

Before we start, let us recall some notation. For a linear operator $A : \mathbb{E}_1 \to \mathbb{E}_2$, its *adjoint operator* $A^* : \mathbb{E}_2^* \to \mathbb{E}_1^*$ is defined as follows:

$$\langle s, Ax\rangle = \langle A^* s, x\rangle, \quad \forall x \in \mathbb{E}_1, \; s \in \mathbb{E}_2^*.$$

For measuring distances in $\mathbb{E}_1$ and $\mathbb{E}_2$, we introduce the norms $\|\cdot\|_{\mathbb{E}_1}$ and $\|\cdot\|_{\mathbb{E}_2}$. In the dual spaces, the norms are defined in the standard way. For example,

$$\|s\|_{\mathbb{E}_1^*} = \max_{x\in\mathbb{E}_1}\{\langle s, x\rangle : \; \|x\|_{\mathbb{E}_1} \le 1\}, \quad s \in \mathbb{E}_1^*.$$

If no ambiguity occurs, we drop subindexes of the norms since they are always defined by the spaces containing the arguments. For example, $\|s\| \equiv \|s\|_{\mathbb{E}_1^*}$ for $s \in \mathbb{E}_1^*$.

For $A : \mathbb{E}_1 \to \mathbb{E}_2$, we define the *minimal singular value* as follows:

$$\sigma_{\min}(A) = \min_{x\in\mathbb{E}_1}\{\|Ax\| : \; \|x\| = 1\} \quad \Rightarrow \quad \|Ax\| \ge \sigma_{\min}(A)\|x\| \quad \forall x \in \mathbb{E}_1.$$

For invertible $A$, we have $\sigma_{\min}(A) = 1/\|A^{-1}\|$. Note that for two linear operators $A_1$ and $A_2$,

$$\sigma_{\min}(A_1A_2) \ge \sigma_{\min}(A_1)\cdot\sigma_{\min}(A_2).$$

If $\sigma_{\min}(A) > 0$, then we say that the operator $A$ possesses *primal non-degeneracy*. If $\sigma_{\min}(A^*) > 0$, then we say that $A$ possesses *dual non-degeneracy*.

Finally, for a non-linear function $F(\cdot) : \mathbb{E}_1 \to \mathbb{E}_2$ we denote by $F'(x)$ its *Jacobian*, which is a linear operator from $\mathbb{E}_1$ to $\mathbb{E}_2$:

$$F'(x)h = \lim_{\alpha \to 0} \tfrac{1}{\alpha}[F(x + \alpha h) - F(x)] \in \mathbb{E}_2, \quad h \in \mathbb{E}_1.$$

In the special case $f(\cdot) : \mathbb{E}_1 \to \mathbb{E}_2 \equiv \mathbb{R}$, we have $f'(x)h = \langle \nabla f(x), h \rangle$ for all $h \in \mathbb{E}_1$.

Consider a smooth non-linear function $F(\cdot) : \mathbb{E}_1 \to \mathbb{E}_2$. Our main problem of interest is to find an approximate solution to the following system of equations:

$$F(x) = 0, \quad x \in \mathbb{E}_1. \tag{4.4.4}$$

In order to measure the quality of such a solution, we introduce a (sharp) *merit* function $\phi(u)$, $u \in \mathbb{E}_2$, which satisfies the following conditions:

- It is convex, non-negative and vanishes only at the origin. (Hence, its level sets are bounded.)
- It is Lipschitz-continuous with unit Lipschitz constant:

$$|\phi(u) - \phi(v)| \le \|u - v\|, \quad \forall u, v \in \mathbb{E}_2. \tag{4.4.5}$$

- It has a sharp minimum at the origin:

$$\phi(u) \ge \gamma_\phi \|u\|, \quad \forall u \in \mathbb{E}_2, \tag{4.4.6}$$

  for a certain $\gamma_\phi \in (0, 1]$.

For example, we can take $\phi(u) = \|u\|_{\mathbb{E}_2}$. Then $\gamma_\phi = 1$.

We can use this merit function to transform the problem (4.4.4) into the following *unconstrained minimization* problem:

$$\min_{x \in \mathbb{E}_1} \{\, f(x) \equiv \phi(F(x)) \,\} \overset{\mathrm{def}}{=} f^*. \tag{4.4.7}$$

Clearly, the solution $x^*$ to the system (4.4.4) exists if and only if the optimal value $f^*$ of the problem (4.4.7) is equal to zero. The iterative scheme proposed below can be seen as a minimization method for problem (4.4.7), which employs a special structure of the objective function. Function $f$ can even be non-smooth. However, we will see that it is possible to decrease its value at any point $x \in \mathbb{E}_1$ excluding the stationary points of the problem (4.4.7).

Let us fix some $x \in \mathbb{E}_1$. Consider the following *local model* of our objective function:

$$\psi(x; y) = \phi\left(F(x) + F'(x)(y - x)\right), \quad y \in \mathbb{E}_1.$$

Note that $\psi(x; y)$ is convex in $y$. Therefore it looks natural to choose the next approximation of the solution to problem (4.4.7) from the set

$$\operatorname{Arg}\min_{y \in \mathbb{E}_1} \psi(x; y).$$

Such schemes are very well studied in the literature. For example, if choosing $\phi$ as in (4.4.3), we get the classical *Gauss–Newton method*. However, in what follows we see that a simple regularization of this approach leads to another scheme, for which we can speak about global efficiency of the process.

Let us introduce the following smoothness assumption. Denote by $\mathscr{F}$ a closed convex set in $\mathbb{E}_1$ with non-empty interior.

**Assumption 4.4.1** *The function $F(\cdot)$ is differentiable on the set $\mathscr{F}$ and its derivative is Lipschitz-continuous:*

$$\| F'(x) - F'(y) \| \leq L \| x - y \|, \quad \forall x, y \in \mathscr{F}, \tag{4.4.8}$$

*with some $L > 0$.*

A straightforward consequence of this assumption is as follows:

$$\| F(y) - F(x) - F'(x)(y - x) \| \leq \tfrac{1}{2} L \| y - x \|^2, \quad x, y \in \mathscr{F}. \tag{4.4.9}$$

We skip its proof since it is very similar to the proof of inequality (1.2.13). In the remaining part of this section, we always assume that Assumption 4.4.1 is satisfied.

**Lemma 4.4.1** *For any $x$ and $y$ from $\mathscr{F}$, we have*

$$| f(y) - \psi(x; y) | \leq \tfrac{1}{2} L \| y - x \|^2. \tag{4.4.10}$$

*Proof* Let $d(x, y) = F(y) - F(x) - F'(x)(y - x) \in \mathbb{E}_2$. By inequality (4.4.9),

$$\| d(x, y) \| \leq \frac{1}{2} L \| x - y \|^2.$$

Since both $x$ and $y$ belong to $\mathscr{F}$, we have

$$\begin{aligned} | f(y) - \psi(x; y) | &= | \phi(F(y)) - \phi(F(x) + F'(x)(y - x)) | \\ &\overset{(4.4.5)}{\leq} \| d(x, y) \| \leq \frac{1}{2} L \| y - x \|^2. \qquad \square \end{aligned}$$

Inequality (4.4.10) provides us with an *upper* approximation of function $f$:

$$f(y) \leq \psi(x; y) + \tfrac{1}{2} L \| y - x \|^2, \quad \forall x, y \in \mathscr{F}.$$

Let us use it for constructing a minimization scheme. Let $M$ be a positive parameter. For the problem (4.4.7), define a *modified Gauss–Newton iterate* from a point $x \in \mathscr{F}$ as follows:

$$\boxed{V_M(x) \in \operatorname{Arg}\min_{y \in \mathbb{E}_1} \left[ \psi(x; y) + \tfrac{1}{2} M \|y - x\|^2 \right],} \tag{4.4.11}$$

where "Arg" indicates that $V_M(x)$ is chosen from the set of global minima of the corresponding minimization problem.[3] Note that the auxiliary optimization problem in (4.4.11) is *convex* in $y$. We postpone a discussion on the complexity of finding the point $V_M(x)$ until Sect. 4.4.4.

Let us prove several auxiliary results. Define

$$r_M(x) = \|V_M(x) - x\|,$$

$$f_M(x) = \psi(x; V_M(x)) + \tfrac{1}{2} M r_M^2(x),$$

$$\delta_M(x) = f(x) - f_M(x).$$

For a fixed $x$, the value $f_M(x)$ is a *concave function* in $M$ since it can be represented as a minimum of functions linear in $M$ (see Theorem 3.1.8):

$$f_M(x) = \min_{y \in \mathbb{E}_1} \left[ \psi(x; y) + \tfrac{1}{2} M \|y - x\|^2 \right].$$

Consequently, the value $\frac{1}{2} r_M^2(x)$, which is equal to the derivative of $f_M(x)$ in $M$ (see Lemma 3.1.14), is a *decreasing* function of $M$.

**Lemma 4.4.2** *For any $x \in \mathbb{E}_1$ we have*

$$\delta_M(x) \geq \frac{1}{2} M r_M^2(x). \tag{4.4.12}$$

*Proof* Let us fix an arbitrary $x \in \mathbb{E}_1$. Let $\psi_0(y) = \frac{1}{2} M \|y - x\|^2$ and

$$\psi_1(y) = \psi(x; y) + \psi_0(y).$$

In view of Theorem 3.1.24, there exists $g_1 \in \partial_y \psi(x; V_M(x))$ and $g_2 \in \partial \psi_0(V_M(x))$ such that

$$\langle g_1 + g_2, y - V_M(x) \rangle \geq 0 \quad \forall y \in \mathbb{E}_1. \tag{4.4.13}$$

[3] Since we do not assume that the norm $\|x\|$, $x \in \mathbb{E}_1$, is strongly convex, this problem may have a non-trivial convex set of global solutions.

At the same time, in view of identity (3.1.39), we have $\langle g_2, V_M(x) - x\rangle = Mr_M^2(x)$. Hence,

$$\begin{aligned} f(x) &= \psi(x;x) \overset{(3.1.23)}{\geq} \psi(x, V_M(x)) + \langle g_1, x - V_M(x)\rangle \\ &\overset{(4.4.13)}{\geq} \psi(x, V_M(x)) + \langle g_2, V_M(x) - x\rangle \\ &= \psi(x, V_M(x)) + Mr_M^2(x) \;=\; f_M(x) + \tfrac{1}{2}Mr_M^2(x). \end{aligned}$$

This is exactly inequality (4.4.12). □

Let us compare $\delta_M(x)$ with another natural measure of local decrease of the model $\psi(x;\cdot)$. For $r > 0$ define

$$\Delta_r(x) = f(x) - \min_{y\in\mathbb{E}_1}\{\psi(x;y) : \; \|y - x\| \leq r\}.$$

**Lemma 4.4.3** *For any $x \in \mathbb{E}_1$ and $r > 0$ we have*

$$\delta_M(x) \geq Mr^2 \cdot \varkappa\left(\tfrac{1}{Mr^2}\Delta_r(x)\right), \tag{4.4.14}$$

*where*

$$\varkappa(t) = \begin{cases} t - \frac{1}{2}, \; t \geq 1, \\ \frac{1}{2}t^2, \;\; t \in [0,1]. \end{cases}$$

*The right-hand side of the bound (4.4.14) is a decreasing function of $M$.*

*Proof* Let us choose $h_r \in \operatorname{Arg}\min\limits_{h\in\mathbb{E}_1}\{\psi(x; x + h) : \; \|h\| \leq r\}$. Then

$$\begin{aligned} f_M(x) &\leq \min_{\tau}\{\phi(F(x) + \tau F'(x)h_r) + \tfrac{1}{2}M\tau^2r^2 : \; \tau \in [0,1]\} \\ &= \min_{\tau}\{\phi((1-\tau)F(x) + \tau(F(x) + F'(x)h_r)) + \tfrac{1}{2}M\tau^2r^2 : \; \tau \in [0,1]\} \\ &\leq \min_{\tau}\{(1-\tau)\phi(F(x)) + \tau\phi(F(x) + F'(x)h_r)) + \tfrac{1}{2}M\tau^2r^2 : \; \tau \in [0,1]\} \\ &= \min_{\tau}\{f(x) - \tau\Delta_r(x) + \tfrac{1}{2}M\tau^2r^2 : \; \tau \in [0,1]\}. \end{aligned}$$

Thus,

$$\delta_M(x) \geq \max_{\tau\in[0,1]}\{\tau\Delta_r(x) - \tfrac{1}{2}M\tau^2r^2\} = Mr^2 \cdot \varkappa\left(\tfrac{1}{Mr^2}\Delta_r(x)\right).$$

Note that the right-hand side of this inequality is decreasing in $M$. □

Define

$$\mathscr{L}(\tau) = \{y \in \mathbb{E}_1 : \ f(y) \le \tau\}.$$

**Lemma 4.4.4** *Let $\mathscr{L}(f(x)) \subseteq int\ \mathscr{F}$ and $M \ge L$. Then $V_M(x) \in \mathscr{L}(f(x))$.*

*Proof* Let $V_M(x) \notin \mathscr{L}(f(x))$. Consider the points

$$y(\alpha) = x + \alpha \cdot (V_M(x) - x), \quad \alpha \in [0, 1].$$

Since $y(0) = x \in \text{int } \mathscr{F}$, we can define the value $\bar{\alpha} \in (0, 1)$ such that $y(\bar{\alpha})$ lies at the boundary of the set $\mathscr{F}$. Note that

$$f(y(\bar{\alpha})) \ge f(x) \ge f_M(x),$$

and $r_M(x) > 0$. By our assumption, $\bar{\alpha} \in (0, 1)$. Define

$$d = F(y(\bar{\alpha})) - F(x) - \bar{\alpha} F'(x)(V_M(x) - x) \in \mathbb{E}_2.$$

In view of inequality (4.4.9), $\|d\| \le \frac{L}{2} \bar{\alpha}^2 r_M^2(x)$. Therefore,

$$\begin{aligned} f(x) \le f(y(\bar{\alpha})) &= \phi(F(x) + \bar{\alpha} F'(x)(y(1) - x) + d) \\ &\le \phi((F(x) + \bar{\alpha} F'(x)(V_M(x) - x)) + \|d\| \\ &\le (1 - \bar{\alpha}) f(x) + \bar{\alpha}\phi((F(x) + F'(x)(V_M(x) - x)) + \tfrac{1}{2} M \bar{\alpha}^2 r_M^2(x) \\ &\le (1 - \bar{\alpha}) f(x) + \bar{\alpha} f_M(x) - \tfrac{1}{2} M \bar{\alpha}(1 - \bar{\alpha}) r_M^2(x). \end{aligned}$$

Thus, $f(x) \le f_M(x) - \frac{1}{2} M (1 - \bar{\alpha}) r_M^2(x)$, which is a contradiction to (4.4.12). □

**Lemma 4.4.5** *Let both $x$ and $V_M(x)$ belong to $\mathscr{F}$. Then*

$$f_M(x) \le \min_{y \in \mathscr{F}} \left[ f(y) + \frac{1}{2}(L + M)\|y - x\|^2 \right]. \tag{4.4.15}$$

*Proof* For $y \in \mathscr{F}$ let $d(x, y) = F(y) - F(x) - F'(x)(y - x) \in \mathbb{E}_2$. By inequality (4.4.9),

$$\|d(x, y)\| \le \frac{1}{2} L \|x - y\|^2.$$

Hence, since both $x$ and $V_M(x)$ belong to $\mathscr{F}$, we have

$$\begin{aligned} f_M(x) &= \min_{y\in\mathscr{F}}\left[\phi(F(x)+F'(x)(y-x))+\frac{1}{2}M\|y-x\|^2\right] \\ &= \min_{y\in\mathscr{F}}\left[\phi(F(y)-d(x,y))+\frac{1}{2}M\|y-x\|^2\right] \\ &\leq \min_{y\in\mathscr{F}}\left[f(y)+\frac{1}{2}(L+M)\|y-x\|^2\right]. \qquad \square \end{aligned}$$

**Corollary 4.4.1** *Let $x^*$ be a solution to problem (4.4.7) and $\mathscr{L}(f(x)) \subseteq \mathscr{F}$. Then*

$$f_M(x) \leq f^* + \frac{1}{2}(L+M)\|x-x^*\|^2. \tag{4.4.16}$$

*Proof* It is enough to substitute $y = x^*$ in the right-hand side of (4.4.15). □

## 4.4.2 *The Modified Gauss–Newton Process*

Now we can analyze the convergence of the following process. Let us fix $L_0 \in (0, L]$.

```
Modified Gauss–Newton method

Initialization: Choose $x_0 \in \mathbb{R}^n$.

Iteration $k$, $(k \geq 0)$ :                                        (4.4.17)

    1. Find $M_k \in [L_0, 2L]$ such that

           $f(V_{M_k}(x_k)) \leq f_{M_k}(x_k)$.

    2. Set $x_{k+1} = V_{M_k}(x_k)$.
```

Since $f_M(x) \leq f(x)$, this process is monotone:

$$f(x_{k+1}) \leq f(x_k). \tag{4.4.18}$$

If the constant $L$ is known, then in Item 1 of this scheme we can use $M_k \equiv L$. In the opposite case, it is possible to apply a simple search procedure (see, for example, Sect. 4.1.4). Let us now present the convergence results.

Let $x_0 \in \operatorname{int} \mathscr{F}$ be a starting point for the above minimization process. We need to assume the following.

**Assumption 4.4.2** *The set $\mathscr{F}$ is big enough: $\mathscr{L}(f(x_0)) \subseteq \mathscr{F}$.*

In what follows, we always suppose that Assumption 4.4.2 is satisfied. In view of (4.4.18,) this assumption implies that $\mathscr{L}(f(x_k)) \subseteq \mathscr{F}$ for any $k \ge 0$.

**Theorem 4.4.1** *For any $k \ge 0$ and $r > 0$ we have*

$$\begin{aligned} f(x_k) - f^* &\ge \tfrac{1}{2} L_0 \sum_{i=k}^{\infty} r^2_{M_i}(x_i) \ge \tfrac{1}{2} L_0 \sum_{i=k}^{\infty} r^2_{2L}(x_i), \\ f(x_k) - f^* &\ge r^2 \sum_{i=k}^{\infty} M_i \varkappa\left(\tfrac{1}{M_i r^2} \Delta_r(x)\right) \ge 2Lr^2 \sum_{i=k}^{\infty} \varkappa\left(\tfrac{1}{2Lr^2} \Delta_r(x)\right). \end{aligned} \tag{4.4.19}$$

*Proof* Indeed, in view of the rules of Step 1 in (4.4.17),

$$f_{M_i}(x_i) \ge f(x_{i+1}), \quad M_i \ge L_0, \quad r_{M_i}(x_i) \ge r_{2L}(x_i).$$

Thus, inequality (4.4.12) justifies the first inequality in (4.4.19). In order to prove the second one, we apply (4.4.14) and use the bound $M_i \le 2L$ imposed by (4.4.17). □

**Corollary 4.4.2** *Let the sequence $\{x_k\}_{k=0}^{\infty}$ be generated by the scheme (4.4.17). Then*

$$\lim_{k\to\infty} \|x_k - x_{k+1}\| = 0, \quad \lim_{k\to\infty} \Delta_r(x_k) = 0,$$

*and therefore the set of limit points $X^*$ of this sequence is connected. For any $\bar{x}$ from $X^*$, we have $\Delta_r(\bar{x}) = 0$.* □

Let us justify now the local convergence of the scheme (4.4.17).

**Theorem 4.4.2** *Let the point $x^* \in \mathscr{L}(f(x_0))$ with $F(x^*) = 0$ be a non-degenerate solution to problem (4.4.4):*

$$\sigma \equiv \sigma_{\min}(F'(x^*)) > 0.$$

*Let $\gamma_\phi$ be defined by (4.4.6). If $x_k \in \mathscr{L}(f(x_0))$ and*

$$\|x_k - x^*\| \le \tfrac{2}{L} \cdot \tfrac{\sigma\gamma_\phi}{3+5\gamma_\phi},$$

*then* $x_{k+1} \in \mathscr{L}(f(x_0))$ *and*

$$\|x_{k+1} - x^*\| \le \frac{3(1+\gamma_\phi)L\,\|x_k - x^*\|^2}{2\gamma_\phi(\sigma - L\|x_k - x^*\|)} \le \|x_k - x^*\|. \tag{4.4.20}$$

*Proof* Since $f(x^*) = 0$, in view of inequality (4.4.16) and inequality (4.4.9), we have

$$\begin{aligned}
\tfrac{3L}{2}\|x_k - x^*\|^2 \ge f_{M_k}(x_k) \ge \psi(x_k; x_{k+1}) \;\ge\; & \gamma_\phi \|F(x_k) + F'(x_k)(x_{k+1} - x_k)\| \\
&= \gamma_\phi \|F'(x^*)(x_{k+1} - x^*) + \big(F(x_k) - F(x^*) - F'(x^*)(x_k - x^*)\big) \\
&\quad + (F'(x_k) - F'(x^*))(x_{k+1} - x_k)\| \\
&\ge \gamma_\phi[\|F'(x^*)(x_{k+1} - x^*)\| - \tfrac{L}{2}\|x_k - x^*\|^2 \\
&\quad - L\|x_k - x^*\| \cdot \|x_{k+1} - x_k\|] \\
&\ge \gamma_\phi \left[(\sigma - L\|x_k - x^*\|) \cdot \|x_{k+1} - x^*\| - \tfrac{3L}{2}\|x_k - x^*\|^2\right]. \qquad \Box
\end{aligned}$$

### 4.4.3 Global Rate of Convergence

In order to get global complexity results for method (4.4.17), we need to introduce an additional non-degeneracy assumption.

**Assumption 4.4.3** *The operator* $F'(x) : \mathbb{E}_1 \to \mathbb{E}_2$ *possesses a uniform* dual *non-degeneracy:*

$$\sigma_{\min}(F'(x)^*) \ge \sigma > 0 \quad \forall x \in \mathscr{L}(f(x_0)).$$

Note that this assumption implies $\dim \mathbb{E}_2 \le \dim \mathbb{E}_1$. The role of Assumption 4.4.3 in our analysis can be seen from the following standard result.

**Lemma 4.4.6** *Let the linear operator* $A : \mathbb{E}_1 \to \mathbb{E}_2$ *possess dual non-degeneracy:*

$$\sigma_{\min}(A^*) > 0.$$

*Then for any* $b \in \mathbb{E}_2$ *there exists a point* $x(b) \in \mathbb{E}_1$ *such that*

$$Ax(b) = b, \quad \|x(b)\| \le \frac{\|b\|}{\sigma_{\min}(A^*)}.$$

*Proof* Consider the following optimization problem:

$$\min_x \{ f(x) = \|x\| : \ Ax = b \}.$$

Since the level sets of its objective function are bounded, its solution $x^*$ exists. In view of the statement (3.1.59), there exists a $y^* \in \mathbb{E}_2^*$ such that $g^* = A^* y^* \in \partial f(x^*)$. Using inequality (3.1.42) and Lemma 3.1.15, we conclude that $\|g^*\| \leq 1$. Thus,

$$1 \geq \|A^* y^*\| \ \geq \ \sigma_{\min}(A^*) \|y^*\|. \tag{4.4.21}$$

On the other hand,

$$\|x^*\| \overset{(3.1.40)}{=} \langle g^*, x^* \rangle \ = \ \langle Ax^*, y^* \rangle \ = \ \langle b, y^* \rangle \ \leq \ \|b\| \cdot \|y^*\|.$$

It remains to apply inequality (4.4.21). □

An important consequence of Lemma 4.4.6 is as follows.

**Lemma 4.4.7** *Let the operator $F'(x)$ possess dual non-degeneracy: $\sigma_{\min}(F'(x)^*) > 0$. Then for any $M > 0$ we have*

$$r_M(x) \leq \tfrac{\|F(x)\|}{\sigma_{\min}(F'(x)^*)}. \tag{4.4.22}$$

*Proof* Indeed, in view of Lemma 4.4.6 there exists an $h^*$ such that

$$F(x) + F'(x) h^* = 0$$

and $\|h^*\| \leq \frac{\|F(x)\|}{\sigma_{\min}(F'(x)^*)}$. Therefore

$$\begin{aligned} \frac{M}{2} r_M^2(x) &\leq \psi(x; V_M(x)) + \frac{M}{2} r_M^2(x) = \min_{h \in \mathbb{E}_1} \left[ \psi(x; x+h) + \frac{M}{2} \|h\|^2 \right] \\ &\leq \frac{M}{2} \|h^*\|^2 \leq \frac{M \|F(x)\|^2}{2\sigma_{\min}^2(F'(x)^*)}. \qquad \Box \end{aligned}$$

Now we can justify the global rate of convergence of scheme (4.4.17).

**Theorem 4.4.3** *Let Assumptions 4.4.1, 4.4.2 and 4.4.3 be satisfied.*

1) *Suppose that the sequence $\{x_k\}_{k=0}^{\infty}$ is generated by method (4.4.17). If $f(x_k) \geq \frac{\sigma^2}{2L} \gamma_\phi^2$, then*

$$f(x_{k+1}) \leq f(x_k) - \tfrac{\sigma^2}{4L} \gamma_\phi^2. \tag{4.4.23}$$

*Otherwise,*

$$f(x_{k+1}) \leq \frac{L}{\sigma^2\gamma_\phi^2} f^2(x_k) \leq \tfrac{1}{2} f(x_k). \tag{4.4.24}$$

2) *Suppose that the sequence* $\{x_k\}_{k=0}^\infty$ *is generated by method (4.4.17) with* $M_k \equiv L$. *If* $f(x_k) \geq \frac{\sigma^2}{L}\gamma_\phi^2$, *then*

$$f(x_{k+1}) \leq f(x_k) - \frac{\sigma^2}{2L}\gamma_\phi^2. \tag{4.4.25}$$

*Otherwise,*

$$f(x_{k+1}) \leq \frac{L}{2\sigma^2\gamma_\phi^2} f^2(x_k) \leq \tfrac{1}{2} f(x_k). \tag{4.4.26}$$

*Proof* Let us prove the first part of the theorem. Since the operator $F'(x_k)$ is nondegenerate, in view of Lemma 4.4.6 there exists a solution $h_k^*$ to the system of linear equations $F(x_k) + F'(x_k)h = 0$ with a bounded norm:

$$\| h_k^* \| \leq \tfrac{1}{\sigma} \| F(x_k) \| \leq \tfrac{1}{\sigma\gamma_\phi} f(x_k).$$

Therefore, in view of the step-size rules in the scheme (4.4.17) and the upper bound on the values $M_k$, we have

$$\begin{aligned} f(x_{k+1}) &\leq \min_{h \in \mathbb{E}_1} \left[ \phi(F(x_k) + F'(x_k)h) + \tfrac{1}{2} M_k \|h\|^2 \right] \\ &\leq \min_{t\in[0,1]} \left[ \phi(F(x_k) + tF'(x_k)h_k^*) + L \| t h_k^* \|^2 \right] \\ &\leq \min_{t\in[0,1]} \left[ \phi((1-t)F(x_k)) + \tfrac{L}{\sigma^2\gamma_\phi^2} t^2 f^2(x_k) \right] \\ &\leq \min_{t\in[0,1]} \left[ (1-t) f(x_k) + \tfrac{L}{\sigma^2\gamma_\phi^2} t^2 f^2(x_k) \right]. \end{aligned}$$

Thus, if $f(x_k) \leq \frac{\sigma^2}{2L}\gamma_\phi^2$, then the minimum in the latter univariate problem is attained at $t = 1$ and we get inequalities (4.4.24). In the opposite case, the minimum is attained at $t = \frac{\sigma^2\gamma_\phi^2}{2Lf(x_k)}$ and we get estimate (4.4.23).

The second part of the theorem can be proved in a similar way. □

Using Theorem 4.4.3, we can establish some properties of problem (4.4.7).

**Theorem 4.4.4** *Let Assumptions 4.4.1, 4.4.2 and 4.4.3 be satisfied. Then there exists a solution $x^*$ to problem (4.4.7) such that $f(x^*) = 0$ and*

$$\| x^* - x_0 \| \leq \tfrac{2}{\sigma} \| F(x_0) \|. \tag{4.4.27}$$

*Proof* Let us choose $\phi(u) = \|u\|$. Then $\gamma_\phi = 1$. Let us now apply method (4.4.17) with $M_k \equiv L$ to the corresponding problem (4.4.7) with $f(x) = \|F(x)\|$.

Assume first that $f(x_0) > \frac{\sigma^2}{L}$. In accordance with the second statement of Theorem 4.4.3, as far as $f(x_k) \geq \frac{\sigma^2}{L}$ we have

$$f(x_k) - f(x_{k+1}) \geq \tfrac{\sigma^2}{2L}. \tag{4.4.28}$$

Denote by $N$ the length of the first stage of the process:

$$f(x_N) \geq \tfrac{\sigma^2}{L} \geq f(x_{N+1}).$$

Summing up inequalities (4.4.28) for $k = 0, \dots, N$, we get

$$N + 1 \leq \tfrac{2L}{\sigma^2}(f(x_0) - f(x_{N+1})). \tag{4.4.29}$$

On the other hand, in view of inequality (4.4.12) we have

$$f(x_k) - f(x_{k+1}) \geq \tfrac{L}{2} \| x_k - x_{k+1} \|^2. \tag{4.4.30}$$

Summing up these inequalities for $k = 0, \dots, N$, we get

$$\begin{aligned} f(x_0) - f(x_{N+1}) &\geq \tfrac{L}{2} \sum_{k=0}^{N} \| x_k - x_{k+1} \|^2 \geq \tfrac{L}{2(N+1)} \left( \sum_{k=0}^{N} \| x_k - x_{k+1} \| \right)^2 \\ &\geq \tfrac{L}{2(N+1)} \| x_0 - x_{N+1} \|^2. \end{aligned}$$

Now, using estimate (4.4.29), we obtain

$$\| x_0 - x_{N+1} \| \leq \left[ \tfrac{2(N+1)}{L}(f(x_0) - f(x_{N+1})) \right]^{1/2} \leq \tfrac{2}{\sigma}(f(x_0) - f(x_{N+1})). \tag{4.4.31}$$

Further, in view of Theorem 4.4.3, at the second stage of the process we can guarantee that

$$f(x_{k+1}) \leq \tfrac{L}{2\sigma^2} f^2(x_k) \leq \tfrac{1}{2} f(x_k), \quad k \geq N + 1. \tag{4.4.32}$$

Thus, $f(x_{N+k+1}) \leq (\frac{1}{2})^k f(x_{N+1})$ for $k \geq 0$. Hence, in view of inequality (4.4.22) we have

$$\|x_{N+k+2} - x_{N+k+1}\| \leq \tfrac{1}{\sigma}(\tfrac{1}{2})^k f(x_{N+1}), \quad k \geq 0.$$

Thus, the sequence $\{x_k\}_{k=0}^{\infty}$ converges to a point $x^*$ with $F(x^*) = 0$ and

$$\|x^* - x_{N+1}\| \leq \tfrac{2}{\sigma} f(x_{N+1}).$$

Taking into account this inequality and (4.4.31), we get (4.4.27).

If $f(x_0) \leq \frac{\sigma^2}{L}$, then we can apply the latter reasoning from the very beginning:

$$\sum_{k=0}^{\infty} \|x_{k+1} - x_k\| \leq \tfrac{1}{\sigma} \sum_{k=0}^{\infty} f(x_k) \leq \tfrac{1}{\sigma} f(x_0) \sum_{k=0}^{\infty} (\tfrac{1}{2})^k = \tfrac{2}{\sigma} f(x_0). \qquad \Box$$

Applying exactly the same arguments as in the proof of Theorem 4.4.4, it is possible to justify the following statement.

**Theorem 4.4.5** *Let Assumptions 4.4.1, 4.4.2 and 4.4.3 be satisfied. Suppose the sequence $\{x_k\}_{k=0}^{\infty}$ is generated by method (4.4.17) as applied to problem (4.4.7). Then this sequence converges to a single point $x^*$ with $F(x^*) = 0$.* □

Let us conclude this section with the following remark. We have seen that Assumptions 4.4.1, 4.4.2 and 4.4.3 guarantee the existence of a solution to problem (4.4.4). Define

$$D = \min_x \{\|x - x_0\| : \ x \in \mathscr{L}(f(x_0)), \ F(x) = 0\}.$$

In view of Corollary 4.4.1 and the bounds on $M_k$ in method (4.4.17), we can always guarantee that

$$f(x_1) \leq \tfrac{3}{2} L D^2. \tag{4.4.33}$$

Thus, in view of Theorem 4.4.3, the number of iterations $N$ of method (4.4.17) which is necessary for reaching the region of quadratic convergence can be bounded as follows:

$$N \leq 1 + \tfrac{4L}{\sigma^2 \gamma_\phi^2} f(x_1) \leq 1 + 6\left(\tfrac{LD}{\sigma \gamma_\phi}\right)^2. \tag{4.4.34}$$

We will refer to this bound as an upper complexity estimate of the class of problems described by Assumptions 4.4.1, 4.4.2 and 4.4.3. This bound is justified by the modified Gauss–Newton method (4.4.17).

## 4.4.4 Discussion

### 4.4.4.1 A Comparative Analysis of Scheme (4.4.17)

Let us compare the efficiency of method (4.4.17) with the Cubic Newton Method for unconstrained minimization (see Sect. 4.1). Note that the fields of applications of both methods intersect. Indeed, any problem of solving a system of non-linear equations can be transformed into a problem of unconstrained minimization using some merit function. On the other hand, any unconstrained minimization problem can be reduced to a system of non-linear equations, which corresponds to the first-order optimality conditions (1.2.4).

Consider the following unconstrained minimization problem:

$$\min_{x \in \mathbb{E}_1} \varphi(x), \tag{4.4.35}$$

where $\varphi(\cdot)$ is a twice differentiable strongly convex function whose Hessian is Lipschitz continuous. In this subsection, we assume that all norms are Euclidean. Suppose that there exist positive $\sigma$ and $L$ such that the conditions

$$\begin{aligned} \langle \nabla^2 \varphi(x) h, h \rangle &\geq \sigma \|h\|^2, \\ \|\nabla^2 \varphi(x+h) - \nabla^2 \varphi(x)\| &\leq L \|h\|, \end{aligned} \tag{4.4.36}$$

are satisfied for any $x$ and $h$ from $\mathbb{E}_1$. Let $D = \|x_0 - x^*\|$. Then in Sect. 4.1.5, we have shown that the complexity of problem (4.4.35) for the Cubic Newton Method (4.1.16) depends on the characteristic

$$\zeta = \tfrac{LD}{\sigma}$$

(we use the notation of this section). If $\zeta < 1$, then problem (4.4.35) is easy. In the opposite case, the number of iterations of the modified Newton scheme which is necessary to come to the region of quadratic convergence is essentially bounded by

$$N_1 = 6.25\sqrt{\zeta}, \tag{4.4.37}$$

(see (4.1.57)).

Note that problem (4.4.35) can be posed in the form (4.4.4):

$$\text{Find } x : \; F(x) \overset{\text{def}}{=} \nabla \varphi(x) = 0. \tag{4.4.38}$$

In this case, $F'(x) = \nabla^2 \varphi(x)$. Therefore, in view of conditions (4.4.36), our problem (4.4.38) satisfies Assumptions 4.4.1, 4.4.2 and 4.4.3. Let us choose $f(x) = \|F(x)\|$. Then, in view of (4.4.34), the number of iterations of the modified Gauss–Newton scheme (4.4.17) required to come to the region of quadratic convergence is bounded by

$$N_2 = 1 + 6\zeta^2. \tag{4.4.39}$$

Clearly, the estimate (4.4.37) is much better than (4.4.39). However, this observation just confirms a standard rule that the specialized procedures are usually more efficient than a general purpose scheme. However, at this moment we cannot come to a definitive answer since the lower complexity bounds for the problem class described by Assumptions 4.4.1, 4.4.2 and 4.4.3 are not known. So, there is a chance that the complexity (4.4.39) can be improved by other methods.

In fact, as compared with the Cubic Newton Method (4.1.16), the scheme (4.4.17) has one important advantage. The auxiliary problem for computing the new test point at each iteration of method (4.1.16) is solvable in polynomial time only if this method is based on the Euclidean norm. On the contrary, in the modified Gauss–Newton scheme we are absolutely free in the choice of norms in the spaces $\mathbb{E}_1$ and $\mathbb{E}_2$. As we will see in Sect. 4.4.4.2, any choice results in a convex auxiliary problem. Therefore, it is possible to choose the norms in a reasonable way, which makes the ratio $\frac{L}{\sigma}$ as small as possible.

#### 4.4.4.2 Implementation Issues

Let us study the complexity of auxiliary problem (4.4.11). For simplicity, let us assume that we choose $f(x) = \|F(x)\|$. So, our problem is as follows:

$$\text{Find } f_M(x) = \min_{h\in\mathbb{E}_1} \left[ \|F(x) + F'(x)h\| + \frac{1}{2}M\|h\|^2 \right]. \tag{4.4.40}$$

Note that sometimes this problem looks easier in its dual form:

$$\begin{aligned}
&\min_{h\in\mathbb{E}_1} \left[ \|F(x) + F'(x)h\| + \frac{1}{2}M\|h\|^2 \right] \\
&= \min_{h\in\mathbb{E}_1} \max_{\substack{s\in\mathbb{E}_2^* \\ \|s\|\le 1}} \left[ \langle s, F(x) + F'(x)h\rangle + \frac{1}{2}M\|h\|^2 \right] \\
&= \max_{\substack{s\in\mathbb{E}_2^* \\ \|s\|\le 1}} \min_{h\in\mathbb{E}_1} \left[ \langle s, F(x) + F'(x)h\rangle + \frac{1}{2}M\|h\|^2 \right] \\
&= \max_{s\in\mathbb{E}_2^*} \left[ \langle s, F(x)\rangle - \frac{1}{2M}\|F'(x)^* s\|_*^2 :\ \|s\| \le 1 \right].
\end{aligned}$$

Since this problem is convex, it can be solved by the efficient optimization schemes of Convex Optimization.

Let us show that for Euclidean norms, problem (4.4.40) can be solved by the standard Linear Algebra technique.

**Lemma 4.4.8** *Let us introduce in* $\mathbb{E}_1$ *and* $\mathbb{E}_2$ *the Euclidean norms:*

$$\|x\| = \langle B_1 x, x\rangle^{1/2}, \; x \in \mathbb{E}_1, \quad \|u\| = \langle B_2 u, u\rangle^{1/2}, \; u \in \mathbb{E}_2,$$

*where* $B_1 = B_1^* \succeq 0$*, and* $B_2 = B_2^* \succeq 0$*. Then the solution of the problem (4.4.40) can be found by the following univariate convex optimization problem:*

$$f_M(x) = \min_{\tau \geq 0} \left[ \tau + \tfrac{1}{\tau} \|F(x)\|^2 - \langle [\tau F'(x)^* B_2 F'(x) + \tau^2 M B_1]^{-1} g, g\rangle \right], \tag{4.4.41}$$

*where* $g = F'(x)^* B_2 F(x)$*. If* $\tau^*$ *is an optimal solution to this problem, then the solution to (4.4.40) is given by*

$$h^* = -\left[F'(x)^* B_2 F'(x) + \tau^* M B_1\right]^{-1} F'(x)^* B_2 F(x). \tag{4.4.42}$$

*Proof* Indeed

$$\begin{aligned} f_M(x) &= \min_{h \in \mathbb{E}_1} \min_{\tau \geq 0} \left[ \tfrac{1}{2}\tau + \tfrac{1}{2\tau} \|F(x) + F'(x)h\|^2 + \tfrac{M}{2}\|h\|^2 \right] \\ &= \min_{\tau \geq 0} \min_{h \in \mathbb{E}_1} \left[ \tfrac{1}{2}\tau + \tfrac{1}{2\tau} \|F(x) + F'(x)h\|^2 + \tfrac{M}{2}\|h\|^2 \right] \\ &= \min_{\tau \geq 0} \min_{h \in \mathbb{E}_1} \left[ \tfrac{1}{2}\tau + \tfrac{1}{2\tau} \|F(x)\|^2 + \tfrac{1}{\tau}\langle B_2 F(x), F'(x)h\rangle \right. \\ &\qquad \left. + \tfrac{1}{2\tau}\langle B_2 F'(x)h, F'(x)h\rangle + \tfrac{M}{2}\langle B_1 h, h\rangle \right]. \end{aligned}$$

The minimum of the internal minimization problem is achieved at

$$\begin{aligned} h^*(\tau) &= -\left[\tfrac{1}{\tau} F'(x)^* B_2 F'(x) + M B_1\right]^{-1} \tfrac{1}{\tau} F'(x)^* B_2 F(x) \\ &= -\left[F'(x)^* B_2 F'(x) + \tau M B_1\right]^{-1} F'(x)^* B_2 F(x). \end{aligned}$$

With the notation $g = F'(x)^* B_2 F(x)$, the objective function of the optimization problem in $\tau$ is as follows:

$$\tfrac{1}{2}\tau + \tfrac{1}{2\tau}\|F(x)\|^2 - \tfrac{1}{2\tau^2}\langle \Big[\tfrac{1}{\tau}F'(x)^* B_2 F'(x) + MB_1\Big]^{-1} g, g\rangle$$

$$= \tfrac{1}{2}\tau + \tfrac{1}{2\tau}\|F(x)\|^2 - \tfrac{1}{2}\langle \big[\tau F'(x)^* B_2 F'(x) + \tau^2 MB_1\big]^{-1} g, g\rangle.$$

In view of Theorem 3.1.7, this function is convex in $\tau$. □

Note that the univariate optimization problem in (4.4.41) can be solved efficiently by one-dimensional search procedures (see, for example, Sect. A.1).

# Part II
# Structural Optimization

# Chapter 5
# Polynomial-Time Interior-Point Methods

In this section, we present the problem classes and complexity bounds of polynomial-time interior-point methods. These methods are based on the notion of a self-concordant function. It appears that such a function can be easily minimized by the Newton's Method. On the other hand, an important subclass of these functions, the self-concordant barriers, can be used in the framework of path-following schemes. Moreover, it can be proved that we can follow the corresponding central path with polynomial-time complexity. The size of the steps in the penalty coefficient of the central path depends on the corresponding barrier parameter. It appears that for any convex set there exists a self-concordant barrier with parameter proportional to the dimension of the space of variables. On the other hand, for any convex set with explicit structure, such a barrier with a reasonable value of parameter can be constructed by simple combination rules. We present applications of this technique to Linear and Quadratic Optimization, Linear Matrix Inequalities and other optimization problems.

## 5.1 Self-concordant Functions

(Do we really have a Black Box? What does the Newton method actually do? Definition of self-concordant functions; Main properties; The Implicit Function Theorem; Minimizing self-concordant functions; Relations with the standard second-order methods.)

### *5.1.1 The Black Box Concept in Convex Optimization*

In this chapter, we are going to present the main ideas underlying the modern polynomial-time interior-point methods in Nonlinear Optimization. In order to start, let us look first at the traditional formulation of a minimization problem.

Y. Nesterov, *Lectures on Convex Optimization*, Springer Optimization
and Its Applications 137, https://doi.org/10.1007/978-3-319-91578-4_5

Suppose we want to solve a minimization problem in the following form:

$$\min_{x \in \mathbb{R}^n} \{ f_0(x) : \; f_j(x) \leq 0, \; j = 1 \ldots m \}.$$

We assume that the functional components of this problem are convex. Note that all standard convex optimization schemes for solving this problem are based on the Black-Box concept. This means that we assume our problem to be equipped with an oracle, which provides us with some information on the functional components of the problem at some test point $x$. This oracle is local: If we change the shape of the component far enough from the test point, the answer of the oracle does not change. These answers comprise the only information available for numerical methods.[1]

However, looking carefully at the above situation, we can discover a certain contradiction. Indeed, in order to apply the convex optimization methods, we need to be *sure* that our functional components are convex. However, we can check convexity only by analyzing the *structure* of these functions[2]: If our function is obtained from the *basic* convex functions by *convex* operations (summation, maximum, etc.), we conclude that it is convex.

Thus, the functional components of the problem are not in the Black Box at the moment we are checking their convexity and choose the minimization scheme. However, we lock them in the Black Box for numerical methods. This is the main conceptual contradiction of the standard Convex Optimization theory.[3]

The above observation gives us hope that the structure of the problem could be used to improve performance of convex minimization schemes. Unfortunately, structure is a very fuzzy notion, which is quite difficult to formalize. One possible way to describe the structure is to fix the *analytical type* of functional components. For example, we can consider the problems with linear functions $f_j(\cdot)$ only. This works, but note that this approach is very fragile: If we introduce in our problem just a single functional component of different type, we get another problem class and all the theory must be redone from scratch.

Alternatively, it is clear that having the structure at hand, we can play with the *analytical form* of the problem. We can rewrite the problem in many equivalent forms using nontrivial transformations of variables or constraints, introducing additional variables, etc. However, this would serve no purpose without realizing the final goal of such transformations. So, let us try to find such a goal.

At this moment, it is better to look at classical examples. In many situations, the sequential reformulations of the initial problem can be seen as a part of the numerical method. We start from a complicated problem $\mathscr{P}$ and, step by step, simplify its structure up to the moment we get a trivial problem (or, a problem

[1] We have already discussed this concept and the corresponding methods in Part I of the book.

[2] A numerical verification of convexity is a hopeless computational task.

[3] Nevertheless, the conclusions of the theory concerning the oracle-based minimization schemes remain valid, of course, for the methods which are *designed* in accordance with the Black-Box principles.

which we know how to solve):

$$\mathscr{P} \longrightarrow \ldots \longrightarrow (f^*, x^*).$$

Let us look at the standard approach for solving the system of linear equations, namely,

$$Ax = b.$$

We can proceed as follows:

1. Check that matrix $A$ is symmetric and positive definite. Sometimes this is clear from its origin.
2. Compute the Cholesky factorization of the matrix:

   $$A = LL^T,$$

   where $L$ is a lower-triangular matrix. Form two auxiliary systems

   $$Ly = b, \quad L^T x = y.$$

3. Solve the auxiliary systems.

This process can be seen as a sequence of equivalent transformations of the initial problem.

Imagine for a moment that we do not know how to solve the systems of linear equation. In order to discover the above technology, we should perform the following steps:

1. Find a class of problems which can be solved very efficiently (linear systems with triangular matrices in our example).
2. Describe the transformation rules for converting our initial problem into the desired form.
3. Describe the class of problems for which these transformation rules are applicable.

We are ready to explain how it works in Convex Optimization. First of all, we need to find a *basic* numerical scheme and problem formulation at which this scheme is very efficient. We will see that for our goals the most appropriate candidate is the *Newton's method* (see Sect. 1.2.4 and Chap. 4) as applied in the framework of *Sequential Unconstrained Minimization* (see Sect. 1.3.3).

In the next section, we will analyze some drawbacks of the standard theory on the Newton's method. From this analysis, we derive a family of very special convex functions, so-called *self-concordant functions* and *self-concordant barriers*, which can be efficiently minimized by the Newton's method. We use these objects in the description of a transformed version of the initial problem. In the sequel, we refer to this description as a *barrier model* of our problem. This model will replace the standard functional model of the optimization problem used in the previous chapters.

## 5.1.2 *What Does the Newton's Method Actually Do?*

Let us look at the standard result on the local convergence of Newton's method (we have proved it as Theorem 1.2.5). We need to find an unconstrained local minimum $x^*$ of the twice differentiable function $f(\cdot)$:

$$\min_{x \in \mathbb{R}^n} f(x), \tag{5.1.1}$$

For the moment, all the norms we use are standard Euclidean. Assume that:

- $\nabla^2 f(x^*) \succeq \mu I_n$ with some constant $\mu > 0$,
- $\| \nabla^2 f(x) - \nabla^2 f(y) \| \leq M \| x - y \|$ for all $x$ and $y \in \mathbb{R}^n$.

Assume also that the starting point of the Newton process $x_0$ is close enough to $x^*$:

$$\| x_0 - x^* \| < \bar{r} = \tfrac{2\mu}{3M}. \tag{5.1.2}$$

Then we can prove (see Theorem 1.2.5) that the sequence

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k), \quad k \geq 0, \tag{5.1.3}$$

is well defined. Moreover, $\| x_k - x^* \| < \bar{r}$ for all $k \geq 0$ and the Newton's method (5.1.3) converges quadratically:

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

What is wrong with this result? Note that the description of the *region of quadratic convergence* (5.1.2) for this method is given in terms of the *standard* inner product

$$\langle x, y \rangle = \sum_{i=1}^{n} x^{(i)} y^{(i)}, \quad x, y \in \mathbb{R}^n.$$

If we choose a new basis in $\mathbb{R}^n$, then all objects in our description change: the metric, the Hessians, the bounds $\mu$ and $M$. However, let us see what happens in this situation with the Newton process. Namely, let $B$ be a nondegenerate $(n \times n)$-matrix. Consider the function

$$\phi(y) = f(By), \quad y \in \mathbb{R}^n.$$

The following result is very important for understanding the nature of the Newton's method.

**Lemma 5.1.1** *Let the sequence $\{x_k\}$ be generated by the Newton's method as applied to the function $f$:*

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1}\nabla f(x_k), \quad k \geq 0.$$

*Consider the sequence $\{y_k\}$, generated by the Newton's method for the function $\phi$:*

$$y_{k+1} = y_k - [\nabla^2 \phi(y_k)]^{-1}\nabla \phi(y_k), \quad k \geq 0,$$

*with $y_0 = B^{-1}x_0$. Then $y_k = B^{-1}x_k$ for all $k \geq 0$.*

*Proof* Let $y_k = B^{-1}x_k$ for some $k \geq 0$. Then

$$\begin{aligned} y_{k+1} &= y_k - [\nabla^2\phi(y_k)]^{-1}\nabla\phi(y_k) = y_k - [B^T\nabla^2 f(By_k)B]^{-1}B^T\nabla f(By_k) \\ &= B^{-1}x_k - B^{-1}[\nabla^2 f(x_k)]^{-1}\nabla f(x_k) = B^{-1}x_{k+1}. \end{aligned}$$ □

Thus, the Newton's method is *affine invariant* with respect to affine transformations of variables. Therefore, its actual region of quadratic convergence *does not depend* on a particular choice of the basis. It depends only on the local topological structure of the function $f(\cdot)$.

Let us try to understand what was wrong in our assumptions. The main assumption is related to the Lipschitz continuity of the Hessians:

$$\| \nabla^2 f(x) - \nabla^2 f(y) \| \leq M \| x - y \|, \quad \forall x, y \in \mathbb{R}^n.$$

Let us assume that $f \in C^3(\mathbb{R}^n)$. Define

$$f'''(x)[u] = \lim_{\alpha\to 0} \frac{1}{\alpha}[\nabla^2 f(x+\alpha u) - \nabla^2 f(x)] \;\equiv\; D^3 f(x)[h].$$

The object in the right-hand side of this equality (and, consequently, in its left-hand side) is an $(n \times n)$-matrix. Thus, our assumption is equivalent to the condition

$$\| f'''(x)[u] \| \leq M \| u \| .$$

This means that at any point $x \in \mathbb{R}^n$, we have

$$\langle f'''(x)[u]v, v\rangle \;\equiv\; D^3 f(x)[u, v, v] \leq M \| u \| \cdot \| v \|^2 \quad \forall u, v \in \mathbb{R}^n.$$

Note that the value in the left-hand side of this inequality is invariant with respect to affine transformations of variables (since this is just a third directional derivative along direction $u$ and twice along direction $v$). However, its right-hand side does depend on the choice of coordinates. Therefore, the most natural way to improve our situation consists in finding an affine-invariant replacement for the standard Euclidean norm $\| \cdot \|$. The most natural candidate for such a replacement is quite evident: This is the norm defined by the Hessian $\nabla^2 f(x)$ itself, namely,

$$\| u \|^2_{\nabla^2 f(x)} = \langle \nabla^2 f(x) u, u \rangle \; \equiv \; D^2 f(x)[h, h].$$

This choice results in the definition of a *self-concordant function*.

### 5.1.3 Definition of Self-concordant Functions

Since we are going to work with affine-invariant objects, it is natural to get rid of coordinate representations and denote by $\mathbb{E}$ a real vector space for our variables, and by $\mathbb{E}^*$ the dual space (see Sect. 4.2.1).

Let us consider a *closed convex* function $f(\cdot) \in C^3(\mathrm{dom}\, f)$ with *open* domain. By fixing a point $x \in \mathrm{dom}\, f$ and direction $u \in \mathbb{E}$, we define a function

$$\phi(x; t) = f(x + tu),$$

dependent on the variable $t \in \mathrm{dom}\, \phi(x; \cdot) \subseteq \mathbb{R}$. Define

$$Df(x)[u] = \phi'(x; 0) = \langle \nabla f(x), u \rangle,$$

$$D^2 f(x)[u, u] = \phi''(x; 0) = \langle \nabla^2 f(x) u, u \rangle = \| u \|^2_{\nabla^2 f(x)},$$

$$D^3 f(x)[u, u, u] = \phi'''(x; 0) = \langle D^3 f(x)[u] u, u \rangle.$$

**Definition 5.1.1** A function $f$ is called *self-concordant* if there exists a constant $M_f \geq 0$ such that the inequality

$$|D^3 f(x)[u, u, u]| \leq 2 M_f \| u \|^3_{\nabla^2 f(x)} \tag{5.1.4}$$

holds for all $x \in \mathrm{dom}\, f$ and $u \in \mathbb{E}$. If $M_f = 1$, the function is called *standard self-concordant*.

Note that we are going to use these functions to construct a barrier model of our problem. Our main hope is that they can be easily minimized by the Newton's method.

Let us point out an equivalent definition of self-concordant functions.

**Lemma 5.1.2** *A function $f$ is self-concordant if and only if for any $x \in dom\, f$ and any triple of directions $u_1$, $u_2$, $u_3 \in \mathbb{E}$ we have*

$$| D^3 f(x)[u_1, u_2, u_3] | \leq 2M_f \prod_{i=1}^{3} \| u_i \|_{\nabla^2 f(x)} . \tag{5.1.5}$$

We accept this statement without proof since it needs some special facts from the theory of tri-linear symmetric forms. For the same reason, we accept without proof the following corollary.

**Corollary 5.1.1** *A function $f$ is self-concordant if and only if for any $x \in dom\, f$ and any direction $u \in \mathbb{R}^n$ we have*

$$D^3 f(x)[u] \preceq 2M_f \|u\|_{\nabla^2 f(x)} \nabla^2 f(x). \tag{5.1.6}$$

In what follows, we often use Definition 5.1.1 in order to prove that some $f$ is self-concordant. In contrast, Lemma 5.1.2 is useful for establishing different properties of self-concordant functions.

Let us consider several examples.

*Example 5.1.1*

1. *Linear function*. Consider the function

$$f(x) = \alpha + \langle a, x \rangle, \quad \text{dom}\, f = \mathbb{E}.$$

Then

$$\nabla f(x) = a, \quad \nabla^2 f(x) = 0, \quad \nabla^3 f(x) = 0,$$

and we conclude that $M_f = 0$.

2. *Convex quadratic function*. Consider the function

$$f(x) = \alpha + \langle a, x \rangle + \frac{1}{2}\langle Ax, x \rangle, \quad \text{dom}\, f = \mathbb{E},$$

where $A = A^* \succeq 0$. Then

$$\nabla f(x) = a + Ax, \quad \nabla^2 f(x) = A, \quad \nabla^3 f(x) = 0,$$

and we conclude that $M_f = 0$.

3. *Logarithmic barrier for a ray*. Consider a univariate function

$$f(x) = -\ln x, \quad \text{dom}\, f = \{x \in \mathbb{R} \mid \ x > 0\}.$$

Then

$$f'(x) = -\tfrac{1}{x}, \quad f''(x) = \tfrac{1}{x^2}, \quad f'''(x) = -\tfrac{2}{x^3}.$$

Therefore, $f(\cdot)$ is self-concordant with $M_f = 1$.

4. *Logarithmic barrier for an ellipsoid.* Let $A = A^* \succeq 0$. Consider the *concave* function

$$\phi(x) = \alpha + \langle a, x\rangle - \frac{1}{2}\langle Ax, x\rangle.$$

Define $f(x) = -\ln \phi(x)$, with $\mathrm{dom}\, f = \{x \in \mathbb{E} : \ \phi(x) > 0\}$. Then

$$\begin{aligned} Df(x)[u] &= -\tfrac{1}{\phi(x)}[\langle a, u\rangle - \langle Ax, u\rangle], \\ D^2 f(x)[u,u] &= \tfrac{1}{\phi^2(x)}[\langle a, u\rangle - \langle Ax, u\rangle]^2 + \tfrac{1}{\phi(x)}\langle Au, u\rangle, \\ D^3 f(x)[u,u,u] &= -\tfrac{2}{\phi^3(x)}[\langle a, u\rangle - \langle Ax, u\rangle]^3 \\ &\quad -\tfrac{3}{\phi^2(x)}[\langle a, u\rangle - \langle Ax, u\rangle]\langle Au, u\rangle. \end{aligned}$$

Let $\omega_1 = Df(x)[u]$ and $\omega_2 = \frac{1}{\phi(x)}\langle Au, u\rangle$. Then

$$D^2 f(x)[u,u] = \omega_1^2 + \omega_2 \geq 0,$$

$$\mid D^3 f(x)[u,u,u] \mid = \mid 2\omega_1^3 + 3\omega_1\omega_2 \mid.$$

The only nontrivial case is $\omega_1 \neq 0$. Let $\xi = \omega_2/\omega_1^2$. Then

$$\tfrac{|D^3 f(x)[u,u,u]|}{(D^2 f(x)[u,u])^{3/2}} \leq \tfrac{2|\omega_1|^3 + 3|\omega_1|\omega_2}{(\omega_1^2+\omega_2)^{3/2}} = \tfrac{2(1+\frac{3}{2}\xi)}{(1+\xi)^{3/2}} \leq 2,$$

where the last inequality follows from the convexity of the function $(1+\xi)^{3/2}$ for $\xi \geq -1$. Thus, the function $f$ is self-concordant and $M_f = 1$.

5. It is easy to verify that none of the following univariate functions is self-concordant:

$$f(x) = e^x, \quad f(x) = \tfrac{1}{x^p}, \ x > 0, \ p > 0, \quad f(x) = \mid x \mid^p, \ p > 2.$$

However the function $f_p(x) = \frac{1}{2}x^2 + \frac{1}{px^p} - \frac{1}{p}$ with $p > 0$ is self-concordant for $x > 0$. Let us prove this. Indeed,

$$f_p'(x) = x - \tfrac{1}{x^{p+1}}, \quad f_p''(x) = 1 + \tfrac{p+1}{x^{p+2}} \geq 1, \quad f_p'''(x) = -\tfrac{(p+1)(p+2)}{x^{p+3}}.$$

If $x \geq 1$, then

$$|f_p'''(x)| = \tfrac{(p+1)(p+2)}{x^{p+2}} \;\leq\; (p+2)f_p''(x) \;\leq\; (p+2)[f_p''(x)]^{3/2}.$$

If $x \in (0, 1]$, then

$$\begin{aligned} |f_p'''(x)| = \tfrac{(p+1)(p+2)}{x^{p+3}} \;&\leq\; (p+1)(p+2)\left(\tfrac{1}{x^{p+2}}\right)^{3/2} \\ &\leq (p+1)(p+2)\left(\tfrac{f_p''(x)}{p+1}\right)^{3/2}. \end{aligned}$$

Thus, we can take $M_{f_p} = \max\left\{1 + \frac{p}{2}, \frac{p+2}{2\sqrt{p+1}}\right\} = 1 + \frac{p}{2}$. Note that the function $f_p$ is well defined as $p \to 0$. Indeed,

$$\lim_{p\to 0} f_p(x) = \tfrac{1}{2}x^2 + \lim_{p\to 0} \tfrac{1}{p}\left[e^{p \ln \frac{1}{x}} - 1\right] \;=\; \tfrac{1}{2}x^2 - \ln x.$$

6. Let $f \in C^{3,2}_{L_3}(\mathbb{R}^n)$. Assume that it is strongly convex on $\mathbb{R}^n$ with convexity parameter $\sigma_2(f)$. Then, for any $x \in \mathbb{R}^n$ and direction $u \in \mathbb{R}^n$ we have

$$D^3 f(x)[u] \preceq L_3\|u\| I_n \overset{(2.1.28)}{\preceq} L_3\left(\tfrac{1}{\sigma_2(f)}\|u\|^2_{\nabla^2 f(x)}\right)^{1/2} \tfrac{1}{\sigma_2(f)}\nabla^2 f(x).$$

Thus, in view of Corollary 5.1.1, we can take $M_f = \frac{L_3}{2\sigma_2^{3/2}(f)}$. □

Let us now look at the main properties of self-concordant functions.

**Theorem 5.1.1** *Let functions $f_i$ be self-concordant with constants $M_i$, $i = 1, 2$, and let $\alpha$, $\beta > 0$. Then the function $f(x) = \alpha f_1(x) + \beta f_2(x)$ is self-concordant with constant*

$$M_f = \max\left\{\tfrac{1}{\sqrt{\alpha}}M_1, \tfrac{1}{\sqrt{\beta}}M_2\right\}$$

*and dom $f$ = dom $f_1 \bigcap$ dom $f_2$.*

*Proof* In view of Theorem 3.1.5, $f$ is a closed convex function. Let us fix some $x \in \operatorname{dom} f$ and $u \in \mathbb{E}$. Then

$$|\; D^3 f_i(x)[u, u, u] \;| \leq 2M_i\left[D^2 f_i(x)[u, u]\right]^{3/2}, \quad i = 1, 2.$$

Let $\omega_i = D^2 f_i(x)[u, u] \geq 0$. Then

$$\tfrac{|D^3 f(x)[u,u,u]|}{[D^2 f(x)[u,u]]^{3/2}} \leq \tfrac{\alpha|D^3 f_1(x)[u,u,u]|+\beta|D^3 f_2(x)[u,u,u]|}{[\alpha D^2 f_1(x)[u,u]+\beta D^2 f_2(x)[u,u]]^{3/2}} \;\leq\; \tfrac{\alpha M_1\omega_1^{3/2}+\beta M_2\omega_2^{3/2}}{[\alpha\omega_1+\beta\omega_2]^{3/2}}. \tag{5.1.7}$$

The right-hand side of this inequality does not change when we replace $(\omega_1, \omega_2)$ by $(t\omega_1, t\omega_2)$ with $t > 0$. Therefore, we can assume that

$$\alpha\omega_1 + \beta\omega_2 = 1.$$

Let $\xi = \alpha\omega_1$. Then the right-hand side of inequality (5.1.7) becomes equal to

$$\tfrac{M_1}{\sqrt{\alpha}}\xi^{3/2} + \tfrac{M_2}{\sqrt{\beta}}(1-\xi)^{3/2}, \quad \xi \in [0, 1].$$

This function is convex in $\xi$. Therefore it attains its maximum at the end points of the interval (see Corollary 3.1.1). □

**Corollary 5.1.2** *Let a function $f$ be self-concordant with some constant $M_f$. If $A = A^* \succeq 0$, then the function*

$$\phi(x) = \alpha + \langle a, x\rangle + \frac{1}{2}\langle Ax, x\rangle + f(x)$$

*is also self-concordant with constant $M_\phi = M_f$.*

*Proof* We have seen that any convex quadratic function is self-concordant with zero constant. □

**Corollary 5.1.3** *Let a function $f$ be self-concordant with some constant $M_f$ and $\alpha > 0$. Then the function $\phi(x) = \alpha f(x)$ is also self-concordant with constant $M_\phi = \frac{1}{\sqrt{\alpha}}M_f$.* □

Let us now prove that self-concordance is an affine-invariant property.

**Theorem 5.1.2** *Let $\mathscr{A}(x) = Ax + b$: $\mathbb{E} \to \mathbb{E}_1$ be a linear operator. Assume that a function $f(\cdot)$ is self-concordant with constant $M_f$. Then the function*

$$\phi(x) = f(\mathscr{A}(x))$$

*is also self-concordant and $M_\phi = M_f$.*

*Proof* The function $\phi(\cdot)$ is closed and convex in view of Theorem 3.1.6. Let us fix some $x \in \operatorname{dom}\phi = \{x : \mathscr{A}(x) \in \operatorname{dom} f\}$ and $u \in \mathbb{E}$. Define $y = \mathscr{A}(x)$, $v = Au$. Then

$$D\phi(x)[u] = \langle \nabla f(\mathscr{A}(x)), Au\rangle = \langle \nabla f(y), v\rangle,$$

$$D^2\phi(x)[u, u] = \langle \nabla^2 f(\mathscr{A}(x))Au, Au\rangle = \langle \nabla^2 f(y)v, v\rangle,$$

$$D^3\phi(x)[u, u, u] = D^3 f(\mathscr{A}(x))[Au, Au, Au] = D^3 f(y)[v, v, v].$$

Therefore,

$$\begin{aligned} \mid D^3\phi(x)[u,u,u] \mid &= \mid D^3 f(y)[v,v,v] \mid \leq 2M_f \langle \nabla^2 f(y)v, v\rangle^{3/2} \\ &= 2M_f (D^2\phi(x)[u,u])^{3/2}. \qquad \square \end{aligned}$$

Finally, let us describe the behavior of a self-concordant function near the boundary of its domain.

**Theorem 5.1.3** *Let $f$ be a self-concordant function. Then for any $\bar{x} \in \partial(dom\, f)$ and any sequence*

$$\{x_k\} \subset dom\, f : \quad x_k \to \bar{x}$$

*we have $f(x_k) \to +\infty$.*

*Proof* Since $f$ is a closed convex function with open domain, this statement follows from Item 2 of Theorem 3.1.4. □

Thus, $f$ is a *barrier function* for $\mathrm{cl}\,(\mathrm{dom}\, f)$ (see Sect. 1.3.3). Finally, let us establish the self-concordance of a logarithmic barrier for the level set of self-concordant function.

**Theorem 5.1.4** *Let a function $f$ be self-concordant with constant $M_f$ and $f(x) \geq f^*$ for all $x \in dom\, f$. For arbitrary $\beta > f^*$, consider the function*

$$\phi(x) = -\ln(\beta - f(x)).$$

*Then*

1. *$\phi$ is well defined on $dom\,\phi = \{x \in dom\, f : \ f(x) < \beta\}$.*
2. *For any $x \in dom\,\phi$ and $h \in \mathbb{E}$ we have*

$$\langle \nabla^2\phi(x)h, h\rangle \geq \langle \nabla\phi(x), h\rangle^2. \tag{5.1.8}$$

3. *$\phi$ is self-concordant with constant $M_\phi = \sqrt{1 + M_f^2(\beta - f^*)}$.*

*Proof* Let us fix $x \in \mathrm{dom}\,\phi$ and $h \in \mathbb{E}$. Consider the function $\psi(\tau) = \phi(x + \tau h)$. Define $\omega = \beta - f(x)$. Then

$$\psi'(0) = \tfrac{1}{\omega}\langle \nabla f(x), h\rangle, \quad \psi''(0) = \tfrac{1}{\omega}\langle \nabla^2 f(x)h, h\rangle + \tfrac{1}{\omega^2}\langle \nabla f(x), h\rangle^2,$$

$$\psi'''(0) = \tfrac{1}{\omega} D^3 f(x)[h,h,h] + \tfrac{3}{\omega^2}\langle \nabla^2 f(x)h, h\rangle\langle \nabla f(x), h\rangle + \tfrac{2}{\omega^3}\langle \nabla f(x), h\rangle^3.$$

Thus, $\psi''(0) \geq (\psi'(0))^2$, and this is inequality (5.1.8).

Further, we need to bound $\psi'''(0)$ from above by $\psi''(0)^{3/2}$. Since $f$ is self-concordant, we have

$$\psi'''(0) \overset{(5.1.4)}{\leq} \tfrac{2M_f}{\omega} \langle \nabla^2 f(x)h, h\rangle^{3/2} + \tfrac{3}{\omega^2}\langle \nabla^2 f(x)h, h\rangle \langle \nabla f(x), h\rangle$$

$$+\tfrac{2}{\omega^3}\langle \nabla f(x), h\rangle^3.$$

The right-hand side of this inequality is homogeneous in $h$ of degree three. Therefore, let us find an upper bound for it assuming that $\psi''(0) = 1$. Defining

$$\tau = \left(\tfrac{1}{\omega}\langle \nabla^2 f(x)h, h\rangle\right)^{1/2}, \quad \xi = \tfrac{1}{\omega}\langle \nabla f(x), h\rangle,$$

we come to the following maximization problem:

$$\max_{\tau,\xi\in\mathbb{R}} \left\{2\hat{\omega}^{1/2}\tau^3 + 3\tau^2\xi + 2\xi^3 : \ \tau^2 + \xi^2 = 1\right\},$$

where $\hat{\omega} = M_f^2\omega$. Note that the optimal values of $\tau$ and $\xi$ in this problem are non-negative. Therefore, in view of the equality constraint, we can rewrite the objective function as follows.

$$2\hat{\omega}^{1/2}\tau^3 + 3\tau^2\xi + 2\xi^3 = 2\hat{\omega}^{1/2}\tau^3 + \tau^2\xi + 2\xi(\tau^2 + \xi^2) \ = \ 2\hat{\omega}^{1/2}\tau^3 + (\tau^2 + 2)\xi$$

$$= 2\hat{\omega}^{1/2}\tau^3 + (\tau^2 + 2)\sqrt{1 - \tau^2}.$$

The first-order optimality condition for this univariate function can be written as follows:

$$0 = 6\hat{\omega}^{1/2}\tau^2 + 2\tau\sqrt{1-\tau^2} - (\tau^2 + 2)\tfrac{\tau}{\sqrt{1-\tau^2}} \ = \ 6\hat{\omega}^{1/2}\tau^2 - \tfrac{3\tau^3}{\sqrt{1-\tau^2}}.$$

Thus, the optimal value $\tau_*$ satisfies equation $2\hat{\omega}^{1/2} = \frac{\tau_*}{\sqrt{1-\tau_*^2}}$. Hence, $\tau_* = \sqrt{\frac{4\hat{\omega}}{1+4\hat{\omega}}}$. Substituting this value into the objective function, we come to the following bound:

$$2\hat{\omega}^{1/2}\left(\tfrac{4\hat{\omega}}{1+4\hat{\omega}}\right)^{3/2} + \tfrac{2+12\hat{\omega}}{(1+4\hat{\omega})^{3/2}} \ = \ \tfrac{2+12\hat{\omega}+16\omega^2}{(1+4\hat{\omega})^{3/2}} \ = \ 2\tfrac{1+2\hat{\omega}}{(1+4\hat{\omega})^{1/2}} \ \leq \ 2\sqrt{1+\hat{\omega}}.$$

It remains to note that $\hat{\omega} \leq M_f^2(\beta - f^*)$. □

### 5.1.4 Main Inequalities

Let $f$ be a self-concordant function. Define

$$\| h \|_x = \langle \nabla^2 f(x)h, h\rangle^{1/2}.$$

We call $\| h \|_x$ the *(primal) local norm* of direction $h$ with respect to $x$. Let us fix a point $x \in \mathrm{dom}\, f$ and a direction $h \in \mathbb{E}$ such that $\langle \nabla^2 f(x)h, h\rangle > 0$. Consider the univariate function

$$\phi(t) = \tfrac{1}{\langle \nabla^2 f(x+th)h,h\rangle^{1/2}}.$$

In view of the continuity of the second derivative of the function $f$, $0 \in \mathrm{int}\,(\mathrm{dom}\, \phi)$.

**Lemma 5.1.3** *For all feasible $t$, we have $| \phi'(t) | \leq M_f$.*

*Proof* Indeed,

$$\phi'(t) = -\tfrac{D^3 f(x+th)[h,h,h]}{2\langle \nabla^2 f(x+tu)h,h\rangle^{3/2}}.$$

Therefore $| \phi'(t) | \leq M_f$ in view of Definition 5.1.1. □

**Corollary 5.1.4** *The domain of function $\phi(\cdot)$ contains the interval*

$$I_x = \left( -\tfrac{1}{M_f}\phi(0), \tfrac{1}{M_f}\phi(0) \right).$$

*Proof* Indeed, in view of Lemma 5.1.3, the values $\langle \nabla^2 f(x+\tau h)h, h\rangle$ are positive at any subinterval of $I_x$ and $\phi(t) \geq \phi(0) - M_f \mid t \mid$. Moreover, since $f(x+th) \to \infty$ as the points $x+th$ approach the boundary of $\mathrm{dom}\, f$ (see Theorem 5.1.3), the cannot intersect the boundary as $t \in I_x$. □

Let us consider the following ellipsoids:

$$W^0(x;r) = \{y \in \mathbb{E} \mid \| y - x \|_x < r\},$$

$$W(x;r) = \mathrm{cl}\left(W^0(x;r)\right) = \{y \in \mathbb{E} \mid \| y - x \|_x \leq r\}.$$

This set is called the *Dikin ellipsoid* of the function $f$ at $x$.

**Theorem 5.1.5** *1. For any $x \in dom\, f$, we have $W^0(x; \frac{1}{M_f}) \subseteq dom\, f$.*
*2. For all $x$, $y \in dom\, f$, the following inequality holds:*

$$\| y - x \|_y \geq \tfrac{\|y-x\|_x}{1+M_f\|y-x\|_x}. \tag{5.1.9}$$

*3. If* $\| y - x \|_x < \frac{1}{M_f}$, *then*

$$\| y - x \|_y \leq \frac{\|y-x\|_x}{1 - M_f \|y-x\|_x}. \tag{5.1.10}$$

*Proof* 1. Let us choose in $\mathbb{E}$ a Euclidean norm $\| \cdot \|$ and small $\epsilon > 0$. Consider the function $f_\epsilon(x) = f(x) + \frac{1}{2}\epsilon\|x\|^2$. In view of Corollary 5.1.2, it is self-concordant with constant $M_f$. Moreover, for any $h \in \mathbb{E}$ we have $\langle \nabla^2 f_\epsilon(x)h, h\rangle > 0$. Therefore, in view of Corollary 5.1.4, dom $f_\epsilon \equiv$ dom $f$ contains the set

$$\left\{ y = x + th \mid t^2(\| h \|_x^2 + \epsilon\|h\|^2) < \frac{1}{M_f^2} \right\}$$

(since $\phi(0) = 1/\langle \nabla^2 f_\epsilon(x)h, h\rangle^{1/2}$). Since $\epsilon$ can be arbitrarily small, this means that dom $f$ contains $W^0(x; \frac{1}{M_f})$.

2. Let us choose $h = y - x$. Assume for a moment that $\|h\|_x > 0$. Then

$$\phi(1) = \frac{1}{\|y-x\|_y}, \quad \phi(0) = \frac{1}{\|y-x\|_x},$$

and $\phi(1) \leq \phi(0) + M_f$ in view of Lemma 5.1.3. This is inequality (5.1.9).

3. If $\| y - x \|_x < \frac{1}{M_f}$, then $\phi(0) > M_f$, and in view of Lemma 5.1.3 $\phi(1) \geq \phi(0) - M_f$. This is inequality (5.1.10).

In the case when $\|h\|_x = 0$, both items can be justified by the trick used in the proof of Item 1. □

The next statement demonstrates that some local properties of self-concordant functions reflect somehow the global properties of its domain.

**Theorem 5.1.6** *Let a function $f$ be self-concordant and dom $f$ contains no straight lines. Then the Hessian $\nabla^2 f(x)$ is nondegenerate at all points $x \in$ dom $f$.*

*Proof* Assume that $\langle \nabla^2 f(\bar{x})h, h\rangle = 0$ for some $\bar{x} \in$ dom $f$ and direction $h \in \mathbb{E}$, $h \neq 0$. Then, all points of the line $\{x = \bar{x} + \tau h, \ \tau \in \mathbb{R}\}$ belong to the ellipsoid $W^0(x; \frac{1}{M_f})$. However, in view of Item 1 of Theorem 5.1.5, this ellipsoid belongs to dom $f$. This contradicts the conditions of the theorem. □

**Theorem 5.1.7** *Let $x \in$ dom $f$. Then for any $y \in W^0(x; \frac{1}{M_f})$ we have*

$$(1 - M_f r)^2 \, \nabla^2 f(x) \preceq \nabla^2 f(y) \ \preceq \ \frac{1}{(1-M_f r)^2} \nabla^2 f(x), \tag{5.1.11}$$

*where $r = \| y - x \|_x$.*

*Proof* Let us fix an arbitrary direction $h \in \mathbb{E}$, $h \neq 0$. Consider the function

$$\psi(t) = \langle \nabla^2 f(x + t(y - x))h, h\rangle, \quad t \in [0, 1].$$

Define $y_t = x + t(y - x)$ and $r = \|y - x\|_x$. Then, in view of Lemma 5.1.2 and inequality (5.1.10), we have

$$\begin{aligned} \mid \psi'(t) \mid &= \mid D^3 f(y_t)[y - x, h, h] \mid \leq \; 2M_f \; \| \; y - x \; \|_{y_t} \| \; h \; \|_{y_t}^2 \\ &= \tfrac{2M_f}{t} \; \| \; y_t - x \; \|_{y_t} \; \psi(t) \;\; \leq \;\; \tfrac{2M_f}{t} \cdot \tfrac{\| y_t - x \|_x}{1 - M_f \| y_t - x \|_x} \cdot \psi(t) \\ &= \tfrac{2M_f r}{1 - t M_f r} \psi(t). \end{aligned}$$

If $\|y - x\|_x = 0$, then $\psi(t) = \psi(0)$, $t \in [0, 1]$, and therefore

$$(1 - M_f r)^2 \psi(0) \leq \psi(t) \;\leq\; \tfrac{1}{(1 - M_f r)^2} \psi(0). \qquad (5.1.12)$$

If $r > 0$, then $2(\ln(1 - t M_f r))' \leq (\ln \psi(t))' \leq -2(\ln(1 - t M_f r))'$ for all $t \in [0, 1]$. Integrating these inequalities in $t \in [0, 1]$, we get again (5.1.12), which is equivalent to (5.1.11) since $h$ was chosen arbitrarily. □

**Corollary 5.1.5** *Let $x \in \mathrm{dom}\, f$ and $r = \| \; y - x \; \|_x < \frac{1}{M_f}$. Then we can bound the operator*

$$G = \int\limits_0^1 \nabla^2 f(x + \tau(y - x)) d\tau$$

*as follows:*

$$\left(1 - M_f r + \tfrac{1}{3} M_f^2 r^2\right) \; \nabla^2 f(x) \preceq G \preceq \tfrac{1}{1 - M_f r} \nabla^2 f(x).$$

*Proof* Indeed, in view of Theorem 5.1.7 we have

$$\begin{aligned} G &= \int\limits_0^1 \nabla^2 f(x + \tau(y - x)) d\tau \succeq \nabla^2 f(x) \cdot \int\limits_0^1 (1 - \tau M_f r)^2 d\tau \\ &= (1 - M_f r + \tfrac{1}{3} M_f^2 r^2) \nabla^2 f(x), \end{aligned}$$

and $G \;\preceq\; \nabla^2 f(x) \cdot \int\limits_0^1 \frac{d\tau}{(1 - \tau M_f r)^2} \;=\; \frac{1}{1 - M_f r} \nabla^2 f(x)$. □

*Remark 5.1.1* The statement of Corollary 5.1.5 remains valid for $r = \|y - x\|_y$.

Let us now recall the most important facts we have already proved.

- At any point $x \in \operatorname{dom} f$ we can define an *ellipsoid*

$$W^0\left(x; \tfrac{1}{M_f}\right) = \left\{x \in \mathbb{E} \mid \langle \nabla^2 f(x)(y-x), y-x\rangle\rangle < \tfrac{1}{M_f^2}\right\},$$

  belonging to dom $f$.
- Inside the ellipsoid $W(x; r)$ with $r \in [0, \frac{1}{M_f})$ the function $f$ is almost quadratic:

$$(1 - M_f r)^2 \nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \tfrac{1}{(1-M_f r)^2}\nabla^2 f(x)$$

  for all $y \in W(x; r)$. Choosing $r$ small enough, we can make the quality of quadratic approximation acceptable for our goals.

These two facts form the basis for all consequent results.

Let us now prove several inequalities related to the divergence of the value of a self-concordant function with respect to its linear approximation.

**Theorem 5.1.8** *For any $x$, $y \in \operatorname{dom} f$, we have*

$$\langle \nabla f(y) - \nabla f(x), y - x\rangle \geq \tfrac{\|y-x\|_x^2}{1+M_f\|y-x\|_x}, \tag{5.1.13}$$

$$f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle + \tfrac{1}{M_f^2}\omega(M_f \parallel y-x \parallel_x), \tag{5.1.14}$$

*where $\omega(t) = t - \ln(1+t)$.*

*Proof* Let $y_\tau = x+\tau(y-x)$, $\tau \in [0, 1]$, and $r = \parallel y-x \parallel_x$. Then, in view of (5.1.9) we have

$$\begin{aligned}\langle \nabla f(y) - \nabla f(x), y-x\rangle &= \int_0^1 \langle \nabla^2 f(y_\tau)(y-x), y-x\rangle d\tau \\ &= \int_0^1 \tfrac{1}{\tau^2} \parallel y_\tau - x \parallel_{y_\tau}^2 d\tau \\ &\geq \int_0^1 \tfrac{r^2}{(1+\tau M_f r)^2} d\tau = \tfrac{r}{M_f}\int_0^{M_f r} \tfrac{1}{(1+t)^2} dt = \tfrac{r^2}{1+M_f r}.\end{aligned}$$

Further, using (5.1.13), we obtain

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \int_0^1 \langle \nabla f(y_\tau) - \nabla f(x), y - x \rangle d\tau$$

$$= \int_0^1 \tfrac{1}{\tau} \langle \nabla f(y_\tau) - \nabla f(x), y_\tau - x \rangle d\tau \geq \int_0^1 \tfrac{\| y_\tau - x \|_x^2}{\tau(1 + M_f \| y_\tau - x \|_x)} d\tau = \int_0^1 \tfrac{\tau r^2}{1 + \tau M_f r} d\tau$$

$$= \tfrac{1}{M_f^2} \int_0^{M_f r} \tfrac{t dt}{1+t} = \tfrac{1}{M_f^2} \omega(M_f r).$$

□

**Theorem 5.1.9** *Let $x \in dom\, f$ and $\| y - x \|_x < \frac{1}{M_f}$. Then*

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \leq \tfrac{\| y - x \|_x^2}{1 - M_f \| y - x \|_x}, \tag{5.1.15}$$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{M_f^2} \omega_*(M_f \| y - x \|_x), \tag{5.1.16}$$

*where $\omega_*(t) = -t - \ln(1 - t)$.*

*Proof* Let $y_\tau = x + \tau(y - x)$, $\tau \in [0, 1]$, and $r = \| y - x \|_x$. Since $\| y_\tau - x \| < \frac{1}{M_f}$, in view of (5.1.10) we have

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle = \int_0^1 \langle \nabla^2 f(y_\tau)(y - x), y - x \rangle d\tau$$

$$= \int_0^1 \tfrac{1}{\tau^2} \| y_\tau - x \|_{y_\tau}^2 \, d\tau$$

$$\leq \int_0^1 \tfrac{r^2}{(1 - \tau M_f r)^2} d\tau = \tfrac{r}{M_f} \int_0^{M_f r} \tfrac{1}{(1-t)^2} dt = \tfrac{r^2}{1 - M_f r}.$$

Further, using (5.1.15), we obtain

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \int_0^1 \langle \nabla f(y_\tau) - \nabla f(x), y - x \rangle d\tau$$

$$= \int_0^1 \tfrac{1}{\tau} \langle \nabla f(y_\tau) - \nabla f(x), y_\tau - x \rangle d\tau \leq \int_0^1 \tfrac{\| y_\tau - x \|_x^2}{\tau(1 - M_f \| y_\tau - x \|_x)} d\tau = \int_0^1 \tfrac{\tau r^2}{1 - \tau M_f r} d\tau$$

$$= \tfrac{1}{M_f^2} \int_0^{M_f r} \tfrac{t dt}{1-t} = \tfrac{1}{M_f^2} \omega_*(M_f r).$$

□

**Theorem 5.1.10** *Inequalities (5.1.9), (5.1.10), (5.1.13), (5.1.14), (5.1.15) and (5.1.16) are necessary and sufficient characteristics of self-concordant functions.*

*Proof* We have already justified two sequences of implications:

$$\text{Definition 5.1.1} \Rightarrow (5.1.9) \Rightarrow (5.1.13) \Rightarrow (5.1.14),$$

$$\text{Definition 5.1.1} \Rightarrow (5.1.10) \Rightarrow (5.1.15) \Rightarrow (5.1.16).$$

Let us prove the implication (5.1.14) ⇒ Definition 5.1.1. Let $x \in \operatorname{dom} f$ and $x - \alpha u \in \operatorname{dom} f$ for $\alpha \in [0, \epsilon)$. Consider the function

$$\psi(\alpha) = f(x - \alpha u), \quad \alpha \in [0, \epsilon).$$

Let $r = \|u\|_x \equiv [\psi''(0)]^{1/2}$. Assuming that (5.1.14) holds for all $x$ and $y$ from $\operatorname{dom} f$, we have

$$\psi(\alpha) - \psi(0) - \psi'(0)\alpha - \tfrac{1}{2}\psi''(0)\alpha^2 \geq \tfrac{1}{M_f^2}\omega(\alpha M_f r) - \tfrac{1}{2}\alpha^2 r^2.$$

Therefore

$$\begin{aligned} \tfrac{1}{6}\psi'''(0) &= \lim_{\alpha \downarrow 0} \tfrac{1}{\alpha^3}\left[\psi(\alpha) - \psi(0) - \psi'(0)\alpha - \tfrac{1}{2}\psi''(0)\alpha^2\right] \\ &\geq \lim_{\alpha \downarrow 0} \tfrac{1}{\alpha^3}\left[\tfrac{1}{M_f^2}\omega(\alpha M_f r) - \tfrac{1}{2}\alpha^2 r^2\right] = \lim_{\alpha \downarrow 0} \tfrac{r}{3\alpha^2}\left[\tfrac{1}{M_f}\omega'(\alpha M_f r) - \alpha r\right] \\ &= \lim_{\alpha \downarrow 0} \tfrac{r}{3\alpha^2}\left[\tfrac{\alpha r}{1+\alpha M_f r} - \alpha r\right] = -\tfrac{1}{3}M_f r^3. \end{aligned}$$

Therefore, $D^3 f(x)[u, u, u] = -\psi'''(0) \leq 2M_f[\psi''(0)]^{3/2}$ and this is Definition 5.1.1. Implication (5.1.16) ⇒ Definition 5.1.1 can be proved in a similar way. □

Sometimes Theorem 5.1.10 is convenient for establishing self-concordance of certain functions. Let us demonstrate this with an *Implicit Function Theorem.*

Let us assume that $\mathbb{E} = \mathbb{E}_1 \times \mathbb{E}_2$. Thus, we have a corresponding partition of variable $z = (x, y) \in \mathbb{E}$. Let $\Phi$ be a self-concordant function with $\operatorname{dom} \Phi \subseteq \mathbb{E}$. Consider the following implicit function:

$$f(x) = \min_y \{\Phi(x, y) : (x, y) \in \operatorname{dom} \Phi\}. \tag{5.1.17}$$

In order to simplify the situation, let us assume that for any $x$ such that the set $Q(x) = \{y : (x, y) \in \mathrm{dom}\, \Phi\}$ is nonempty, it does not contain a straight line. Then simple conditions, like boundedness of $\Phi$ from below, guarantee existence of the unique solution $y(x)$ of the optimization problem in (5.1.17) (see Sect. 5.1.5).

Anyway, let us assume existence of point $y(x)$. Then it is characterized by the first-order optimality condition:

$$\nabla_y \Phi(x, y(x)) = 0. \tag{5.1.18}$$

Moreover, by Theorem 3.1.25 and Lemma 3.1.10, we have

$$\nabla f(x) = \nabla_x \Phi(x, y(x)). \tag{5.1.19}$$

Let us compute the Hessian of the function $f$. Differentiating equation (5.1.18) along direction $h \in \mathbb{E}_1$, we get

$$\nabla^2_{yx} \Phi(x, y(x))h + \nabla^2_{yy} \Phi(x, y(x)) y'(x)h = 0.$$

Therefore, by differentiating equality (5.1.19) along direction $h$, we obtain

$$\begin{aligned} \nabla^2 f(x)h &= \nabla^2_{xx} \Phi(x, y(x))h + \nabla^2_{xy} \Phi(x, y(x)) y'(x)h \\ &= \nabla^2_{xx} \Phi(x, y(x))h - \nabla^2_{xy} \Phi(x, y(x))[\nabla^2_{yy} \Phi(x, y(x))]^{-1} \nabla^2_{yx} \Phi(x, y(x))h. \end{aligned} \tag{5.1.20}$$

**Theorem 5.1.11** *Let $\Phi$ be a self-concordant function. Then the function $f$ defined by (5.1.17) is also self-concordant with constant $M_\Phi$.*

*Proof* Let us fix $\bar{x} \in \mathrm{dom}\, f$. Define $\bar{z} = (\bar{x}, y(\bar{x}))$ and let $x \in \mathrm{dom}\, f$. Then with $z = (x, y)$, we have

$$\begin{aligned} f(x) &= \min_{y \in Q(x)} \Phi(x, y) \\ &\overset{(5.1.14)}{\geq} \min_{y \in Q(x)} \left\{ \Phi(\bar{x}, y(\bar{x})) + \langle \nabla \Phi(\bar{x}, y(\bar{x})), z - \bar{z} \rangle + \tfrac{1}{M_f^2} \omega(M_f \| z - \bar{z} \|_{\bar{z}}) \right\} \\ &\overset{(5.1.19)}{=} f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle_{E_1} + \tfrac{1}{M_f^2} \omega\left( M_f \min_{y \in Q(x)} \| z - \bar{z} \|_{\bar{z}} \right). \end{aligned}$$

It remains to compute the minimum in the last line. Let $h = x - \bar{x}$. Then

$$\begin{aligned}
\min_{y \in Q(x)} \|z - \bar{z}\|_{\bar{z}}^2 &= \langle \nabla_{xx}^2 \Phi(\bar{z}) h, h \rangle_{\mathbb{E}_1} \\
&+ \min_{y \in Q(x)} \left\{ 2\langle \nabla_{xy}^2 \Phi(\bar{z})(y - \bar{y}), h \rangle_{\mathbb{E}_1} + \langle \nabla_{yy}^2 \Phi(\bar{z})(y - \bar{y}), y - \bar{y} \rangle_{\mathbb{E}_2} \right\} \\
&\geq \langle \nabla_{xx}^2 \Phi(\bar{z}) h, h \rangle_{\mathbb{E}_1} + \min_{\delta \in \mathbb{E}_2} \left\{ 2\langle \nabla_{xy}^2 \Phi(\bar{z})\delta, h \rangle_{\mathbb{E}_1} + \langle \nabla_{yy}^2 \Phi(\bar{z})\delta, \delta \rangle_{\mathbb{E}_2} \right\} \\
&= \langle \nabla_{xx}^2 \Phi(\bar{z}) h, h \rangle_{\mathbb{E}_1} - \langle [\nabla_{yy}^2 \Phi(\bar{z})]^{-1} \nabla_{yx}^2 \Phi(\bar{z}) h, \nabla_{yx}^2 \Phi(\bar{z}) h \rangle_{\mathbb{E}_1} \\
&\overset{(5.1.20)}{=} \langle \nabla^2 f(\bar{x}) h, h \rangle.
\end{aligned}$$

It remains to apply Theorem 5.1.10. □

Let us prove two more inequalities. From now on, we assume that dom $f$ contains no straight lines. In this case, in view of Theorem 5.1.6, all Hessians $\nabla^2 f(x)$ with $x \in \text{dom } f$ are nondegenerate. Denote by

$$\| g \|_x^* = \langle g, [\nabla^2 f(x)]^{-1} g \rangle^{1/2}, \quad g \in \mathbb{E}^*,$$

the *dual local norm*. Clearly, $| \langle g, h \rangle | \leq \| g \|_x^* \cdot \| h \|_x$.

**Theorem 5.1.12** *For any $x$ and $y$ from dom $f$ we have*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{M_f^2} \omega(M_f \|\nabla f(y) - \nabla f(x)\|_y^*). \tag{5.1.21}$$

*If in addition $\|\nabla f(y) - \nabla f(x)\|_y^* < \frac{1}{M_f}$, then*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{M_f^2} \omega_*(M_f \|\nabla f(y) - \nabla f(x)\|_y^*). \tag{5.1.22}$$

*Proof* Let us fix arbitrary points $x$ and $y$ from dom $f$. Consider the function

$$\phi(z) = f(z) - \langle \nabla f(x), z \rangle, \quad z \in \text{dom } f.$$

Note that this function is self-concordant and $\nabla \phi(x) = 0$. Therefore, using inequality (5.1.16), we get

$$\begin{aligned}
f(x) - \langle \nabla f(x), x \rangle &= \phi(x) = \min_{z \in \text{dom } f} \phi(z) \\
&\leq \min_z \left\{ \phi(y) + \langle \nabla \phi(y), z - y \rangle + \tfrac{1}{M_f^2} \omega_*(M_f \|z - y\|_y) : \|z - y\|_y < \tfrac{1}{M_f} \right\}
\end{aligned}$$

$$= \min_{0\le\tau<1}\left\{\phi(y) - \tfrac{\tau}{M_f}\|\nabla\phi(y)\|_y^* + \tfrac{1}{M_f^2}\omega_*(\tau)\right\} = \phi(y) - \tfrac{1}{M_f^2}\omega(M_f\|\nabla\phi(y)\|_y^*)$$

$$= f(y) - \langle\nabla f(x), y\rangle - \tfrac{1}{M_f^2}\omega(M_f\|\nabla f(y) - \nabla f(x)\|_y^*),$$

and this is inequality (5.1.21). In order to prove inequality (5.1.22), we use a similar reasoning based on inequality (5.1.14). □

All theorems above are written in terms of two auxiliary univariate functions,

$$\omega(t) = t - \ln(1+t), \quad \omega_*(\tau) = -\tau - \ln(1-\tau).$$

Note that

$$\omega'(t) = \tfrac{t}{1+t} \ge 0, \quad \omega''(t) = \tfrac{1}{(1+t)^2} > 0,$$

$$\omega_*'(\tau) = \tfrac{\tau}{1-\tau} \ge 0, \quad \omega_*''(\tau) = \tfrac{1}{(1-\tau)^2} > 0.$$

Therefore, $\omega(\cdot)$ and $\omega_*(\cdot)$ are convex functions. In what follows, we often use different relations between these objects. Let us provide them with a formal justification.

**Lemma 5.1.4** *For any $t \ge 0$ and $\tau \in [0,1)$, we have*

$$\omega'(\omega_*'(\tau)) = \tau, \quad \omega_*'(\omega'(t)) = t,$$

$$\omega(t) = \max_{0\le\xi<1}[\xi t - \omega_*(\xi)], \quad \omega_*(\tau) = \max_{\xi\ge 0}[\xi\tau - \omega(\xi)],$$

$$\omega(t) + \omega_*(\tau) \ge \tau t,$$

$$\omega_*(\tau) = \tau\omega_*'(\tau) - \omega(\omega_*'(\tau)), \quad \omega(t) = t\omega'(t) - \omega_*(\omega'(t)).$$

We leave the proof of this lemma as an exercise for the reader. Note that the main reason for the above relations is that functions $\omega(t)$ and $\omega_*(t)$ are *Fenchel conjugate* (see definition (3.1.27)).

Functions $\omega(\cdot)$ and $\omega_*(\cdot)$ will often be used for estimating the rate of growth of self-concordant functions. Sometimes, it is more convenient to replace them by appropriate lower and upper bounds.

**Lemma 5.1.5** *For any $t \ge 0$ we have*

$$\tfrac{t^2}{2(1+t)} \le \tfrac{t^2}{2\left(1+\frac{2}{3}t\right)} \le \omega(t) \le \tfrac{t^2}{2+t}, \tag{5.1.23}$$

*and for* $t \in [0, 1)$,

$$\tfrac{t^2}{2-t} \;\le\; \omega_*(t) \le \tfrac{t^2}{2(1-t)}. \tag{5.1.24}$$

*Proof* Let $\psi_1(t) = \frac{t^2}{2\left(1+\frac{2}{3}t\right)}$. Note that $\psi_1(0) = \omega(0) = 0$. At the same time,

$$\psi_1'(t) = \tfrac{t}{1+\frac{2}{3}t} - \tfrac{t^2}{3\left(1+\frac{2}{3}t\right)^2} \;=\; \tfrac{t(3+t)}{3\left(1+\frac{2}{3}t\right)^2} \;\le\; \tfrac{t}{1+t} \;=\; \omega'(t).$$

Similarly, for $\psi_2(t) = \frac{t^2}{2+t}$, we have

$$\psi_2'(t) = \tfrac{2t}{2+t} - \tfrac{t^2}{(2+t)^2} \;=\; \tfrac{4t+t^2}{(2+t)^2} \;\ge\; \tfrac{t}{1+t} \;=\; \omega'(t).$$

For the second inequality, let $\psi_3(t) = \frac{t^2}{2-t}$ and $\psi_4(t) = \frac{t^2}{2(1-t)}$. Then

$$\psi_3'(t) = \tfrac{2t}{2-t} + \tfrac{t^2}{(2-t)^2} \;=\; \tfrac{4t-t^2}{(2-t)^2} \;\le\; \tfrac{t}{1-t},$$

$$\psi_4'(t) = \tfrac{t}{1-t} + \tfrac{t^2}{2(1-t)^2} = \tfrac{2t-t^2}{2(1-t)^2} \ge \tfrac{t}{1-t}.$$

Since $\frac{t}{1-t} = \omega_*'(t)$ and $\omega_*(0) = \psi_3(0) = \psi_4(0) = 0$, we get (5.1.24) by integration. □

### 5.1.5 Self-Concordance and Fenchel Duality

Let us start with some preliminary results. Consider the following minimization problem:

$$\min\{f(x) \mid x \in \operatorname{dom} f\}, \tag{5.1.25}$$

where we assume that $f$ is self-concordant and all Hessians $\nabla^2 f(x)$, $x \in \operatorname{dom} f$, are positive definite. In view of Theorem 5.1.6, this can be derived from the fact that $\operatorname{dom} f$ contains no straight lines. Or, we can assume that $f$ is strongly convex.

Define

$$\lambda_f(x) = \langle \nabla f(x), [\nabla^2 f(x)]^{-1} \nabla f(x) \rangle^{1/2}.$$

We call $\lambda_f(x) = \| \nabla f(x) \|_x^*$ the *local norm of the gradient* $\nabla f(x)$.[4]

[4] Sometimes $\lambda_f(x)$ is called the *Newton decrement* of the function $f$ at $x$.

The next theorem provides us with a sufficient condition for existence of solution of problem (5.1.25).

**Theorem 5.1.13** *Let $\lambda_f(x) < \frac{1}{M_f}$ for some $x \in \mathrm{dom}\, f$. Then there exists a unique solution $x_f^*$ of problem (5.1.25) and*

$$f(x) - f(x_f^*) \le \frac{1}{M_f^2}\omega_*(M_f\lambda_f(x)). \tag{5.1.26}$$

*Proof* Indeed, in view of (5.1.14), for any $y \in \mathrm{dom}\, f$ we have

$$\begin{aligned} f(y) &\ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{M_f^2}\omega(M_f \parallel y - x \parallel_x) \\ &\ge f(x) - \lambda_f(x)\cdot \parallel y - x \parallel_x + \frac{1}{M_f^2}\omega(M_f \parallel y - x \parallel_x) \\ &= f(x) + \left(\frac{1}{M_f} - \lambda_f(x)\right)\|y - x\|_x - \frac{1}{M_f^2}\ln(1 + M_f\|y - x\|_x). \end{aligned}$$

Thus, the level set $\mathscr{L}_f(f(x))$ is bounded and therefore $x_f^*$ exists. It is unique since in view of (5.1.14), for all $y \in \mathrm{dom}\, f$ we have

$$f(y) \ge f(x_f^*) + \frac{1}{M_f^2}\omega(M_f \parallel y - x_f^* \parallel_{x_f^*}).$$

Finally, taking in (5.1.22) $x = x^*$ and $y = x$, we get inequality (5.1.26). □

Thus, we have proved that a local condition $\lambda_f(x) < \frac{1}{M_f}$ provides us with some global information on the function $f$, namely, the existence of the minimum $x_f^*$. Note that the result of Theorem 5.1.13 cannot be strengthened.

*Example 5.1.2* Let us fix some $\epsilon > 0$. Consider a function of one variable

$$f_\epsilon(x) = \epsilon x - \ln x, \quad x > 0.$$

This function is self-concordant in view of Example 5.1.1 and Corollary 5.1.2. Note that

$$\nabla f_\epsilon(x) = \epsilon - \frac{1}{x}, \quad \nabla^2 f_\epsilon = \frac{1}{x^2}.$$

Therefore $\lambda_{f_\epsilon}(x) =\mid 1 - \epsilon x \mid$. Thus, for $\epsilon = 0$ we have $\lambda_{f_0}(x) = 1$ for any $x > 0$. Note that the function $f_0$ is not bounded below.

If $\epsilon > 0$, then $x_{f_\epsilon}^* = \frac{1}{\epsilon}$. However, we can guarantee the existence of this point by collecting information at the point $x = 1$ even if $\epsilon$ is arbitrary small. □

Theorem 5.1.13 has several important consequences. One of them is called the *Theorem on Recession Direction*. Note that for its validity, we *do not need* the assumption that all Hessians of the function $f$ are positive definite.

**Theorem 5.1.14** *Let $h \in \mathbb{E}$ be a recession direction of the self-concordant function $f$: for any $x \in dom\, f$ we have*

$$\langle \nabla f(x), h \rangle \le 0,$$

*and there exists a $\tau = \tau(x)$ such that $x - \tau h \in \partial dom\, f$. Then*

$$\langle \nabla^2 f(x)h, h \rangle^{1/2} \le M_f \langle -\nabla f(x), h \rangle, \quad x \in dom\, f. \tag{5.1.27}$$

*Proof* Let us fix an arbitrary $x \in \mathrm{dom}\, f$. Consider a univariate function $\phi(\tau) = f(x + \tau h)$. This function is self-concordant and $0 \in \mathrm{dom}\, \phi$. As $\mathrm{dom}\, \phi$ contains no straight line, by Theorem 5.1.6, $\phi''(\tau) > 0$ for all $\tau \in \mathrm{dom}\, \phi$. Therefore, we must have

$$\lambda_\phi^2(0) \equiv \tfrac{\langle \nabla f(x), h \rangle^2}{\langle \nabla^2 f(x)h, h \rangle} \ge \tfrac{1}{M_f^2}$$

since otherwise, by Theorem 5.1.13, the minimum of $\phi(\cdot)$ exists. Thus,

$$\langle \nabla f(x), h \rangle^2 \ge \tfrac{1}{M_f^2} \langle \nabla^2 f(x)h, h \rangle,$$

and we get (5.1.27) taking into account the sign of the first derivative. □

Let us consider now the scheme of the *Damped Newton's method*.

```
Damped Newton's method

0. Choose $x_0 \in \mathrm{dom}\, f$.
1. Iterate $x_{k+1} = x_k - \frac{1}{1+M_f \lambda_f(x_k)} [\nabla^2 f(x_k)]^{-1} \nabla f(x_k), \; k \ge 0$.
```

(5.1.28)

**Theorem 5.1.15** *For any $k \ge 0$, we have*

$$f(x_{k+1}) \le f(x_k) - \tfrac{1}{M_f^2} \omega(M_f \lambda_f(x_k)). \tag{5.1.29}$$

*Proof* Let $\lambda = \lambda_f(x_k)$. Then $\| x_{k+1} - x_k \|_{x_k} = \frac{\lambda}{1+M_f\lambda} = \frac{1}{M_f}\omega'(M_f\lambda)$. Therefore, in view of (5.1.16) and Lemma 5.1.4, we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \tfrac{1}{M_f^2}\omega_*(M_f \| x_{k+1} - x_k \|_x) \\ &= f(x_k) - \tfrac{\lambda^2}{1+M_f\lambda} + \tfrac{1}{M_f^2}\omega_*(\omega'(M_f\lambda)) \\ &= f(x_k) - \tfrac{\lambda}{M_f}\omega'(M_f\lambda) + \tfrac{1}{M_f^2}\omega_*(\omega'(M_f\lambda)) \\ &= f(x_k) - \tfrac{1}{M_f^2}\omega(M_f\lambda). \qquad \square \end{aligned}$$

Thus, for all $x \in \mathrm{dom}\, f$ with $\lambda_f(x) \geq \beta > 0$, one step of the damped Newton's Method decreases the value of the function $f(\cdot)$ at least by a constant $\frac{1}{M_f^2}\omega(M_f\beta) > 0$. Note that the result of Theorem 5.1.15 is *global*. In Sect. 5.2 it will be used to obtaine a global efficiency bound of the process. However, now we employ it to prove an existence theorem. Recall that we assume that $\mathrm{dom}\, f$ contains no straight line.

**Theorem 5.1.16** *Let a self-concordant function $f$ be bounded below. Then it attains its minimum at a single point.*

*Proof* Indeed, assume that $f(x) \geq f^*$ for all $x \in \mathrm{dom}\, f$. Let us start the process (5.1.28) from some $x_0 \in \mathrm{dom}\, f$. If the number of steps of this method exceeds $M_f^2(f(x_0) - f^*)/\omega(1)$, then in view of (5.1.29) we must get a point $x_k$ with $\lambda_f(x_k) < \frac{1}{M_f}$. However, by Theorem 5.1.13 this implies the existence of a point $x_f^*$. It is unique since all Hessians of the function $f$ are nondegenerate. □

Now we can introduce the *Fenchel dual* of a self-concordant function $f$ (sometimes called a *conjugate function*, or *dual function* of $f$). For $s \in \mathbb{E}^*$, the value of this function is defined as follows:

$$f_*(s) = \sup_{x \in \mathrm{dom}\, f} [\langle s, x \rangle - f(x)]. \tag{5.1.30}$$

Clearly, $\mathrm{dom}\, f_* = \{s \in \mathbb{E}^* : \ f(x) - \langle s, x \rangle \text{ is bounded below on } \mathrm{dom}\, f\}$.

**Lemma 5.1.6** *The function $f_*$ is a closed convex function with nonempty open domain. Moreover, dom $f_* = \{\nabla f(x) : \ x \in \mathrm{dom}\, f\}$.*

*Proof* Indeed, for any $\bar{x} \in \mathrm{dom}\, f$, we have $\nabla f(\bar{x}) \in \mathrm{dom}\, f_*$. On the other hand, if $s \in \mathrm{dom}\, f_*$, then $f(x) - \langle s, x \rangle$ is below bounded. Hence, by Theorem 5.1.16 and the first-order optimality condition, there exists an $x \in \mathrm{dom}\, f$ such that $s = \nabla f(x)$.

Further, the epigraph of the function $f_*$ is an intersection of half-spaces

$$\{(s,\tau)\in \mathbb{E}^*\times\mathbb{R}:\ \tau\geq \langle s,x\rangle - f(x)\},\quad x\in \mathrm{dom}\, f,$$

which are closed and convex. Therefore, the epigraph of $f_*$ is also closed and convex.

Suppose for $s_1$ and $s_2$ from $\mathrm{dom}\, f_*$ we have

$$f(x)-\langle s_1,x\rangle \geq f_1^*,\quad f(x)-\langle s_2,x\rangle \ \geq\ f_2^*$$

for all $x\in \mathrm{dom}\, f$. Then, for any $\alpha\in[0,1]$

$$\begin{aligned} f(x)-\langle \alpha s_1+(1-\alpha)s_2,x\rangle &= \alpha(f(x)-\langle s_1,x\rangle)+(1-\alpha)(f(x)-\langle s_2,x\rangle)\\ &\geq \alpha f_1^*+(1-\alpha)f_2^*,\quad x\in \mathrm{dom}\, f. \end{aligned}$$

Thus, $\alpha s_1+(1-\alpha)s_2\in \mathrm{dom}\, f_*$.

Finally, let $s\in \mathrm{dom}\, f_*$. Denote by $x(s)\in \mathrm{dom}\, f$ the unique solution of the equation

$$s=\nabla f(x(s)).$$

Let $\delta\in\mathbb{E}^*$ be small enough: $\|\delta\|^*_{x(s)}<\frac{1}{M_f}$. Consider the function

$$f_\delta(x)=f(x)-\langle s+\delta,x\rangle.$$

Then $\nabla f_\delta(x(s))=\nabla f(x(s))-s-\delta=-\delta$. Therefore, $\lambda_{f_\delta}(x(s))=\|\delta\|^*_{x(s)}<\frac{1}{M_f}$. Thus, in view of Theorem 5.1.13 the function $f_\delta$ attains its minimum. Consequently, $s+\delta\in \mathrm{dom}\, f_*$, and we conclude that $s$ is an interior point of $\mathrm{dom}\, f_*$. □

*Example 5.1.3* Note that in general, the structure of the set $\{\nabla f(x):\ x\in \mathrm{dom}\, f\}$ can be quite complicated. Consider the function

$$f(x)=\frac{1}{x^{(1)}}\left(x^{(2)}\right)^2,\quad \mathrm{dom}\, f=\{x\in\mathbb{R}^2:\ x^{(1)}>0\}\bigcup\{0\},\quad f(0)=0.$$

In Example 3.1.2(5) we have seen that this is a closed convex function. However,

$$\nabla f(x)=\left(-\left(\frac{x^{(2)}}{x^{(1)}}\right)^2, 2\frac{x^{(2)}}{x^{(1)}}\right),\ x\neq 0,\quad \nabla f(0)=0.$$

Thus, $\{\nabla f(x):\ x\in \mathrm{dom}\, f\}=\{g\in\mathbb{R}^2:\ g^{(1)}=-\frac{1}{2}(g^{(2)})^2\}$. □

Let us now look at the derivatives of the function $f_*$. Since $f$ is self-concordant, for any $s\in \mathrm{dom}\, f_*$, the supremum in (5.1.30) is attained (see Theorem 5.1.16).

Define

$$x(s) = \arg\max_{x\in\mathrm{dom}\, f} [\langle s, x\rangle - f(x)].$$

Thus,

$$\nabla f(x(s)) = s. \tag{5.1.31}$$

In view of Lemma 3.1.14, we have $x(s) \in \partial f_*(s)$. On the other hand, for $s_1$ and $s_2$ from dom $f_*$ we have

$$\begin{aligned}
\frac{\|x(s_1)-x(s_2)\|^2_{x(s_1)}}{1+M_f\|x(s_1)-x(s_2)\|_{x(s_1)}} &\overset{(5.1.13)}{\leq} \langle \nabla f(x(s_1)) - \nabla f(x(s_2)), x(s_1) - x(s_2)\rangle \\
&\overset{(5.1.31)}{=} \langle s_1 - s_2, x(s_1) - x(s_2)\rangle \\
&\leq \|s_1 - s_2\|^*_{x(s_1)} \|x(s_1) - x(s_2)\|_{x(s_1)}.
\end{aligned}$$

Thus, $x(s)$ is a continuous function of $s$ and by Lemma 3.1.10 we conclude that

$$\nabla f_*(s) = x(s). \tag{5.1.32}$$

Let us differentiate identities (5.1.31) and (5.1.32) along direction $h \in \mathbb{E}^*$:

$$\nabla^2 f(x(s))x'(s)h = h, \quad \nabla^2 f_*(s)h \; = \; x'(s)h.$$

Thus,

$$\nabla^2 f_*(s) = [\nabla^2 f(x(s))]^{-1}, \quad s \in \mathrm{dom}\, f_*. \tag{5.1.33}$$

In other words, if $s = \nabla f(x)$, then

$$\nabla^2 f_*(s) = [\nabla^2 f(x)]^{-1}, \quad x \in \mathrm{dom}\, f. \tag{5.1.34}$$

Let us compute the third derivative of the dual function $f_*$ along direction $h \in \mathbb{E}^*$ using the representation (5.1.33).

$$\begin{aligned}
D^3 f_*(s)[h] \; &= \; \lim_{\alpha\to 0} \tfrac{1}{\alpha} \left([\nabla^2 f(x(s+\alpha h))]^{-1} - [\nabla^2 f(x(s))]^{-1}\right) \\
&= \lim_{\alpha\to 0} \tfrac{1}{\alpha} [\nabla^2 f(x(s))]^{-1} \left(\nabla^2 f(x(s)) - \nabla^2 f(x(s+\alpha h))\right) [\nabla^2 f(x(s+\alpha h))]^{-1} \\
&= -[\nabla^2 f(x(s))]^{-1} D^3 f(x(s))[x'(s)h][\nabla^2 f(x(s))]^{-1}.
\end{aligned}$$

Thus, we have proved the following representation:

$$D^3 f_*(s)[h] = \nabla^2 f_*(s) D^3 f(x(s)) \left[-\nabla^2 f_*(s)h\right] \nabla^2 f_*(s), \tag{5.1.35}$$

which is valid for all $s \in \mathrm{dom}\, f_*$ and $h \in \mathbb{E}^*$. Now we can prove our main statement.

**Theorem 5.1.17** *The function $f_*$ is self-concordant with $M_{f_*} = M_f$.*

*Proof* Indeed, in view of Lemma 5.1.6, $f_*$ is a closed convex function with open domain. Further, for any $s \in \mathrm{dom}\, f_*$ and $h \in \mathbb{E}^*$ we have

$$\|\nabla^2 f_*(s)h\|^2_{x(s)} \overset{(5.1.33)}{=} \langle h, \nabla^2 f_*(s)h\rangle \overset{\text{def}}{=} r^2.$$

Therefore, in view of (5.1.35),

$$D^3 f_*(s)[h] \overset{(5.1.6)}{\preceq} 2M_f r\, \nabla^2 f_*(s)\, \nabla^2 f(x(s))\, \nabla^2 f_*(s) \overset{(5.1.33)}{=} 2M_f r\, \nabla^2 f_*(s).$$

It remains to use Corollary 5.1.1. □

As an example of application of Theorem 5.1.17, let us prove the following result.

**Lemma 5.1.7** *Let $x, y \in \mathrm{dom}\, f$ and $d = \|\nabla f(x) - \nabla f(y)\|^*_x < \frac{1}{M_f}$. Then*

$$(1 - M_f d)^2 \nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \tfrac{1}{(1-M_f d)^2} \nabla^2 f(x). \tag{5.1.36}$$

*Proof* Let $u = \nabla f(x)$ and $v = \nabla f(y)$. In view of Lemma 5.1.6, both points belong to $\mathrm{dom}\, f_*$. Note that

$$d^2 = (\|\nabla f(x) - \nabla f(y)\|^*_x)^2 = \langle u - v, \nabla^2 f_*(u)(u - v)\rangle.$$

Since $f_*$ is self-concordant with constant $M_f$, by Theorem 5.1.7 we have

$$(1 - M_f d)^2 \nabla^2 f_*(u) \preceq \nabla^2 f_*(v) \preceq \tfrac{1}{(1-M_f d)^2} \nabla^2 f_*(u).$$

In view of (5.1.33), this is exactly (5.1.36). □

*Remark 5.1.2* Some results on self-concordant functions have a more natural *dual* interpretation. Let us look at the statement of Theorem 5.1.13. Since the function $f_*$ is self-concordant, for any $\bar{s} \in \mathrm{dom}\, f_*$ the ellipsoid

$$W^0_*(\bar{s}) = \left\{ s \in \mathbb{E}^* :\ \langle s - \bar{s}, \nabla^2 f_*(\bar{s})(s - \bar{s})\rangle < \tfrac{1}{M_f^2} \right\}$$

belongs to dom $f_*$. Note that for $\bar{s} = \nabla f(x)$, in view of (5.1.33), condition $\lambda_f(x) < \frac{1}{M_f}$ is equivalent to

$$\langle \bar{s}, \nabla^2 f_*(\bar{s})\bar{s} \rangle < \frac{1}{M_f^2}.$$

This guarantees that $0 \in W_*^0(\bar{s})$ . Consequently, $0 \in \operatorname{dom} f_*$ and consequently the function $f_*$ is below bounded. □

## 5.2 Minimizing Self-concordant Functions

(Local convergence of different variants of Newton's Method; Path-following method; Minimization of strongly convex functions.)

### *5.2.1 Local Convergence of Newton's Methods*

In this section, we are going to study the complexity of solving the problem (5.1.25) by different optimization strategies. Let us look first at different variants of Newton's Method.

**Variants of Newton's Method**

**0.** Choose $x_0 \in \operatorname{dom} f$.
**1.** For $k \geq 0$, iterate

$$x_{k+1} = x_k - \tfrac{1}{1+\xi_k}[\nabla^2 f(x_k)]^{-1}\nabla f(x_k), \tag{5.2.1}$$

where $\xi_k$ is chosen in one of the following ways:

(A) $\xi_k = 0$ (this is the *Standard* Newton's Method),

(B) $\xi_k = M_f\lambda_k$ (this is the *Damped* Newton's Method (5.1.28)),

(C) $\xi_k = \frac{M_f^2\lambda_k^2}{1+M_f\lambda_k}$ (this is the *Intermediate* Newton's Method),

where $\lambda_k = \lambda_f(x_k)$.

We call method $(5.2.1)_C$ intermediate since for big $\lambda_k$ it is close to variant B, and for small values of $\lambda_k$ it is very close to variant A. However, note that its step size is always bigger than the step size of variant B, which was obtained by minimizing an upper bound for the self-concordant function (see the proof of Theorem 5.1.15). Nevertheless, method $(5.2.1)_C$ ensures a monotone decrease of the value of objective function in problem (5.1.25).

**Lemma 5.2.1** *Let points $\{x_k\}_{k\geq 0}$ be generated by method $(5.2.1)_C$. Then, for any $k \geq 0$ we have*

$$f(x_k) - f(x_{k+1}) \geq \frac{\lambda_k^2}{2(1+M_f\lambda_k+M_f^2\lambda_k^2)} + \frac{M_f\lambda_k^3}{2(1+M_f\lambda_k)(3+2M_f\lambda_k)}. \tag{5.2.2}$$

*Proof* Indeed, in view of inequality (5.1.16), we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \tfrac{\lambda_k^2}{1+\xi_k} + \tfrac{1}{M_f^2}\omega_*\left(\tfrac{M_f\lambda_k}{1+\xi_k}\right) \\ &= f(x_k) - \tfrac{\lambda_k^2(1+M_f\lambda_k)}{1+M_f\lambda_k+M_f^2\lambda_k^2} + \tfrac{1}{M_f^2}\left[-\tfrac{M_f\lambda_k(1+M_f\lambda_k)}{1+M_f\lambda_k+M_f^2\lambda_k^2} + \ln\left(1+M_f\lambda_k+M_f^2\lambda_k^2\right)\right]. \end{aligned}$$

Defining $\tau_k = M_f\lambda_k$, we have

$$\begin{aligned} \tfrac{\tau_k(1+\tau_k)^2}{1+\tau_k+\tau_k^2} - \ln\left(1+\tau_k+\tau_k^2\right) &= \tfrac{\tau_k(1+\tau_k)^2}{1+\tau_k+\tau_k^2} - \tau_k + \omega(\tau_k) - \ln\left(1+\tfrac{\tau_k^2}{1+\tau_k}\right) \\ \overset{(5.1.23)}{\geq} \tfrac{\tau_k^2}{1+\tau_k+\tau_k^2} + \tfrac{\tau_k^2}{2\left(1+\frac{2}{3}\tau_k\right)} - \ln\left(1+\tfrac{\tau_k^2}{1+\tau_k}\right) &= \tfrac{\tau_k^2}{2\left(1+\frac{2}{3}\tau_k\right)} - \xi_k + \tfrac{\xi_k}{1+\xi_k} + \omega(\xi_k). \end{aligned}$$

It remains to note that

$$\frac{\tau_k^2}{2\left(1+\frac{2}{3}\tau_k\right)} - \frac{1}{2}\xi_k = \frac{\tau_k^2}{2}\left(\frac{1}{1+\frac{2}{3}\tau_k} - \frac{1}{1+\tau_k}\right) = \frac{\tau_k^3}{2(1+\tau_k)(3+2\tau_k)},$$

and $-\frac{\xi_k}{2} + \omega(\xi_k) \overset{(5.1.23)}{\geq} -\frac{\xi_k}{2} + \frac{\xi_k^2}{2(1+\xi_k)} = -\frac{\xi_k}{2(1+\xi_k)}$. □

Let us describe now the *local* convergence of different variants of the Newton's Method. Note that we can measure the convergence of these schemes in four different ways. We can estimate the rate of convergence for the *functional gap* $f(x_k) - f(x_f^*)$, or for the local norm of the gradient $\lambda_f(x_k) = \parallel \nabla f(x_k) \parallel_{x_k}^*$, or for the *local distance to the minimum* $\parallel x_k - x_f^* \parallel_{x_k}$. Finally, we can look at the distance to the minimum in a fixed metric

$$r_*(x_k) = \|x_k - x_f^*\|_{x_f^*},$$

defined by the minimum itself. Let us prove that locally all these measures are equivalent.

**Theorem 5.2.1** *Let $\lambda_f(x) < \frac{1}{M_f}$. Then*

$$\omega(M_f\lambda_f(x)) \;\le\; M_f^2(f(x) - f(x_f^*)) \;\le\; \omega_*(M_f\lambda_f(x)), \tag{5.2.3}$$

$$\omega'(M_f\lambda_f(x)) \;\le\; M_f \parallel x - x_f^* \parallel_x \le\; \omega_*'(M_f\lambda_f(x)), \tag{5.2.4}$$

$$\omega(M_f r_*(x)) \;\le\; M_f^2(f(x) - f(x_f^*)) \;\le\; \omega_*(M_f r_*(x)), \tag{5.2.5}$$

*where the last inequality is valid for $r_*(x) < \frac{1}{M_f}$.*

*Proof* Let $r =\parallel x - x_f^* \parallel_x$ and $\lambda = \lambda_f(x)$. Inequalities (5.2.3) follow from Theorem 5.1.12. Further, in view of (5.1.13), we have

$$\tfrac{r^2}{1+M_f r} \le \langle \nabla f(x), x - x_f^* \rangle \;\le\; \lambda r.$$

Applying the function $\omega_*'(\cdot)$ to both sides of inequality $\frac{M_f r}{1+M_f r} \le M_f\lambda$, we get the right-hand side of inequality (5.2.4). If $r \ge \frac{1}{M_f}$, then the left-hand side of this inequality is trivial. Suppose that $r < \frac{1}{M_f}$. Then $\nabla f(x) = G(x - x_f^*)$ with

$$G = \int\limits_0^1 \nabla^2 f(x_f^* + \tau(x - x_f^*))d\tau \;\succ\; 0,$$

and $\lambda_f^2(x) = \langle G[\nabla^2 f(x)]^{-1} G(x - x_f^*), x - x_f^* \rangle$. Let us introduce in $\mathbb{E}$ a canonical basis. Then all self-adjoint operators from $\mathbb{E}$ to $\mathbb{E}^*$ can be represented by symmetric matrices (we do not change the existing notation). Define

$$H = \nabla^2 f(x), \quad S \;=\; H^{-1/2} G H^{-1} G H^{-1/2} \;=\; \left(H^{-1/2} G H^{-1/2}\right)^2 \;\overset{\text{def}}{=}\; P^2 \succ 0.$$

Then $\| H^{1/2}(x - x_f^*) \|_2 = \| x - x_f^* \|_x = r$, where $\| \cdot \|_2$ is the standard Euclidean norm, and

$$\lambda_f(x) \;=\; \langle H^{1/2} S H^{1/2}(x - x^*), x - x^* \rangle^{1/2} \;\le \parallel P \parallel_2 \| H^{1/2}(x - x^*) \|_2 \;= \parallel P \parallel_2 r.$$

In view of Corollary 5.1.5 (see Remark 5.1.1), we have

$$G \preceq \tfrac{1}{1-M_f r} H.$$

Therefore, $\parallel P \parallel_2 \le \frac{1}{1-M_f r}$ and we conclude that

$$M_f \lambda_f(x) \le \tfrac{M_f r}{1-M_f r} \;=\; \omega_*'(M_f r).$$

Applying the function $\omega'(\cdot)$ to both sides of this inequality, we get the remaining part of (5.2.4). Finally, inequalities (5.2.5) follow from (5.1.14) and (5.1.16). □

We are going to estimate the local rate of convergence of different variants of the Newton's method (5.2.1) in terms of $\lambda_f(\cdot)$, the local norm of the gradient.

**Theorem 5.2.2** *Let $x \in dom\, f$ and $\lambda = \lambda_f(x)$.*

*1. If $\lambda < \frac{1}{M_f}$ and the point $x_+$ is generated by variant A of method (5.2.1), then $x_+ \in dom\, f$ and*

$$\lambda_f(x_+) \le \frac{M_f\lambda^2}{(1-M_f\lambda)^2}. \tag{5.2.6}$$

*2. If point $x_+$ is generated by variant B of method (5.2.1), then $x_+ \in dom\, f$ and*

$$\lambda_f(x_+) \le M_f\lambda^2\left(1 + \frac{1}{1+M_f\lambda}\right). \tag{5.2.7}$$

*3. If $M_f\lambda + M_f^2\lambda^2 + M_f^3\lambda^3 \le 1$ and point $x_+$ is generated by method (5.2.1)$_C$, then $x_+ \in dom\, f$ and*

$$\lambda_f(x_+) \le M_f\lambda^2\left(1 + M_f\lambda + \frac{M_f\lambda}{1+M_f\lambda+M_f^2\lambda^2}\right) \;\le\; M_f\lambda^2\left(1+2M_f\lambda\right). \tag{5.2.8}$$

*Proof* Let $h = x_+ - x$, $\lambda = \lambda_f(x)$, and $r = \|h\|_x$. Then $r = \frac{\lambda}{1+\xi}$. Note that for all variants of method (5.2.1), we have $M_f\lambda < 1+\xi$. Therefore, in all cases, $M_f r < 1$ and $x_+ \in \operatorname{dom} f$ (see Theorem 5.1.5). Hence, in view of Theorem 5.1.7 we have

$$\lambda_f(x_+) = \langle \nabla f(x_+), [\nabla^2 f(x_+)]^{-1}\nabla f(x_+)\rangle^{1/2} \;\le\; \frac{1}{1-M_f r} \,\|\, \nabla f(x_+) \,\|_x^* .$$

Further, by (5.2.1)

$$\nabla f(x_+) \;=\; \nabla f(x) + \int\limits_0^1 \nabla^2 f(x+\tau h)h d\tau \;=\; Gh,$$

where $G = \int\limits_0^1 [\nabla^2 f(x+\tau h) - (1+\xi)\nabla^2 f(x)]d\tau$. As in the proof of Theorem 5.2.1, let us pass to matrices. Define

$$H = \nabla^2 f(x), \quad S \;=\; H^{-1/2}GH^{-1}GH^{-1/2} \;\overset{\mathrm{def}}{=}\; P^2,$$

where $P = H^{-1/2}GH^{-1/2}$. Then $\|H^{1/2}h\|_2 = \|h\|_x = r$, and

$$\|\, \nabla f(x_+) \,\|_x^* = \langle Gh, H^{-1}Gh\rangle^{1/2} \;=\; \langle H^{1/2}SH^{1/2}h, h\rangle^{1/2} \;\le \|\, P \,\|_2\, r.$$

In view of Corollary 5.1.5,

$$\left(-\xi - M_f r + \tfrac{1}{3}M_f^2 r^2\right) H \preceq G \ \preceq \ \left(\tfrac{1}{1-M_f r} - (1+\xi)\right) H.$$

Therefore, $\| \ P \ \|_2 \leq \max\left\{\frac{M_f r}{1-M_f r} - \xi, M_f r + \xi\right\}$.

For the variant A, $\xi = 0$. Thus, $r = \lambda$ and we get $\|P\|_2 \leq \frac{M_f\lambda}{1-M_f\lambda}$. Therefore,

$$\lambda_f(x_+) \leq \tfrac{\lambda}{1-M_f\lambda}\|P\|_2 \ \leq \ \tfrac{M_f\lambda^2}{(1-M_f\lambda)^2}.$$

For the variant B, $\xi = M_f\lambda$. Therefore, $r = \frac{\lambda}{1+M_f\lambda}$, and we get $\|P\|_2 \leq M_f\lambda + \frac{M_f\lambda}{1+M_f\lambda}$. Consequently,

$$\lambda_f(x_+) \leq \tfrac{r}{1-M_f r}\|P\|_2 \ \leq \ M_f\lambda^2\left(1 + \tfrac{1}{1+M_f\lambda}\right).$$

Finally, for variant C, $\xi = \frac{M_f^2\lambda^2}{1+M_f\lambda}$. Then, $r = \frac{\lambda(1+M_f\lambda)}{1+M_f\lambda+M_f^2\lambda^2}$, and we have

$$\begin{aligned}\tfrac{M_f r}{1-M_f r} - M_f r - \xi \ &= \ \tfrac{M_f^2 r^2}{1-M_f r} - \xi \ = \ \tfrac{M_f^2\lambda^2(1+M_f\lambda)^2}{1+M_f\lambda+M_f^2\lambda^2} - \tfrac{M_f^2\lambda^2}{1+M_f\lambda}\\ &= \tfrac{M_f^2\lambda^2(2M_f\lambda+2M_f^2\lambda^2+M_f^3\lambda^3)}{(1+M_f\lambda+M_f^2\lambda^2)(1+M_f\lambda)} \ = \ \tfrac{\xi(2M_f\lambda+2M_f^2\lambda^2+M_f^3\lambda^3)}{1+M_f\lambda+M_f^2\lambda^2} \ \leq \ \xi\end{aligned}$$

in view of the condition of this item. Hence

$$\begin{aligned}\lambda_f(x_+) &\leq \tfrac{r}{1-M_f r}\|P\|_2 \ \leq \ \tfrac{r}{1-M_f r}(M_f r + \xi)\\ &= \tfrac{\lambda(1+M_f\lambda)}{1+M_f\lambda+M_f^2\lambda^2}(1 + M_f\lambda + M_f^2\lambda^2)\left(\tfrac{M_f\lambda(1+M_f\lambda)}{1+M_f\lambda+M_f^2\lambda^2} + \tfrac{M_f^2\lambda^2}{1+M_f\lambda}\right)\\ &= M_f\lambda^2\left(\tfrac{(1+M_f\lambda)^2}{1+M_f\lambda+M_f^2\lambda^2} + M_f\lambda\right)\\ &= M_f\lambda^2\left(1 + M_f\lambda + \tfrac{M_f\lambda}{1+M_f\lambda+M_f^2\lambda^2}\right). \qquad \square\end{aligned}$$

Among all variants of the rate of convergence, described in Theorem 5.2.2, the estimate (5.2.8) looks more attractive. It provides us with the following description of the region of quadratic convergence for method (5.2.1)$_C$:

$$\mathcal{Q}_f \overset{\text{def}}{=} \left\{ x \in \text{dom}\, f : \lambda_f(x) < \tfrac{1}{2M_f} \right\}. \tag{5.2.9}$$

In this case, we can guarantee that $\lambda_f(x_+) < \lambda_f(x)$, and then the quadratic convergence starts (see (5.2.8)). Thus, our results lead to the following strategy for solving the initial problem (5.1.25).

- *First stage*: $\lambda_f(x_k) \geq \frac{1}{2M_f}$. At this stage we apply the Damped Newton's Method (5.1.28). At each iteration of this method, we have

  $$f(x_{k+1}) \leq f(x_k) - \tfrac{1}{M_f^2}\omega(\tfrac{1}{2}).$$

  Thus, the number of steps of this stage is bounded as follows:

  $$N \leq M_f^2[f(x_0) - f(x_f^*)]/\omega(\tfrac{1}{2}). \tag{5.2.10}$$

- *Second stage*: $\lambda_f(x_k) < \frac{1}{2M_f}$. At this stage, we apply method (5.2.1)$_C$. This process converges quadratically:

  $$\lambda_f(x_{k+1}) \leq M_f \lambda_f^2(x_k)(1 + 2M_f\lambda_f(x_k)) < \lambda_f(x_k).$$

Since the quadratic convergence is very fast, the main efforts in the above strategy are spent at the first stage. The estimate (5.2.10) shows that the length of this stage is $O(\Delta_f(x_0))$, where

$$\Delta_f(x_0) \overset{\text{def}}{=} M_f^2[f(x_0) - f(x_f^*)]. \tag{5.2.11}$$

Is it possible to reach the region of quadratic convergence in a faster way? In order to answer this question, let us consider an alternative way to solve the problem (5.1.25), based on a *path-following scheme*. In Sect. 5.3 we will see how we can use this idea for solving a *constrained* minimization problem.

### 5.2.2 Path-Following Scheme

Assume that we have $y_0 \in \text{dom}\, f$. Let us define an *auxiliary central path*

$$y(t) = \arg \min_{y \in \text{dom}\, f} \left[ \psi(t; y) \overset{\text{def}}{=} f(y) - t\langle \nabla f(y_0), y \rangle \right], \quad t \in [0, 1]. \tag{5.2.12}$$

This minimization problem corresponds to computation of the value of the dual function $-f_*(s)$ with $s = t\nabla f(y_0)$ (see (5.1.30)). Note that $\nabla f(y_0) \in \text{dom}\, f_*$ and the origin in the dual space also belongs to $\mathrm{dom}\, f_*$ since the problem (5.1.25) is solvable. Therefore, in view of Lemma 5.1.6,

$$t\nabla f(y_0) \in \mathrm{dom}\, f_*, \quad 0 \le t \le 1,$$

and trajectory (5.2.12) is well defined.

We are going to follow the auxiliary central path with parameter $t$ changing from one to zero by updating points satisfying the *approximate centering condition*

$$\lambda_{\psi(t;\cdot)}(y) \stackrel{\mathrm{def}}{=} \|\nabla f(y) - t\nabla f(y_0)\|_y^* \le \tfrac{\beta}{M_f}, \tag{5.2.13}$$

where the *centering parameter* $\beta$ is small enough. Note that the function $\psi(t;\cdot)$ is self-concordant with constant $M_f$ and domain $\mathrm{dom}\, f$ (see Corollary 5.1.2).

Consider the following iterate:

$$(t_+, y_+) = \mathscr{P}_\gamma(t,y) \equiv \begin{cases} t_+ = t - \frac{\gamma}{M_f \|\nabla f(y_0)\|_y^*}, \\ \\ y_+ = y - \frac{[\nabla^2 f(y)]^{-1}(\nabla f(y) - t_+\nabla f(y_0))}{1+\xi}, \end{cases} \tag{5.2.14}$$

where $\xi = \frac{M_f^2\lambda^2}{1+M_f\lambda}$ and $\lambda = \lambda_{\psi(t;\cdot)}(y)$ (this is one iteration of method $(5.2.1)_C$). For future use, we allow the parameter $\gamma$ in (5.2.14) to be both positive or negative.

**Lemma 5.2.2** *Let the pair $(t,y)$ satisfy (5.2.13) with $\beta = \tau^2(1+\tau+\frac{\tau}{1+\tau+\tau^2})$, where $\tau \le \frac{1}{2}$. Then the pair $(t_+,y_+)$ satisfies the same condition for $\gamma$ small enough, namely*

$$|\gamma| \le \tau - \tau^2\left(1+\tau+\tfrac{\tau}{1+\tau+\tau^2}\right). \tag{5.2.15}$$

*Proof* Let $\lambda = \|\nabla f(y) - t\nabla f(y_0)\|_y^* \le \frac{\beta}{M_f}$, $\lambda_1 = \|\nabla f(y) - t_+\nabla f(y_0)\|_y^*$, and $\lambda_+ = \|\nabla f(y) - t_+\nabla f(y_0)\|_{y_+}^*$. Then $\lambda_1 \le \lambda + \frac{|\gamma|}{M_f} \le \frac{1}{M_f}(\beta+|\gamma|) \overset{(5.2.15)}{\le} \frac{\tau}{M_f}$. Hence,

$$\lambda_+ \overset{(5.2.8)}{\le} \frac{\tau^2}{M_f}\left(1+\tau+\tfrac{\tau}{1+\tau+\tau^2}\right) = \frac{\beta}{M_f}. \qquad \square$$

Let us derive from this fact a complexity bound of the path-following scheme as applied to problem (5.1.25).

**Theorem 5.2.3** *Consider the following process:*

$$t_0 = 1, \ y_0 \in dom\, f, \quad (t_{k+1}, y_{k+1}) = \mathscr{P}_\gamma(t_k, y_k), \quad k \geq 0, \tag{5.2.16}$$

*where* $\gamma = \gamma(\tau) = \tau - \beta$, $\beta = \beta(\tau) = \tau^2\left(1 + \tau + \frac{\tau}{1+\tau+\tau^2}\right)$, *and* $\tau \leq 0.23$. *Then*

$$\lambda_k \overset{def}{=} \|\nabla f(y_k) - t_k \nabla f(y_0)\|^*_{y_k} \leq \frac{\beta}{M_f}, \quad k \geq 0. \tag{5.2.17}$$

*Assume that* $\lambda_f(y_k) \geq \frac{1}{2M_f}$ *for all* $k = 0, \ldots, N$. *Then*

$$t_N \leq \exp\left\{-\frac{\gamma\varkappa(\tau)N^2}{\Delta_f(x_0)}\right\}, \tag{5.2.18}$$

*where* $\varkappa(\tau) = \frac{(\tau-3\beta)(1+\beta)}{2(1+\beta+\beta^2)}$.

*Proof* Since $\lambda_0 = 0 < \frac{\beta}{M_f}$, by Lemma 5.2.2 we prove that inequality (5.2.17) is valid for all $k \geq 0$. Let $c = -\nabla f(y_0)$. Note that

$$y_k - y_{k+1} \overset{(5.2.14)}{=} \frac{1}{1+\xi_k}[\nabla^2 f(y_k)]^{-1}\left(t_k c + \nabla f(y_k) - \frac{\gamma c}{M_f \|c\|^*_{y_k}}\right), \tag{5.2.19}$$

where $\xi_k = \frac{M_f^2 \lambda_k^2}{1+M_f\lambda_k}$. Therefore,

$$r_k \overset{\text{def}}{=} \|y_k - y_{k+1}\|_{y_k} \leq \frac{\lambda_k}{1+\xi_k} + \frac{\gamma}{M_f(1+\xi_k)} = \frac{\gamma+M_f\lambda_k}{M_f(1+\xi_k)} \overset{(5.2.17)}{\leq} \frac{\tau}{M_f}. \tag{5.2.20}$$

Further,

$$t_{k+1} \overset{(5.2.14)}{=} t_k - \frac{\gamma}{M_f\|c\|^*_{y_k}} = t_k\left(1 - \frac{\gamma}{M_f t_k \|c\|^*_{y_k}}\right) \leq t_k \exp\left\{-\frac{\gamma}{M_f t_k \|c\|^*_{y_k}}\right\}.$$

Thus, $t_N \leq \exp\left\{-\frac{\gamma}{M_f} S_N\right\}$, where $S_N = \sum\limits_{k=0}^{N} \frac{1}{t_k\|c\|^*_{y_k}}$. Let us estimate this value from below.

Since $\frac{\beta^2}{M_f^2} \overset{(5.2.17)}{\geq} \lambda_f^2(y_k) + 2t_k\langle\nabla f(y_k), [\nabla^2 f(y_k)]^{-1}c\rangle + t_k^2(\|c\|^*_{y_k})^2$, we have

$$-\langle\nabla f(y_k), [\nabla^2 f(y_k)]^{-1}c\rangle \geq \frac{1}{2t_k}\left[\lambda_f^2(y_k) + t_k^2(\|c\|^*_{y_k})^2 - \frac{\beta^2}{M_f^2}\right]. \tag{5.2.21}$$

Therefore,

$$
\begin{aligned}
f(y_k) - f(y_{k+1}) &\overset{(5.1.16)}{\geq} \langle \nabla f(y_k), y_k - y_{k+1} \rangle - \tfrac{1}{M_f^2}\omega_*(M_f r_k) \\
&\overset{(5.2.19)}{=} \tfrac{1}{1+\xi_k}\langle \nabla f(y_k), [\nabla^2 f(y_k)]^{-1}\left(t_k c + \nabla f(y_k) - \tfrac{\gamma c}{M_f \|c\|_{y_k}^*}\right)\rangle - \tfrac{1}{M_f^2}\omega_*(M_f r_k) \\
&= \tfrac{\lambda_k^2}{1+\xi_k} - \tfrac{t_k}{1+\xi_k}\langle c, [\nabla^2 f(y_k)]^{-1}(t_k c + \nabla f(y_k))\rangle \\
&\quad + \tfrac{1}{1+\xi_k}\langle \nabla f(y_k), [\nabla^2 f(y_k)]^{-1}\left(\tfrac{-\gamma c}{M_f \|c\|_{y_k}^*}\right)\rangle - \tfrac{1}{M_f^2}\omega_*(M_f r_k) \\
&\overset{(5.2.17)}{\geq} \tfrac{\lambda_k^2 - t_k \|c\|_{y_k}^* \lambda_k}{1+\xi_k} - \tfrac{\gamma}{M_f \|c\|_{y_k}^*(1+\xi_k)}\langle \nabla f(y_k), [\nabla^2 f(y_k)]^{-1} c\rangle - \tfrac{1}{M_f^2}\omega_*(M_f r_k) \\
&\overset{(5.2.21)}{\geq} \tfrac{\lambda_k^2 - t_k \|c\|_{y_k}^* \lambda_k}{1+\xi_k} + \tfrac{\gamma}{2M_f t_k \|c\|_{y_k}^*(1+\xi_k)}\left[\lambda_f^2(y_k) + t_k^2(\|c\|_{y_k}^*)^2 - \tfrac{\beta^2}{M_f^2}\right] \\
&\quad - \tfrac{1}{M_f^2}\omega_*(M_f r_k) \\
&\overset{(5.2.20)}{\geq} \tfrac{\gamma - 2M_f\lambda_k}{2M_f(1+\xi_k)} t_k \|c\|_{y_k}^* + \rho_k,
\end{aligned}
$$

where $\rho_k = \frac{\gamma}{2M_f t_k \|c\|_{y_k}^*(1+\xi_k)}\left[\lambda_f^2(y_k) - \frac{\beta^2}{M_f^2}\right] - \frac{1}{M_f^2}\omega_*(\tau)$.

Our next goal is to show that $\rho_k \geq 0$. Note that $t_k \|c\|_{y_k}^* \overset{(5.2.17)}{\leq} \lambda_f(y_k) + \frac{\beta}{M_f}$. Since $\lambda_f(y_k) \geq \frac{1}{2M_f}$, we have

$$
\begin{aligned}
\rho_k &\geq \tfrac{\gamma}{2M_f(1+\xi_k)}\left[\lambda_f(y_k) - \tfrac{\beta}{M_f}\right] - \tfrac{1}{M_f^2}\omega_*(\tau) \geq \tfrac{\gamma(1-2\beta)}{4M_f^2(1+\xi_k)} - \tfrac{1}{M_f^2}\omega_*(\tau) \\
&\overset{(5.2.17)}{\geq} \tfrac{1}{M_f^2}\left[\tfrac{\gamma(1-2\beta)(1+\beta)}{4(1+\beta+\beta^2)} - \omega_*(\tau)\right].
\end{aligned}
$$

Note that $\gamma = O(\tau)$, $\beta = O(\tau^2)$, and $\omega_*(\tau) = O(\tau^2)$. Therefore, for $\tau$ small enough we have $\rho_k \geq 0$. By numerical evaluation, it is easy to check that this can be achieved by taking $\tau \leq 0.23$.

Further,

$$
\tfrac{\gamma - 2M_f\lambda_k}{2(1+\xi_k)} \overset{(5.2.17)}{\geq} \tfrac{(\gamma-2\beta)(1+\beta)}{2(1+\beta+\beta^2)} = \tfrac{(\tau-3\beta)(1+\beta)}{2(1+\beta+\beta^2)} \overset{\text{def}}{=} \varkappa(\tau).
$$

Again, it is easy to check that $\varkappa(\tau) > 0$ for $\tau \in (0, 0.23]$. Thus, we have proved that $f(y_k) - f(y_{k+1}) \geq \frac{\varkappa(\tau)}{M_f} t_k \|c\|_{y_k}$. Therefore,

$$S_N \geq \sum_{k=0}^{N} \frac{\varkappa(\tau)}{M_f(f(y_k)-f(y_{k+1}))} \geq \frac{\varkappa(\tau)\Lambda^*(N)}{M_f(f(y_0)-f(y_{N+1}))},$$

where $\Lambda^*(N) = \min\limits_{\lambda \in \mathbb{R}^{N+1}_+} \left\{ \sum\limits_{i=1}^{N+1} \frac{1}{\lambda^{(i)}} : \sum\limits_{i=1}^{N+1} \lambda^{(i)} = 1 \right\} = (N+1)^2$. □

Let us estimate now the number of iterations, which are necessary for method (5.2.16) to enter the region of quadratic convergence $\mathscr{Q}_f$. Define

$$D = \max_{x,y \in \mathrm{dom}\, f} \{ \|x - y\|_{y_0} : \ f(x) \leq f(y_0), \ f(y) \leq f(y_0) \}.$$

**Theorem 5.2.4** *Let the sequence $\{y_k\}_{k \geq 0}$ be generated by method (5.2.16). Then for all*

$$N \geq \left[ \frac{\Delta_f(x_0)}{\gamma \varkappa(\tau)} \ln \left( \frac{M_f D \omega^{-1}(\Delta_f(x_0))}{\omega(\frac{(1-\beta)(1-2\beta)}{2})} \right) \right]^{1/2} \tag{5.2.22}$$

*and we have $y_N \in \mathscr{Q}_f$.*

*Proof* Indeed,

$$\begin{aligned} f(y(t_k)) - f^* &\leq \langle \nabla f(y(t_k)), y(t_k) - x^* \rangle \overset{(5.2.12)}{=} t_k \langle \nabla f(y_0), y(t_k) - x^* \rangle \\ &\leq t_k \lambda_f(y_0) D. \end{aligned}$$

Note that $\omega(M_f \lambda_f(y_0)) \overset{(5.1.29)}{\leq} M_f^2(f(y_0) - f^*) = \Delta_f(y_0)$. Thus,

$$\frac{1}{M_f^2} \omega(M_f \lambda_f(y(t_k))) \overset{(5.1.29)}{\leq} f(y(t_k)) - f^* \leq \frac{t_k}{M_f} \omega^{-1}(\Delta_f(y_0)) D.$$

Since $\|\nabla f(y_k) - \nabla f(y(t_k))\|^*_{y_k} \overset{(5.2.12)}{=} \|\nabla f(y_k) - t_k \nabla f(y_0)\|^*_{y_k} \leq \frac{\beta}{M_f}$, we have

$$\begin{aligned} \lambda_f(y_k) &\overset{(5.2.17)}{\leq} t_k \|\nabla f(y_0)\|^*_{y_k} + \frac{\beta}{M_f} = \langle \nabla f(y(t_k)), [\nabla^2 f(y_k)]^{-1} \nabla f(y(t_k)) \rangle^{1/2} \\ &\quad + \frac{\beta}{M_f} \\ &\overset{(5.1.36)}{\leq} \frac{1}{1-\beta} \lambda_f(y(t_k)) + \frac{\beta}{M_f}. \end{aligned}$$

Thus, inclusion $y_k \in \mathscr{Q}_f$ is ensured by the inequality $\lambda_f(y(t_k)) \leq \frac{(1-\beta)(1-2\beta)}{2M_f}$. Consequently, we need to ensure the inequality

$$\frac{t_k}{M_f}\omega^{-1}(\Delta_f(x_0))D \leq \frac{1}{M_f^2}\omega\left(\frac{(1-\beta)(1-2\beta)}{2}\right).$$

It remains to use the estimate (5.2.18). □

As we can see from the estimate (5.2.22), up to a logarithmic factor, the number of iterations of the path-following scheme is proportional to $\Delta_f^{1/2}(y_0)$. This is much better than the guarantee (5.2.10) obtained for the Damped Newton's Method (5.1.28). However, as we will see in Sect. 5.2.3, for some special subclasses of self-concordant functions the performance estimate (5.2.22) can be significantly improved.

From the practical point of view, reasonable values of parameters for path-following scheme (5.2.16) correspond to $\tau = 0.15$. In this case, $\left[\frac{1}{\gamma(\tau)\varkappa(\tau)}\right]^{1/2} \leq$ 16.1.

*Remark 5.2.1* The dual interpretation of the central path (5.2.12) is quite straightforward: it is just a straight line. We follow the primal image of the *dual* central path

$$s(t) = t\nabla f(y_0) \in \operatorname{dom} f_*, \quad 0 \leq t \leq 1,$$

by generating points $s_k = \nabla f(y_k)$ in a small neighborhood of this trajectory:

$$\langle s_k - s(t_k), \nabla^2 f_*(s_k)(s_k - s(t_k))\rangle \overset{(5.2.13)}{\leq} \frac{\beta^2}{M_f^2}. \qquad \square$$

## 5.2.3 Minimizing Strongly Convex Functions

Let $B = B^* \succ 0$ map $\mathbb{E}$ to $\mathbb{E}^*$. Define the Euclidean metric

$$\|x\|^2 = \langle Bx, x\rangle^{1/2}, \quad x \in \mathbb{E}.$$

In this section, we consider the following minimization problem

$$\min_{x\in\mathbb{E}} f(x), \tag{5.2.23}$$

where $f$ is a strongly convex function:

$$f(y) \geq f(x) + \langle\nabla f(x), y-x\rangle + \tfrac{1}{2}\sigma_2(f)\|y-x\|^2, \quad x, y \in \mathbb{E}, \tag{5.2.24}$$

where $\sigma_2(f) > 0$. We also assume that the function $f$ belongs to $\mathbb{C}^3(\mathbb{E})$ and its Hessian is Lipschitz continuous:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L_3(f)\|x-y\|, \quad x, y \in \mathbb{E}. \tag{5.2.25}$$

As we have seen in Example 5.1.1 (6), this function is self-concordant on $\mathbb{E}$ with the constant

$$M_f = \frac{L_3(f)}{2\sigma_2^{3/2}(f)}. \tag{5.2.26}$$

Thus, problem (5.2.23) can be solved by methods (5.1.28) and (5.2.16). The corresponding complexity bounds can be given in terms of the complexity measure

$$\Delta_f(x_0) = \frac{L_3(f)}{2\sigma_2^{3/2}(f)}(f(x_0) - f^*).$$

As we have seen, the first method needs $O(\Delta_f(x_0))$ iterations. The complexity bound for the second scheme is of the order $\tilde{O}(\Delta_f^{1/2}(x_0))$, where $\tilde{O}(\cdot)$ denotes the hidden logarithmic factors. Let us show that for our particular subclass of self-concordant functions these bounds can be significantly improved.

We will do this by the second-order methods based on cubic regularization of the Newton's Method (see Sect. 4.2). In view of (4.2.60), the region of quadratic convergence of the Cubic Newton's Method (4.2.33) in terms of function value is defined as

$$\mathbb{Q}_f = \left\{ x \in \mathbb{E} : \ f(x) - f^* \le \frac{\sigma_2^3(f)}{2L_3^2(f)} = \frac{1}{8M_f^2} \right\}.$$

Let us check how many iterations we need to enter this region by different schemes based on the cubic Newton step.

Assume our method has the following rate of convergence:

$$f(x_k) - f^* \le \frac{cL_3(f)D^3}{k^p},$$

where $c$ is an absolute constant, $p > 0$, and $D = \max\limits_{x \in \mathbb{E}}\{\|x - x^*\| : \ f(x) \le f(x^0)\}$. Since $f$ is strongly convex, for all $x$ with $f(x) \le f(x_0)$ we have

$$\tfrac{1}{2}\sigma_2(f)\|x - x^*\|^2 \overset{(5.2.24)}{\le} f(x) - f^* \ \le \ f(x_0) - f^*.$$

Therefore,

$$\begin{aligned} f(x_k) - f^* &\leq \tfrac{cL_3(f)}{k^p}\left(\tfrac{2}{\sigma_2(f)}(f(x_0) - f^*)\right)^{3/2} \\ &\overset{(5.2.26)}{=} \tfrac{2^{5/2}cM_f}{k^p}(f(x_0) - f^*)^{3/2}. \end{aligned} \tag{5.2.27}$$

Thus, we need $O\left(\left[M_f^3(f(x_0) - f^*)^{3/2}\right]^{1/p}\right) = O\left(\Delta_f^{\frac{3}{2p}}(x_0)\right)$ iterations to enter the region of quadratic convergence $\mathbb{Q}_f$. For the Cubic Newton's method (4.2.33) we have $p = 2$. Thus, it ensures complexity $O(\Delta_f^{3/4}(x_0))$. For the accelerated Cubic Newton's method (4.2.46) we have $p = 3$. Thus, it needs $O(\Delta^{1/2}(x_0))$ iterations (which is slightly better than (5.2.22)). However, note that for these methods there exists a powerful acceleration tool based on a *restarting procedure*.

Let us define $k_p$ as the first integer for which the right-hand side of inequality (5.2.27) is smaller than $\frac{1}{2}(f(x_0) - f^*)$:

$$\tfrac{2^{5/2}cM_f}{k^p}(f(x_0) - f^*)^{3/2} \leq \tfrac{1}{2}(f(x_0) - f^*).$$

Clearly $k_p = O\left(\left[M_f(f(x_0) - f^*)^{1/2}\right]^{1/p}\right) = O\left(\Delta_f^{\frac{1}{2p}}(x_0)\right)$. This value can be used in the following multi-stage scheme.

**Multi-stage Acceleration Scheme**

Set $y_0 = x_0$

At the $k$th stage ($k \geq 1$) the method starts from the point $y_{k-1}$.

After $t_k = \left\lceil \frac{k_p}{2^{(k-1)/(2p)}} \right\rceil$ steps it generates the output $y_k$.

The method stops when $y_k \in \mathbb{Q}_f$.

(5.2.28)

**Theorem 5.2.5** *The total number of stages $T$ in the optimizations strategy (5.2.28) satisfies the inequality*

$$T \leq 4 + \log_2 \Delta_f(x_0). \tag{5.2.29}$$

*The total number of lower-level iterations $N$ in this scheme does not exceed*

$$4 + \log_2 \Delta_f(x_0) + \tfrac{2^{1/(2p)}}{2^{1/(2p)}-1}k_p.$$

*Proof* Let us prove by induction that $f(y_k) - f^* \le (\frac{1}{2})^k (f(y_0) - f^*)$. For $k = 0$ this is true. Assume that this is also true for some $k \ge 0$. Note that $t_{k+1}^p \ge (\frac{1}{2})^{k/2} k_p^p$. Therefore,

$$\frac{f(y_{k+1})-f^*}{f(y_k)-f^*} \le \frac{2^{5/2}cM_f}{t_{k+1}^p}(f(y_k) - f^*)^{1/2} \le \frac{k_p^p(f(y_k)-f^*)^{1/2}}{2t_{k+1}^p(f(x_0)-f^*)^{1/2}}$$

$$\le \frac{1}{2}\left[\frac{2^k(f(y_k)-f^*)}{f(x_0)-f^*}\right]^{1/2} \le \frac{1}{2}.$$

Thus, the total number of stages satisfies inequality $\left(\frac{1}{2}\right)^{T-1}(f(x_0) - f^*) \ge \frac{1}{8M_f^2}$. Finally,

$$N = \sum_{k=1}^{T} t_k \le T + k_p \sum_{k=0}^{T-1}\left(\frac{1}{2}\right)^{\frac{k}{2p}} \le T + k_p \sum_{k=0}^{\infty}\left(\frac{1}{2}\right)^{\frac{k}{2p}}$$

$$= T + \frac{k_p}{1 - \left(\frac{1}{2}\right)^{1/(2p)}}. \qquad \square$$

Applying Theorem 5.2.5 to different second-order methods based on Cubic Regularization, we get the following complexity bounds.

- **Cubic Newton's Method (4.2.33).** For this method $p = 2$. Therefore, the complexity bound of this scheme, used in the framework of multi-stage method (5.2.28), is of the order

  $$O\left(\Delta_f^{1/4}(x_0)\right).$$

  In fact, this method does not need a restarting strategy. Thus, Theorem 5.2.5 provides the Cubic Newton method with a better way of estimating its rate of convergence.
- **Accelerated Newton's Method (4.2.46).** For this method $p = 3$. Hence, the complexity bound of the corresponding multi-stage scheme (5.2.28) becomes

  $$O\left(\Delta^{1/6}(x_0)\right).$$

- **Optimal second-order method** (see Sect. 4.3.2). For this method $p = 3.5$. Therefore, the corresponding complexity bound is

  $$\tilde{O}\left(\Delta^{1/7}(x_0)\right).$$

  However, note that this method includes an expensive line-search procedure. Consequently, its practical efficiency should be worse that the efficiency of the method from the previous item. Note that the theoretical gap in the complexity estimates of these methods is negligibly small, of the order of $O\left(\varDelta_f^{1/42}(x_0)\right)$. For all reasonable values of the complexity measure $\varDelta_f(x_0)$, feasible for modern computers, it should be much smaller than the logarithmic factors coming from the line search.

## 5.3 Self-concordant Barriers

(Motivation; Definition of self-concordant barriers; Barriers related to self-concordant functions; The implicit barrier theorem; Main properties; Standard minimization problems; The central path; The path-following method; How to initialize the process? Problems with functional constraints.)

### *5.3.1 Motivation*

In the previous section, we have seen that the Newton's Method is very efficient in minimizing *self-concordant* functions. Such a function is always a barrier for its domain. Let us check what can be proved about the Sequential Unconstrained Minimization approach (Sect. 1.3.3) based on these barriers. From now on, we are always working with *standard* self-concordant functions, which means that

$$M_f = 1. \tag{5.3.1}$$

In what follows, we deal with constrained minimization problems of a special type. Let $\mathrm{Dom}\, f = \mathrm{cl}\,(\mathrm{dom}\, f)$.

**Definition 5.3.1** A constrained minimization problem is called *standard* if it has the following form:

$$\min\{\langle c, x\rangle \mid x \in Q\}, \tag{5.3.2}$$

where $Q$ is a closed convex set. It is also assumed that we know a *standard* self-concordant function $f$ such that $\mathrm{Dom}\, f = Q$.

Note that the assumption $M_f = 1$ is not binding since otherwise we can multiply $f$ by an appropriate constant (see Corollary 5.1.3).

Let us introduce a parametric family of penalty functions

$$f(t; x) = t\langle c, x\rangle + f(x)$$

with $t \geq 0$. Note that $f(t;x)$ is self-concordant in $x$ (see Corollary 5.1.2). Define

$$x^*(t) = \arg \min_{x \in \mathrm{dom}\, f} f(t;x).$$

This trajectory is called the *central path* of problem (5.3.2). We can expect that $x^*(t) \to x^*$ as $t \to \infty$ (see Sect. 1.3.3). Therefore, it should be a good idea to keep our test points close to this trajectory.

Recall that the Newton's Methods, as applied to the minimization of the function $f(t;\cdot)$, have local quadratic convergence (Theorem 5.2.2). Our subsequent analysis is based on the Intermediate Newton Method $(5.2.1)_C$, which has the following region of quadratic convergence:

$$\lambda_{f(t;\cdot)}(x) \leq \beta < \tfrac{1}{2}.$$

Let us study our possibilities to move forward in $t$, assuming that we know exactly $x = x^*(t)$ for some $t > 0$.

Thus, we are going to increase $t$:

$$t_+ = t + \Delta, \quad \Delta > 0.$$

However, we need to keep $x$ in the region of quadratic convergence of Newton's Method for the function $f(t+\Delta;\cdot)$:

$$\lambda_{f(t+\Delta;\cdot)}(x) \leq \beta \; < \; \tfrac{1}{2}.$$

Note that the update $t \to t_+$ does not change the Hessian of the barrier function:

$$\nabla^2 f(t+\Delta;x) = \nabla^2 f(t;x).$$

Therefore, it is easy to estimate how big the step $\Delta$ can be. Indeed, the first-order optimality condition (1.2.4) provides us with the following *central path equation:*

$$tc + \nabla f(x^*(t)) = 0. \tag{5.3.3}$$

Since $tc + \nabla f(x) = 0$, we obtain

$$\lambda_{f(t+\Delta;\cdot)}(x) = \| \, t_+ c + \nabla f(x) \, \|_x^* \overset{(5.3.3)}{=} \Delta \, \| \, c \, \|_x^* = \tfrac{\Delta}{t} \, \| \, \nabla f(x) \, \|_x^* \leq \; \beta.$$

Hence, if we want to increase $t$ at some *linear rate*, we need to assume that the value

$$\lambda_f^2(x) \; = \; (\| \, \nabla f(x) \, \|_x^*)^2 \; \equiv \; \langle \nabla f(x), [\nabla^2 f(x)]^{-1} \nabla f(x) \rangle$$

is *uniformly bounded* on $\mathrm{dom}\, f$. Without this assumption, we can have only a sublinear rate of convergence of the process (see Sect. 5.2.2).

Thus, we come to a definition of a *self-concordant barrier*.

### 5.3.2 Definition of a Self-concordant Barrier

**Definition 5.3.2** Let $F(\cdot)$ be a standard self-concordant function. We call it a $\nu$-*self-concordant barrier* for the set Dom $F$, if

$$\sup_{u\in\mathbb{E}} [2\langle \nabla F(x), u\rangle - \langle \nabla^2 F(x)u, u\rangle] \leq \nu \tag{5.3.4}$$

for all $x \in \operatorname{dom} F$. The value $\nu$ is called the *parameter* of the barrier.

Note that we do not assume $\nabla^2 F(x)$ to be nondegenerate. However, if this is the case, then inequality (5.3.4) is equivalent to the following:

$$\langle \nabla F(x), [\nabla^2 F(x)]^{-1} \nabla F(x)\rangle \leq \nu. \tag{5.3.5}$$

We will also use another equivalent form of inequality (5.3.4):

$$\langle \nabla F(x), u\rangle^2 \leq \nu \langle \nabla^2 F(x)u, u\rangle \quad \forall u \in \mathbb{E}. \tag{5.3.6}$$

(To see this for $u$ with $\langle \nabla^2 F(x)u, u\rangle > 0$, replace $u$ in (5.3.4) by $\tau u$ and find the maximum of the left-hand side in $\tau$.) Note that the condition (5.3.6) can be rewritten in matrix notation:

$$\nabla^2 F(x) \succeq \tfrac{1}{\nu} \nabla F(x) \nabla F(x)^T. \tag{5.3.7}$$

**Lemma 5.3.1** *Let $F$ be a $\nu$-self-concordant barrier. Then for any $p \geq \nu$ the function $\xi_p(x) = \exp\left\{-\frac{1}{p} F(x)\right\}$ is concave on* dom $F$*. On the other hand, if function $\xi_\nu(\cdot)$ is concave on* dom $F$*, then $F$ is a self-concordant barrier.*

*Proof* Indeed, for any $x \in \operatorname{dom} F$ and $h \in \mathbb{E}$, we have

$$\langle \nabla \xi_p(x), h\rangle = -\tfrac{1}{p} \langle \nabla F(x), h\rangle \xi_p(x),$$

$$\langle \nabla^2 \xi_p(x)h, h\rangle = \tfrac{1}{p^2} \langle \nabla F(x), h\rangle^2 \xi_p(x) - \tfrac{1}{p} \langle \nabla^2 F(x)h, h\rangle \xi_p(x).$$

It remains to use definition (5.3.6). □

Note that condition (5.3.5) has interesting dual interpretation. In view of relation (5.1.34), definition (5.3.5) is equivalent to the following condition:

$$\langle s, \nabla^2 F_*(s)s\rangle \leq \nu, \quad s \in \operatorname{dom} F_*. \tag{5.3.8}$$

In other words, at any feasible $s$, the distance to the origin is proportional to the size of the unit Dikin ellipsoid, which describes an ellipsoidal neighborhood in dom $f_*$ with similar Hessians.

Let us now check which self-concordant functions presented in Example 5.1.1 are also self-concordant barriers.

*Example 5.3.1*

1. *Linear function*: $f(x) = \alpha + \langle a, x\rangle$, $\mathrm{dom}\, f = \mathbb{E}$. Clearly, for $a \neq 0$ this function is not a self-concordant barrier since $\nabla^2 F(x) = 0$.
2. *Convex quadratic function*. Let $A = A^T \succ 0$. Consider the function
$$f(x) = \alpha + \langle a, x\rangle + \frac{1}{2}\langle Ax, x\rangle, \quad \mathrm{dom}\, f = \mathbb{R}^n.$$
Then $\nabla f(x) = a + Ax$ and $\nabla^2 f(x) = A$. Therefore,
$$\begin{aligned}\langle [\nabla^2 f(x)]^{-1}\nabla f(x), \nabla f(x)\rangle &= \langle A^{-1}(Ax + a), Ax + a\rangle \\ &= \langle Ax, x\rangle + 2\langle a, x\rangle + \langle A^{-1}a, a\rangle.\end{aligned}$$
Clearly, this value is unbounded from above on $\mathbb{R}^n$. Thus, a quadratic function is not a self-concordant barrier.
3. *Logarithmic barrier for a ray*. Consider the following function of one variable:
$$F(x) = -\ln x, \quad \mathrm{dom}\, F = \{x \in \mathbb{R} \mid x > 0\}.$$
Then $\nabla F(x) = -\frac{1}{x}$ and $\nabla^2 F(x) = \frac{1}{x^2} > 0$. Therefore
$$\frac{(\nabla F(x))^2}{\nabla^2 F(x)} = \frac{1}{x^2} \cdot x^2 = 1.$$
Thus, $F(\cdot)$ is a $\nu$-self-concordant barrier for the set $\{x \geq 0\}$ with $\nu = 1$.
4. *Logarithmic barrier for a second-order region*. Let $A = A^T \succeq 0$. Consider the *concave* quadratic function
$$\phi(x) = \alpha + \langle a, x\rangle - \frac{1}{2}\langle Ax, x\rangle.$$
Define $F(x) = -\ln \phi(x)$, $\mathrm{dom}\, f = \{x \in \mathbb{R}^n \mid \phi(x) > 0\}$. Then
$$\langle \nabla F(x), u\rangle = -\frac{1}{\phi(x)}[\langle a, u\rangle - \langle Ax, u\rangle],$$
$$\langle \nabla^2 F(x)u, u\rangle = \frac{1}{\phi^2(x)}[\langle a, u\rangle - \langle Ax, u\rangle]^2 + \frac{1}{\phi(x)}\langle Au, u\rangle.$$
Let $\omega_1 = \langle \nabla F(x), u\rangle$ and $\omega_2 = \frac{1}{\phi(x)}\langle Au, u\rangle$. Then
$$\langle \nabla^2 F(x)u, u\rangle = \omega_1^2 + \omega_2 \geq \omega_1^2.$$

Therefore $2\langle \nabla F(x), u\rangle - \langle \nabla^2 F(x)u, u\rangle \le 2\omega_1 - \omega_1^2 \le 1$. Thus, $F(\cdot)$ is a $\nu$-self-concordant barrier with $\nu = 1$. □

Let us now check the results of some simple operations with self-concordant barriers.

**Theorem 5.3.1** *Let $F(\cdot)$ be a self-concordant barrier. Then the function $\langle c, x\rangle + F(x)$ is a standard self-concordant function on dom $F$.*

*Proof* Since $F(\cdot)$ is a self-concordant function, we just apply Corollary 5.1.2. □

Note that this property is important for path-following schemes.

**Theorem 5.3.2** *Let $F_i$ be $\nu_i$-self-concordant barriers, $i = 1, 2$. Then the function*

$$F(x) = F_1(x) + F_2(x)$$

*is a self-concordant barrier for a convex set Dom $F$ = Dom $F_1 \bigcap$ Dom $F_2$ with the parameter $\nu = \nu_1 + \nu_2$.*

*Proof* In view of Theorem 5.1.1, $F$ is a standard self-concordant function. Let us fix $x \in \operatorname{dom} F$. Then

$$\begin{aligned}
&\max_{u\in\mathbb{R}^n} [2\langle \nabla F(x), u\rangle - \langle \nabla^2 F(x)u, u\rangle]\\
&= \max_{u\in\mathbb{R}^n} [2\langle \nabla F_1(x), u\rangle - \langle \nabla^2 F_1(x)u, u\rangle + 2\langle \nabla F_2(x), u\rangle - \langle \nabla^2 F_2(x)u, u\rangle]\\
&\le \max_{u\in\mathbb{R}^n} [2\langle \nabla F_1(x), u\rangle - \langle \nabla^2 F_1(x)u, u\rangle] + \max_{u\in\mathbb{R}^n} [2\langle \nabla F_2(x), u\rangle - \langle \nabla^2 F_2(x)u, u\rangle]\\
&\le \nu_1 + \nu_2.
\end{aligned}$$

□

It is easy to see that the value of the parameter of a self-concordant barrier is invariant with respect to an affine transformation of variables.

**Theorem 5.3.3** *Let $\mathscr{A}(x) = Ax + b$ be a linear operator, $\mathscr{A} : \mathbb{E} \to \mathbb{E}_1$. Assume that function $F$ is a $\nu$-self-concordant barrier with Dom $F \subset \mathbb{E}_1$. Then the function*

$$\Phi(x) = F(\mathscr{A}(x))$$

*is a $\nu$-self-concordant barrier for the set Dom $\Phi = \{x \in \mathbb{E} :\ \mathscr{A}(x) \in$ Dom $F\}$.*

*Proof* The function $\Phi(\cdot)$ is a standard self-concordant function in view of Theorem 5.1.2. Let us fix $x \in \operatorname{dom}\Phi$. Then $y = \mathscr{A}(x) \in \operatorname{dom} F$. Note that for any $u \in \mathbb{E}$ we have

$$\langle \nabla\Phi(x), u\rangle = \langle \nabla F(y), Au\rangle, \quad \langle \nabla^2\Phi(x)u, u\rangle = \langle \nabla^2 F(y)Au, Au\rangle.$$

Therefore

$$\max_{u\in\mathbb{E}} [2\langle\nabla\Phi(x), u\rangle - \langle\nabla^2\Phi(x)u, u\rangle] = \max_{u\in\mathbb{E}} [2\langle\nabla F(y), Au\rangle - \langle\nabla^2 F(y)Au, Au\rangle]$$

$$\leq \max_{w\in\mathbb{E}_1} [2\langle\nabla F(y), w\rangle - \langle\nabla^2 F(y)w, w\rangle] \leq \nu. \qquad \square$$

To conclude this section, let us show how to construct self-concordant barriers for the level sets of self-concordant functions and for the epigraphs of self-concordant barriers.

**Theorem 5.3.4** *Let the function $f$ be self-concordant with constant $M_f \geq 0$. Suppose that the set*

$$\mathscr{L}(\beta) = \{x \in dom\, f :\ f(x) \leq \beta\}$$

*has nonempty interior and $f(x) \geq f^*$ for all $x \in dom\, f$. Then the function*

$$F(x) = -\nu \ln(\beta - f(x))$$

*with any $\nu \geq 1 + M_f^2(\beta - f^*)$ is a $\nu$-self-concordant barrier for the level set $\mathscr{L}(\beta)$.*

*Proof* Let $\phi(x) = -\ln(\beta - f(x))$. In view of Theorem 5.1.4 and Corollary 5.1.3, the function $F(x) = \nu\phi(x)$ is a standard self-concordant function on dom $f$. On the other hand, for any $h \in \mathbb{E}$ we have

$$\langle\nabla F(x), h\rangle^2 = \nu^2\langle\nabla\phi(x), h\rangle^2 \overset{(5.1.8)}{\leq} \nu^2\langle\nabla^2\phi(x)h, h\rangle = \nu\langle\nabla^2 F(x)h, h\rangle.$$

Thus, by definition (5.3.6), $F$ is a $\nu$-self-concordant barrier for $\mathscr{L}(\beta)$. □

**Theorem 5.3.5** *Let $f$ be a $\nu$-self-concordant barrier. Then the function*

$$F(x, t) = f(x) - \ln(t - f(x))$$

*is a $(\nu + 1)$-self-concordant barrier for the epigraph*

$$\mathscr{E}_f = \{(x, t) \in dom\, f \times \mathbb{R} :\ t \geq f(x)\}.$$

*Proof* Let us fix a direction $h \in \mathbb{E}$ and $\delta \in \mathbb{R}$. Consider the function

$$\phi(\tau) = F(x + \tau h, t + \tau\delta) = f(x + \tau h) - \ln(t + \tau\delta - f(x + \tau h).$$

Let $\omega = t - f(x)$ and $\hat{\omega} = 1 + \frac{1}{\omega}$. Then

$$\phi'(0) = \langle \nabla f(x), h \rangle + \tfrac{1}{\omega}(\langle \nabla f(x), h \rangle - \delta),$$

$$\begin{aligned}\phi''(0) &= \langle \nabla^2 f(x)h, h \rangle + \tfrac{1}{\omega^2}(\langle \nabla f(x), h \rangle - \delta)^2 + \tfrac{1}{\omega}\langle \nabla^2 f(x)h, h \rangle \\ &= \hat{\omega}\langle \nabla^2 f(x)h, h \rangle + \tfrac{1}{\omega^2}(\langle \nabla f(x), h \rangle - \delta)^2.\end{aligned}$$

Define $\xi = \left[\hat{\omega}\langle \nabla^2 f(x)h, h \rangle\right]^{1/2}$ and $\lambda = \frac{1}{\omega}(\langle \nabla f(x), h \rangle - \delta)$. Note that

$$\phi'(0) \overset{(5.3.6)}{\leq} \sqrt{\nu}\langle \nabla^2 f(x)h, h \rangle^{1/2} + \lambda \; = \; \xi\sqrt{\tfrac{\nu}{\hat{\omega}}} + \lambda.$$

It remains to note that the maximum of the right-hand side of this inequality subject to the constraint $\xi^2 + \lambda^2 = 1$ is equal to $\left[\frac{\nu}{\hat{\omega}} + 1\right]^{1/2} \leq \sqrt{\nu + 1}$. Thus, in view of definition (5.3.6), the parameter of the barrier $F$ can be chosen as $\nu + 1$.

Let us estimate now the third derivative of the function $\phi$ at zero, assuming that its second derivative is less or equal to one. Note that

$$\begin{aligned}\phi'''(0) \;&=\; D^3 f(x)[h,h,h] + \tfrac{2}{\omega^3}(\langle \nabla f(x), h \rangle - \delta)^3 \\ &\quad + \tfrac{3}{\omega^2}(\langle \nabla f(x), h \rangle - \delta)\langle \nabla^2 f(x)h, h \rangle + \tfrac{1}{\omega} D^3 f(x)[h,h,h] \\ &\overset{(5.1.4)}{\leq}\; 2\hat{\omega}\langle \nabla^2 f(x)h, h \rangle^{3/2} + \tfrac{2}{\omega^3}(\langle \nabla f(x), h \rangle - \delta)^3 \\ &\quad + \tfrac{3}{\omega^2}(\langle \nabla f(x), h \rangle - \delta)\langle \nabla^2 f(x)h, h \rangle \\ &=\; 2\sqrt{\tfrac{\omega}{1+\omega}}\xi^3 + 2\lambda^3 + \tfrac{3}{1+\omega}\xi^2\lambda \;=\; 2\gamma\xi^3 + 2\lambda^3 + 3(1-\gamma^2)\xi^2\lambda,\end{aligned}$$

where $\gamma^2 = \frac{\omega}{1+\omega}$. We need to maximize the right-hand side of the above inequality subject to constraints $\xi^2 + \lambda^2 \leq 1$ and $\gamma \in [0, 1]$:

$$\varkappa^* = \max_{\gamma,\lambda,\xi}\{2\gamma\xi^3 + 2\lambda^3 + 3(1-\gamma^2)\xi^2\lambda : \; \xi^2 + \lambda^2 \leq 1, \; 0 \leq \gamma \leq 1\}.$$

Let us maximize this objective in $\gamma$. From the first-order optimality condition for $\gamma$,

$$2\xi^3 - 6\gamma\xi^2\lambda = 0,$$

we have $\gamma_* = \min\left\{1, \frac{\xi}{3\lambda}\right\}$. Assume that $\xi \geq 3\lambda$. Then $\gamma_* = 1$ and we need to maximize $2\xi^3 + 2\lambda^3$ with constraints $\xi^2 + \lambda^2 = 1$ and $\xi \geq 3\lambda$. Introducing new variables $\hat{\xi} = \xi^2$ and $\hat{\lambda} = \lambda^2$, we come to the problem

$$\max_{\hat{\xi},\hat{\lambda}\geq 0} \{2\hat{\xi}^{3/2} + 2\hat{\lambda}^{3/2} : \ \hat{\xi} + \hat{\lambda} \leq 1, \ \hat{\xi} \geq 9\hat{\lambda}\}.$$

Its objective is convex. Hence, by inspecting the extreme points of its feasible set we find the optimal solution $\hat{\xi}_* = 1$, $\hat{\lambda}_* = 0$. Thus, the maximal value of this problem is two.

Assume now that $\xi \leq 3\lambda$. Then $\gamma_* = \frac{\xi}{3\lambda}$ and we get the following objective:

$$2\tfrac{\xi}{3\lambda}\xi^3 + 2\lambda^3 + 3\left(1 - \tfrac{\xi^2}{9\lambda^2}\right)\xi^2\lambda = \tfrac{\xi^4}{3\lambda} + 2\lambda^3 + 3\xi^2\lambda.$$

Note that the maximum of this expression is attained at the boundary of the unit circle: $\xi^2 + \lambda^2 = 1$. Thus, we need to show that

$$\tfrac{(1-\lambda^2)^2}{3\lambda} + 2\lambda^3 + 3(1-\lambda^2)\lambda \leq 2,$$

with constraint $3\lambda \geq \sqrt{1-\lambda^2}$. In other words, we need to prove that

$$p(\lambda) \overset{\text{def}}{=} (1-\lambda^2)^2 + 3\lambda(3\lambda - \lambda^3) - 6\lambda \ \leq \ 0, \quad \tfrac{1}{\sqrt{10}} \leq \lambda \leq 1.$$

Note that $p(\lambda) = (1-\lambda)^2(3 - 2(1+\lambda)^2) \leq 0$ for all $\lambda \geq \sqrt{\frac{3}{2}} - 1 = \frac{1}{2+\sqrt{6}}$, and this constant is smaller than our lower bound for $\lambda$: $\frac{1}{\sqrt{10}} > \frac{1}{2+\sqrt{6}}$.

Thus, $\varkappa^* \leq 2$, which means that $F$ is a standard self-concordant function. □

**Corollary 5.3.1** *If $f$ is a standard self-concordant function, then $F$ is also a standard self-concordant function with Dom $F = \mathscr{E}_f$.*

Finally, let us prove the Implicit Barrier Theorem. Let $\Phi$ be a $\nu$-self-concordant barrier for dom $\Phi \subset \mathbb{E}$. We partition the space as follows: $\mathbb{E} = \mathbb{E}_1 \times \mathbb{E}_2$. Define

$$F(x) = \min_y \{\Phi(x, y) : \ (x, y) \in \text{dom}\, \Phi\}. \tag{5.3.9}$$

We assume that for any $x \in \text{dom}\, F \subset \mathbb{E}_1$ the solution $y(x)$ of this optimization problem exists and is unique. Then, as we have seen in the proof of Theorem 5.1.11,

$$\nabla_y \Phi(x, y(x)) = 0, \quad \nabla_x \Phi(x, y(x)) \ = \ \nabla F(x).$$

**Theorem 5.3.6** *The function $F$ defined by (5.3.9) is a $\nu$-self-concordant barrier.*

*Proof* In view of Theorem 5.1.11 the function $F$ is standard self-concordant. Let us fix $x \in \mathrm{dom}\, F$. Then for any direction $z = (h, \delta) \in \mathbb{E}_1 \times \mathbb{E}_2$ we have

$$\begin{aligned} \langle \nabla F(x), h \rangle^2_{\mathbb{E}_1} \;&=\; \langle \nabla_x \Phi(x, y(x)), h \rangle^2_{\mathbb{E}_1} \;=\; \langle \nabla \Phi(x, y(x)), z \rangle^2_{\mathbb{E}} \\ &\overset{(5.3.6)}{\leq}\; \nu \langle \nabla^2 \Phi(x, y(x)) z, z \rangle_{\mathbb{E}}. \end{aligned}$$

As was shown in the proof of Theorem 5.1.11,

$$\min_{\delta \in \mathbb{E}_2} \langle \nabla^2 \Phi(x, y(x)) z, z \rangle_{\mathbb{E}} = \langle \nabla^2 F(x) h, h \rangle_{\mathbb{E}_1}.$$

Thus, $F$ satisfies definition (5.3.6) of a $\nu$-self-concordant barrier. □

### 5.3.3 Main Inequalities

Let us show that the local characteristics of a self-concordant barrier (gradient and Hessian) provide us with *global* information about the structure of its domain.

**Theorem 5.3.7** *1. Let $F$ be a $\nu$-self-concordant barrier. For any $x$ and $y$ from dom $F$, we have*

$$\langle \nabla F(x), y - x \rangle \;<\; \nu. \tag{5.3.10}$$

*Moreover, if $\langle \nabla F(x), y - x \rangle \geq 0$, then*

$$\langle \nabla F(y) - \nabla F(x), y - x \rangle \geq \tfrac{\langle \nabla F(x), y - x \rangle^2}{\nu - \langle \nabla F(x), y - x \rangle}. \tag{5.3.11}$$

*2. A standard self-concordant function $F$ is a $\nu$-self-concordant barrier if and only if*

$$F(y) \geq F(x) - \nu \ln\left(1 - \tfrac{1}{\nu} \langle \nabla F(x), y - x \rangle\right) \quad \forall x, y \in \mathit{dom}\, F. \tag{5.3.12}$$

*Proof* 1. Let us fix two points $x, y \in \mathrm{dom}\, F$. Consider the univariate function

$$\phi(t) = \langle \nabla F(x + t(y - x)), y - x \rangle, \quad t \in [0, 1].$$

If $\phi(0) \leq 0$, then (5.3.10) is trivial. If $\phi(0) = 0$, then (5.3.11) is valid in view of convexity of $f$. Suppose that $\phi(0) > 0$. In view of inequality (5.3.6), we have

$$\begin{aligned} \phi'(t) &= \langle \nabla^2 F(x + t(y - x))(y - x), y - x \rangle \\ &\geq \tfrac{1}{\nu} \langle \nabla F(x + t(y - x)), y - x \rangle^2 \;=\; \tfrac{1}{\nu} \phi^2(t). \end{aligned}$$

Therefore, $\phi(t)$ increases and is positive for $t \in [0,1]$. Moreover, for any $t \in [0,1]$ we have

$$-\frac{1}{\phi(t)} + \frac{1}{\phi(0)} = \int\limits_0^t \frac{\phi'(\tau)}{\phi^2(\tau)}\,d\tau \overset{(5.3.6)}{\geq} \frac{1}{\nu}t.$$

This implies that $\langle \nabla F(x), y-x\rangle = \phi(0) < \frac{\nu}{t}$ for all $t \in [0,1]$. Thus, (5.3.10) is proved. At the same time,

$$\phi(t) - \phi(0) \geq \frac{\nu\phi(0)}{\nu - t\phi(0)} - \phi(0) = \frac{t\phi(0)^2}{\nu - t\phi(0)}, \quad t \in [0,1].$$

Choosing $t = 1$, we get inequality (5.3.11).

2. Let $\psi(x) = e^{-\frac{1}{\nu}F(x)}$. In view of Lemma 5.3.1, this function is concave. It remains to note that inequality (5.3.12) is equivalent to the condition

$$\psi(y) \leq \psi(x) + \langle \nabla\psi(x), y-x\rangle$$

up to a logarithmic transformation of both sides. □

**Corollary 5.3.2** *Let $F$ be a $\nu$-self-concordant barrier and $h \in \mathbb{E}$ be a recession direction of dom $F$: $x + \tau h \in$ dom $F$ for any $x \in$ dom $F$ and $\tau \geq 0$. Then,*

$$\langle \nabla^2 F(x)h, h\rangle^{1/2} \leq \langle -\nabla F(x), h\rangle. \tag{5.3.13}$$

*Proof* In view of inequality (5.3.10), $\langle \nabla F(x), h\rangle \leq 0$. If dom $F$ does not contain the line $\{x + \tau h,\ \tau \in \mathbb{R}\}$, then inequality (5.3.13) follows from (5.1.27). If it contains the line, then $\langle \nabla F(x), h\rangle = 0$ for all $x \in$ dom $F$. This means that $F$ is constant along this line and both sides of inequality (5.3.13) vanish. □

**Corollary 5.3.3** *Let $x, y \in$ dom $F$. Then for any $\alpha \in [0,1)$ we have*

$$F(x + \alpha(y-x)) \leq F(x) - \nu \ln(1-\alpha). \tag{5.3.14}$$

*Proof* Let $y(t) = x + t(y-x)$ and $\phi(t) = F(y(t))$. Then

$$\phi'(t) = \langle \nabla F(y(t)), y-x\rangle = \frac{1}{1-t}\langle \nabla F(y(t)), y - y(\alpha)\rangle \overset{(5.3.10)}{\leq} \frac{\nu}{1-t}.$$

Integrating this inequality in $t \in [0,\alpha)$, we get inequality (5.3.14). □

**Theorem 5.3.8** *Let $F$ be a $\nu$-self-concordant barrier. Then for any $x \in$ dom $F$ and $y \in$ Dom $F$ such that*

$$\langle \nabla F(x), y-x\rangle \geq 0, \tag{5.3.15}$$

*we have*

$$\| y - x \|_x \leq \nu + 2\sqrt{\nu}. \tag{5.3.16}$$

*Proof* Let $r = \| y - x \|_x$ and suppose $r > \sqrt{\nu}$ (otherwise (5.3.16) is trivial). Consider the point $y_\alpha = x + \alpha(y - x)$ with $\alpha = \frac{\sqrt{\nu}}{r} < 1$. In view of our assumption (5.3.15) and inequality (5.1.13) we have

$$\begin{aligned}
\omega \equiv \langle \nabla F(y_\alpha), y - x \rangle &\geq \langle \nabla F(y_\alpha) - \nabla F(x), y - x \rangle \\
&= \tfrac{1}{\alpha} \langle \nabla F(y_\alpha) - \nabla F(x), y_\alpha - x \rangle \\
&\geq \tfrac{1}{\alpha} \cdot \tfrac{\| y_\alpha - x \|_x^2}{1 + \| y_\alpha - x \|_x} = \tfrac{\alpha \| y - x \|_x^2}{1 + \alpha \| y - x \|_x} = \tfrac{r\sqrt{\nu}}{1 + \sqrt{\nu}}.
\end{aligned}$$

On the other hand, in view of (5.3.10), we obtain

$$(1 - \alpha)\omega = \langle \nabla F(y_\alpha), y - y_\alpha \rangle \leq \nu.$$

Thus,

$$\left(1 - \tfrac{\sqrt{\nu}}{r}\right) \tfrac{r\sqrt{\nu}}{1 + \sqrt{\nu}} \leq \nu,$$

and this is exactly (5.3.16). □

We conclude this section by studying the properties of one special point of a convex set.

**Definition 5.3.3** Let $F$ be a $\nu$-self-concordant barrier for the set Dom $F$. The point

$$x_F^* = \arg \min_{x \in \mathrm{dom}\, F} F(x)$$

is called the *analytic center* of the convex set Dom $F$, generated by the barrier $F$.

**Theorem 5.3.9** *Assume that the analytic center of a $\nu$-self-concordant barrier $F$ exists. Then for any $x \in$ Dom $F$ we have*

$$\| x - x_F^* \|_{x_F^*} \leq \nu + 2\sqrt{\nu}.$$

*On the other hand, for any $x \in \mathbb{R}^n$ such that $\| x - x_F^* \|_{x_F^*} \leq 1$, we have $x \in$ Dom $F$.*

*Proof* The first statement follows from Theorem 5.3.8 since $\nabla F(x_F^*) = 0$. The second statement follows from Theorem 5.1.5. □

Thus, the *asphericity* of the set Dom $F$ with respect to $x_F^*$, computed in the metric $\| \cdot \|_{x_F^*}$, does not exceed $\nu + 2\sqrt{\nu}$. It is well known that for any convex set in $\mathbb{R}^n$ there exists a metric in which the asphericity of this set is less than or equal to $n$

(John's Theorem). However, we managed to estimate the asphericity in terms of the *parameter* of the self-concordant barrier. This value does not depend directly on the dimension of the space of variables.

Recall also that if $\mathrm{Dom}\, F$ contains no straight lines the existence of $x_F^*$ implies the boundedness of $\mathrm{Dom}\, F$ (since then $\nabla^2 F(x_F^*)$ is nondegenerate, see Theorem 5.1.6).

**Corollary 5.3.4** *Let Dom $F$ be bounded. Then for any $x \in$ dom $F$ and $v \in \mathbb{R}^n$ we have*

$$\| v \|_x^* \leq \ (\nu + 2\sqrt{\nu}) \, \| v \|_{x_F^*}^* \, .$$

*In other words, for any $x \in$ dom $F$ we have*

$$\nabla^2 F(x) \succeq \tfrac{1}{(\nu+2\sqrt{\nu})^2} \nabla^2 F(x_F^*). \tag{5.3.17}$$

*Proof* By Lemma 3.1.20, we get the following representation:

$$\| v \|_x^* \equiv \ \langle v, [\nabla^2 F(x)]^{-1} v\rangle^{1/2} \ = \ \max\{\langle v, u\rangle \mid \ \langle \nabla^2 F(x) u, u\rangle \leq 1\}.$$

On the other hand, in view of Theorems 5.1.5 and 5.3.9, we have

$$B \equiv \{y \in \mathbb{R}^n \mid \| y - x \|_x \leq 1\} \ \subseteq \ \mathrm{Dom}\, F$$

$$\subseteq \{y \in \mathbb{R}^n \mid \| y - x_F^* \|_{x_F^*} \leq \nu + 2\sqrt{\nu}\} \ \equiv \ B_*.$$

Therefore, using again Theorem 5.3.9, we get the following relations:

$$\| v \|_x^* = \max\{\langle v, y - x\rangle \mid \ y \in B\} \ \leq \ \max\{\langle v, y - x\rangle \mid \ y \in B_*\}$$

$$= \langle v, x_F^* - x\rangle + (\nu + 2\sqrt{\nu}) \, \| v \|_{x_F^*}^* \, .$$

Note that $\| v \|_x^* = \| -v \|_x^*$. Therefore, we can always ensure $\langle v, x_F^* - x\rangle \leq 0$. □

## 5.3.4 The Path-Following Scheme

Now we are ready to describe a *barrier model* of the minimization problem. This is a *standard* minimization problem

$$\min\{\langle c, x\rangle \mid \ x \in Q\} \tag{5.3.18}$$

where $Q$ is a *bounded* closed convex set with nonempty interior, which is a closure of the domain of some $\nu$-self-concordant barrier $F$.

We are going to solve (5.3.18) by tracing the *central path*:

$$x^*(t) \;=\; \arg\min_{x\in \mathrm{dom}\, F} \;\; f(t;x), \tag{5.3.19}$$

where $f(t;x) = t\langle c, x\rangle + F(x)$ and $t \geq 0$. In view of the first-order optimality condition (1.2.4), any point of the central path satisfies the equation

$$tc + \nabla F(x^*(t)) = 0. \tag{5.3.20}$$

Since the set $Q$ is bounded and $F$ is a closed convex function, the *analytic center* of this set $x^*_F$ exists and it is uniquely defined (see Item 4 of Theorems 3.1.4 and 5.1.6). Moreover, it is a starting point for the central path:

$$x^*(0) \;=\; x^*_F. \tag{5.3.21}$$

In order to follow the central path, we are going to update the points satisfying an *approximate centering condition*:

$$\lambda_{f(t;\cdot)}(x) \;\equiv \parallel f'(t;x) \parallel^*_x = \parallel tc + \nabla F(x) \parallel^*_x \leq \; \beta, \tag{5.3.22}$$

where the *centering parameter* $\beta$ is small enough.

Let us show that this is a reasonable goal.

**Theorem 5.3.10** *For any $t > 0$, we have*

$$\langle c, x^*(t)\rangle - c^* \leq \tfrac{\nu}{t}, \tag{5.3.23}$$

*where $c^*$ is the optimal value of problem (5.3.18). If a point $x$ satisfies the approximate centering condition (5.3.22), then*

$$\langle c, x\rangle - c^* \leq \tfrac{1}{t}\left(\nu + \tfrac{(\beta+\sqrt{\nu})\beta}{1-\beta}\right). \tag{5.3.24}$$

*Proof* Let $x^*$ be a solution to (5.3.18). In view of (5.3.20) and (5.3.10), we have

$$\langle c, x^*(t) - x^*\rangle = \tfrac{1}{t}\langle \nabla F(x^*(t)), x^* - x^*(t)\rangle \;\leq\; \tfrac{\nu}{t}.$$

Further, let $x$ satisfy (5.3.22). Let $\lambda = \lambda_{f(t;\cdot)}(x)$. Then, in view of (5.3.5), Theorem 5.2.1, and (5.3.22), we have

$$\begin{aligned} t\langle c, x - x^*(t)\rangle &= \langle f'(t;x) - \nabla F(x), x - x^*(t)\rangle \;\leq\; (\lambda + \sqrt{\nu}) \parallel x - x^*(t) \parallel_x \\ &\leq (\lambda + \sqrt{\nu})\frac{\lambda}{1-\lambda} \;\leq\; \frac{(\beta+\sqrt{\nu})\beta}{1-\beta}. \qquad \square \end{aligned}$$

Let us analyze now one step of a path-following scheme. It differs from the updating rule (5.2.14) only by the origin of the objective vector.

Assume that $x \in \operatorname{dom} F$. Consider the following iterate:

$$\boxed{\begin{aligned} t_+ &= t + \tfrac{\gamma}{\|c\|_x^*}, \\ x_+ &= x - \tfrac{1}{1+\xi}[\nabla^2 F(x)]^{-1}(t_+ c + \nabla F(x)), \\ &\text{where } \xi = \tfrac{\lambda^2}{1+\lambda} \text{ and } \lambda = \|t_+ c + \nabla F(x)\|_x^*. \end{aligned}} \tag{5.3.25}$$

From Lemma 5.2.2 we know that if $\beta = \beta(\tau) = \tau^2(1 + \tau + \frac{\tau}{1+\tau+\tau^2})$ with $\tau \in [0, \frac{1}{2}]$ and $x$ satisfies approximate centering condition (5.3.22), then for $\gamma$, such that

$$| \gamma | \leq \tau - \tau^2(1 + \tau + \tfrac{\tau}{1+\tau+\tau^2}), \tag{5.3.26}$$

we have again $\| t_+ c + \nabla F(x_+) \|_{x_+}^* \leq \beta$.

Let us prove now that the increase of $t$ in the scheme (5.3.25) is sufficiently large.

**Lemma 5.3.2** *Let $x$ satisfy (5.3.22). Then*

$$\| c \|_x^* \leq \tfrac{1}{t}(\beta + \sqrt{\nu}). \tag{5.3.27}$$

*Proof* Indeed, in view of (5.3.22) and (5.3.5), we have

$$\begin{aligned} t \| c \|_x^* &= \| f'(t; x) - \nabla F(x) \|_x^* \leq \| f'(t; x) \|_x^* + \| \nabla F(x) \|_x^* \\ &\leq \beta + \sqrt{\nu}. \qquad \square \end{aligned}$$

Let us now fix some reasonable values of parameters in method (5.3.25). In the remaining part of this chapter we always assume that

$$\begin{aligned} \tau &= 0.29, \quad \beta = \beta(\tau) \approx 0.126, \\ \gamma &= \tau - \beta(\tau) \approx 0.164 \quad \Rightarrow \quad \gamma^{-1} < 6.11. \end{aligned} \tag{5.3.28}$$

We have proved that it is possible to follow the central path, using the rule (5.3.25). Note that we can either increase or decrease the current value of $t$. The lower estimate for the rate of *increasing* $t$ is

$$t_+ \geq \left(1 + \frac{\gamma}{\beta+\sqrt{\nu}}\right) \cdot t,$$

and the upper estimate for the rate of *decreasing* $t$ is

$$t_+ \leq \left(1 - \frac{\gamma}{\beta+\sqrt{\nu}}\right) \cdot t.$$

Thus, the general scheme for solving the problem (5.3.18) is as follows.

**Main path-following scheme**

**0.** Set $t_0 = 0$. Choose an accuracy $\epsilon > 0$ and $x_0 \in \operatorname{dom} F$ such that

$$\| \nabla F(x_0) \|^*_{x_0} \leq \beta.$$

**1.** $k$th iteration ($k \geq 0$). Set

$$t_{k+1} = t_k + \frac{\gamma}{\|c\|^*_{x_k}},$$

$$x_{k+1} = x_k - \frac{1}{1+\xi_k}[\nabla^2 F(x_k)]^{-1}(t_{k+1}c + \nabla F(x_k)),$$

where $\xi_k = \frac{\lambda_k^2}{1+\lambda_k}$, and $\lambda_k = \|t_{k+1}c + \nabla F(x_k)\|^*_{x_k}$.

**2.** Stop the process if $t_k \geq \frac{1}{\epsilon}\left(\nu + \frac{(\beta+\sqrt{\nu})\beta}{1-\beta}\right)$.

(5.3.29)

Let us derive a complexity bound for the above scheme.

**Theorem 5.3.11** *Method (5.3.29) terminates after $N$ steps at most, where*

$$N \leq O\left(\sqrt{\nu}\ln \frac{\nu\|c\|^*_{x^*_F}}{\epsilon}\right).$$

*Moreover, at the moment of termination we have* $\langle c, x_N\rangle - c^* \leq \epsilon$.

*Proof* Note that $r_0 \equiv \| x_0 - x_F^* \|_{x_0} \leq \frac{\beta}{1-\beta}$ (see Theorem 5.2.1). Therefore, in view of Theorem 5.1.7 we have

$$\frac{\gamma}{t_1} = \| c \|_{x_0}^* \leq \frac{1}{1-r_0} \| c \|_{x_F^*}^* \leq \frac{1-\beta}{1-2\beta} \| c \|_{x_F^*}^* .$$

Thus, $t_k \geq \frac{\gamma(1-2\beta)}{(1-\beta)\|c\|_{x_F^*}^*} \left(1 + \frac{\gamma}{\beta+\sqrt{\nu}}\right)^{k-1}$ for all $k \geq 1$. $\square$

Let us discuss now the above complexity bound. The main term there is

$$6.11 \sqrt{\nu} \ln \frac{\nu\|c\|_{x_F^*}^*}{\epsilon}.$$

Note that the value $\nu \| c \|_{x_F^*}^*$ estimates from above the variation of the linear function $\langle c, x\rangle$ over the set Dom $F$ (see Theorem 5.3.9). Thus, the ratio $\frac{\epsilon}{\nu\|c\|_{x_F^*}^*}$ can be seen as the *relative accuracy* of the solution.

The process (5.3.29) has one drawback. Sometimes it is difficult to satisfy its starting condition

$$\| \nabla F(x_0) \|_{x_0}^* \leq \beta.$$

In this case, we need an additional process for *computing* an appropriate starting point. We analyze the corresponding strategies in the next section.

### 5.3.5 *Finding the Analytic Center*

Thus, our current goal is to find an approximation to the *analytic center* of the set Dom $F$. Let us look at the following minimization problem:

$$\min\{F(x) \mid x \in \mathrm{dom}\, F\}, \tag{5.3.30}$$

where $F$ is a $\nu$-self-concordant barrier. In view of the needs of the previous section, we accept an approximate solution $\bar{x} \in \mathrm{dom}\, F$ to this problem, which satisfies the inequality

$$\| \nabla F(\bar{x}) \|_{\bar{x}}^* \leq \beta,$$

for certain $\beta \in (0, 1)$.

As we have already discussed in Sect. 5.2, we can apply two different minimization strategies. The first one is a straightforward implementation of the Intermediate Newton's Method and the second one is based on a path-following approach.

Consider the first scheme.

**Intermediate Newton's Method for finding the analytic center**

**0.** Choose $y_0 \in \mathrm{dom}\, F$.

**1.** ***k*th iteration** ($k \geq 0$). Set

$$y_{k+1} = y_k - \frac{[\nabla^2 F(y_k)]^{-1} \nabla F(y_k)}{1+\xi_k}, \tag{5.3.31}$$

where $\xi_k = \frac{\lambda_k^2}{1+\lambda_k}$ and $\lambda_k = \| \nabla F(y_k) \|_{y_k}^*$ .

**2.** Stop the process if $\| \nabla F(y_k) \|_{y_k}^* \leq \beta$.

As we have seen already, this method needs $O(F(y_0) - F(x_F^*))$ iterations to enter to the region of quadratic convergence.

To implement the path-following approach, we need to choose some $y_0 \in \mathrm{dom}\, F$ and define the *auxiliary central path*:

$$y^*(t) = \arg \min_{y \in \mathrm{dom}\, F} [-t\langle \nabla F(y_0), y\rangle + F(y)],$$

where $t \geq 0$. Since this trajectory satisfies the equation

$$\nabla F(y^*(t)) \;=\; t\nabla F(y_0), \tag{5.3.32}$$

it connects two points, the starting point $y_0$ and the analytic center $x_F^*$:

$$y^*(1) = y_0, \quad y^*(0) \;=\; x_F^*.$$

As was shown in Lemma 5.2.2, we can follow this trajectory by the process (5.3.25) with *decreasing t*.

Let us estimate the rate of convergence of the auxiliary central path $y^*(t)$ to the analytic center in terms of the *barrier parameter*.

**Lemma 5.3.3** *For any $t \geq 0$, we have*

$$\| \nabla F(y^*(t)) \|_{y^*(t)}^* \leq (\nu + 2\sqrt{\nu}) \| \nabla F(y_0) \|_{x_F^*}^* \cdot t.$$

*Proof* This estimate follows from (5.3.32) and Corollary 5.3.4. □

Let us look now at the corresponding algorithmic scheme.

**Auxiliary Path-Following Scheme**

**0.** Choose $y_0 \in \mathrm{dom}\, F$. Set $t_0 = 1$.

**1.** ***k*th iteration** ($k \geq 0$). Set

$$t_{k+1} = t_k - \frac{\gamma}{\|\nabla F(y_0)\|^*_{y_k}},$$

$$y_{k+1} = y_k - \frac{1}{1+\xi_k}[\nabla^2 F(y_k)]^{-1}(-t_{k+1}\nabla F(y_0) + \nabla F(y_k)),$$

$$\text{where } \xi_k = \frac{\lambda_k^2}{1+\lambda_k} \text{ and } \lambda_k = \|t_{k+1}\nabla F(y_0) - \nabla F(y_k)\|^*_{y_k}.$$

**2.** Stop the process if $\| \nabla F(y_k) \|^*_{y_k} \leq \tau$. Set $\xi_k = \frac{\lambda_F(y_k)^2}{1+\lambda_F(y_k)}$ and $\bar{x} = y_k - \frac{1}{1+\xi_k}[\nabla^2 F(y_k)]^{-1}\nabla F(y_k)$.

(5.3.33)

Note that the above scheme follows the auxiliary central path $y^*(t)$ as $t_k \to 0$. It updates the points $\{y_k\}$ satisfying the approximate centering condition

$$\| -t_k\nabla F(y_0) + \nabla F(y_k) \|^*_{y_k} \leq \beta.$$

The termination criterion of this process,

$$\lambda_k = \| \nabla F(y_k) \|^*_{y_k} \leq \tau,$$

guarantees that $\| \nabla F(\bar{x}) \|^*_{\bar{x}} \leq \beta(\tau)$ (see Theorem 5.2.2). Let us derive a complexity bound for this process.

**Theorem 5.3.12** *The process (5.3.33) terminates no later than after*

$$\frac{1}{\gamma}(\beta + \sqrt{\nu}) \ln \left[\frac{1}{\gamma}(\nu + 2\sqrt{\nu}) \| \nabla F(y_0) \|^*_{x^*_F}\right]$$

*iterations.*

*Proof* Recall that our parameters are fixed by (5.3.28). Note that $t_0 = 1$. Therefore, in view of Lemmas 5.2.2 and 5.3.2, we have

$$t_{k+1} \leq \left(1 - \frac{\gamma}{\beta+\sqrt{\nu}}\right) t_k \leq \exp\left(-\frac{\gamma(k+1)}{\beta+\sqrt{\nu}}\right) t_0.$$

Further, in view of Lemma 5.3.3, we obtain

$$\begin{aligned}\| \nabla F(y_k) \|_{y_k}^* &= \| (-t_k \nabla F(x_0) + \nabla F(y_k)) + t_k \nabla F(y_0) \|_{y_k}^* \\ &\leq \beta + t_k \| \nabla F(y_0) \|_{y_k}^* \leq \ \beta + t_k(\nu + 2\sqrt{\nu}) \| \nabla F(y_0) \|_{x_F^*}^* .\end{aligned}$$

Thus, the process is terminated at most when the following inequality holds:

$$t_k(\nu + 2\sqrt{\nu}) \| \nabla F(y_0) \|_{x_F^*}^* \leq \tau - \beta(\tau) \ = \ \gamma. \qquad \Box$$

The principal term in the complexity bound of the auxiliary path-following scheme is

$$6.11\sqrt{\nu}[\ln \nu + \ln \| \nabla F(y_0) \|_{x_F^*}^*]$$

and for the auxiliary Intermediate Newton's method it is $O(F(y_0) - F(x_F^*))$. These estimates cannot be compared directly. However, as we have proved in Sect. 5.2.2 by another reasoning the path-following approach is much more efficient. Note also that its complexity estimate naturally fits the complexity of the main path-following process. Indeed, if we apply (5.3.29) with (5.3.33), we get the following complexity bound for the whole process:

$$6.11\sqrt{\nu}\left[2\ln \nu + \ln \| \nabla F(y_0) \|_{x_F^*}^* + \ln \| c \|_{x_F^*}^* + \ln \tfrac{1}{\epsilon}\right].$$

To conclude this section, note that for some problems it is difficult even to point out a starting point $y_0 \in \operatorname{dom} F$. In such cases, we should apply one more auxiliary minimization process, which is similar to the process (5.3.33). We discuss this situation in the next section.

### *5.3.6 Problems with Functional Constraints*

Let us consider the following minimization problem:

$$\min_{x \in Q} \{ f_0(x) : \ f_j(x) \leq 0, \ j = 1 \ldots m \}, \tag{5.3.34}$$

where $Q$ is a simple bounded closed convex set with nonempty interior and all functions $f_j$, $j = 0 \ldots m$, are convex. We assume that the problem satisfies the Slater condition: There exists an $\bar{x} \in \operatorname{int} Q$ such that $f_j(\bar{x}) < 0$ for all $j = 1 \ldots m$.

Let us assume that we know an upper bound $\bar{\xi}$ such that $f_0(x) < \bar{\xi}$ for all $x \in Q$. Then, introducing two additional variables $\xi$ and $\varkappa$, we can rewrite this problem in the standard form:

$$\min_{\substack{\xi \leq \bar{\xi},\, \varkappa \leq 0,\\ x \in Q}} \{\xi : \ f_0(x) \leq \xi, \ f_j(x) \leq \varkappa, \ j = 1 \dots m\}. \tag{5.3.35}$$

Note that we can apply interior-point methods to this problem only if we are able to construct a self-concordant barrier for the feasible set. In the current situation, this means that we should be able to construct the following barriers:

- A self-concordant barrier $F_Q(x)$ for the set $Q$.
- A self-concordant barrier $F_0(x, \xi)$ for the epigraph of the objective function $f_0(x)$.
- Self-concordant barriers $F_j(x, \varkappa)$ for the epigraphs of functional constraints $f_j(x)$.

Let us assume that we can do that. Then the resulting self-concordant barrier for the feasible set of problem (5.3.35) is as follows:

$$\hat{F}(x, \xi, \varkappa) = F_Q(x) + F_0(x, \xi) + \sum_{j=1}^{m} F_j(x, \varkappa) - \ln(\bar{\xi} - \xi) - \ln(-\varkappa).$$

The parameter of this barrier is

$$\hat{\nu} = \nu_Q + \nu_0 + \sum_{j=1}^{m} \nu_j + 2, \tag{5.3.36}$$

where $\nu_{(\cdot)}$ are the parameters of the corresponding barriers.

Note that it could still be difficult to find a starting point from $\operatorname{dom} \hat{F}$. This domain is an intersection of the set $Q$ with epigraphs of the objective function and constraints, and with two additional linear constraints $\xi \leq \bar{\xi}$ and $\varkappa \leq 0$. If we have a point $x_0 \in \operatorname{int} Q$, then we can choose $\xi_0$ and $\varkappa_0$ large enough to guarantee

$$f_0(x_0) \ < \ \xi_0 \ < \ \bar{\xi}, \quad f_j(x_0) \ < \ \varkappa_0, \ j = 1 \dots m.$$

Then, only constraint $\varkappa \leq 0$ will be violated.

In order to simplify our analysis, let us change the notation. From now on, we consider the problem

$$\min_{z \in S} \{\langle c, z \rangle : \ \langle d, z \rangle \leq 0\}, \tag{5.3.37}$$

where $z = (x, \xi, \varkappa)$, $\langle c, z \rangle \equiv \xi$, $\langle d, z \rangle \equiv \varkappa$ and $S$ is the feasible set of problem (5.3.35) without the constraint $\varkappa \leq 0$. Note that we know a self-concordant barrier $F(z)$ for the set $S$, and we can easily find a point $z_0 \in \operatorname{int} S$. Moreover, in view of our assumptions, the set

$$S(\alpha) \;=\; \{z \in S \mid \; \langle d, z\rangle \le \alpha\}$$

is bounded and, for $\alpha$ large enough, it has nonempty interior.

The process of solving problem (5.3.37) consists of three stages.

1. Choose a starting point $z_0 \in \operatorname{int} S$ and some initial gap $\Delta > 0$. Set $\alpha = \langle d, z_0\rangle + \Delta$. If $\alpha \le 0$, then we can use the two-stage process described in Sect. 5.3.5. Otherwise, we do the following. First, we find an approximate analytic center of the set $S(\alpha)$, generated by the barrier

$$\tilde{F}(z) = F(z) - \ln(\alpha - \langle d, z\rangle).$$

Namely, we find a point $\tilde{z}$ satisfying the condition

$$\lambda_{\tilde{F}}(\tilde{z}) \;\equiv\; \langle \nabla F(\tilde{z}) + \tfrac{d}{\alpha - \langle d, \tilde{z}\rangle}, [\nabla^2 \tilde{F}(\tilde{z})]^{-1} \left(\nabla F(\tilde{z}) + \tfrac{d}{\alpha - \langle d, \tilde{z}\rangle}\right)\rangle^{1/2} \le \beta.$$

In order to generate such a point, we can use the auxiliary schemes discussed in Sect. 5.3.5.

2. The next stage consists in following the central path $z(t)$ defined by the equation

$$td + \nabla \tilde{F}(z(t)) \;=\; 0, \quad t \ge 0.$$

Note that the previous stage provides us with a reasonable approximation to the analytic center $z(0)$. Therefore, we can follow this path, using the process (5.3.25). This trajectory leads us to the solution of the minimization problem

$$\min\{\langle d, z\rangle \mid \; z \in S(\alpha)\}.$$

In view of the Slater condition for problem (5.3.37), the optimal value of this problem is strictly negative.

The goal of this stage consists in finding an approximation to the analytic center of the set

$$\bar{S} = \{z \in S(\alpha) \mid \; \langle d, z\rangle \le 0\}$$

generated by the barrier $\bar{F}(z) = \tilde{F}(z) - \ln(-\langle d, z\rangle)$. This point, $z_*$, satisfies the equation

$$\nabla \tilde{F}(z_*) - \tfrac{d}{\langle d, z_*\rangle} = 0.$$

Therefore, $z_*$ is a point of the central path $z(t)$. The corresponding value of the penalty parameter $t_*$ is

$$t_* = -\tfrac{1}{\langle d, z_* \rangle} > 0.$$

This stage terminates with a point $\bar{z}$ satisfying the condition

$$\lambda_{\tilde{F}}(\bar{z}) \;\equiv\; \langle \nabla \tilde{F}(\bar{z}) - \tfrac{d}{\langle d, \bar{z} \rangle}, [\nabla^2 \tilde{F}(\bar{z})]^{-1} \left( \nabla \tilde{F}(\bar{z}) - \tfrac{d}{\langle d, \bar{z} \rangle} \right) \rangle^{1/2} \leq \beta.$$

3. Note that $\nabla^2 \bar{F}(z) \succeq \nabla^2 \tilde{F}(z)$. Therefore, the point $\bar{z}$, computed at the previous stage, satisfies the inequality

$$\lambda_{\bar{F}}(\bar{z}) \;\equiv\; \langle \nabla \tilde{F}(\bar{z}) - \tfrac{d}{\langle d, \bar{z} \rangle}, [\nabla^2 \bar{F}(\bar{z})]^{-1} \left( \nabla \tilde{F}(\bar{z}) - \tfrac{d}{\langle d, \bar{z} \rangle} \right) \rangle^{1/2} \leq \beta.$$

This means that we have a good approximation of the analytic center of the set $\bar{S}$, and we can apply the main path-following scheme (5.3.29) to solve the problem

$$\min\{\langle c, z \rangle : \; z \in \bar{S}\}.$$

Clearly, this problem is equivalent to (5.3.37).

We omit the detailed complexity analysis of the above three-stage scheme. It can be done similarly to the analysis of Sect. 5.3.5. The main term in the complexity of this scheme is proportional to the product of $\sqrt{\hat{\nu}}$ (see (5.3.36)) and the sum of the logarithm of the desired accuracy $\epsilon$ with logarithms of some structural characteristics of the problem (size of the region, depth of Slater condition, etc.).

Thus, we have shown that the interior point methods can be applied to all problems, for which we can point out some self-concordant barriers for the basic feasible set $Q$ and for the epigraphs of functional constraints. Our main goal now is to describe the classes of convex problems for which such barriers can be constructed in a computable form. Note that we have an exact characteristic of the quality of self-concordant barrier. This is the value of its parameter. The smaller it is, the more efficient will be the corresponding path-following scheme. In the next section, we discuss our possibilities in applying the developed theory to particular convex problems.

## 5.4 Applications to Problems with Explicit Structure

(Bounds on parameters of self-concordant barriers; Linear and quadratic optimization; Semidefinite optimization; Extremal ellipsoids; Constructing self-concordant barriers for particular sets; Separable problems; Geometric optimization; Approximation in $\ell_p$-norms; Choice of optimization scheme.)

### 5.4.1 Lower Bounds for the Parameter of a Self-concordant Barrier

In the previous section, we discussed a path-following scheme for solving the following problem:

$$\min_{x \in Q} \langle c, x \rangle, \tag{5.4.1}$$

where $Q$ is a closed convex set with nonempty interior, for which we know a $\nu$-self-concordant barrier $F(\cdot)$. Using such a barrier, we can solve (5.4.1) in $O\left(\sqrt{\nu} \cdot \ln \frac{\nu}{\epsilon}\right)$ iterations of a path-following scheme. Recall that the most difficult part of each iteration is the solution of a system of linear equations.

In this section, we study the limits of applicability of this approach. We discuss the lower and upper bounds for the parameters of self-concordant barriers. We also discuss some classes of convex problems for which the model (5.4.1) can be created in a computable form.

Let us start from the lower bounds on the barrier parameters.

**Lemma 5.4.1** *Let $f$ be a $\nu$-self-concordant barrier for the interval $(\alpha, \beta) \subset \mathbb{R}$, $\alpha < \beta < \infty$, where we admit the value $\alpha = -\infty$. Then*

$$\nu \geq \varkappa \overset{def}{=} \sup_{t \in (\alpha,\beta)} \frac{(f'(t))^2}{f''(t)} \geq 1.$$

*Proof* Note that $\nu \geq \varkappa$ by definition. Let us assume that $\varkappa < 1$. Since $f$ is a convex barrier function for $(\alpha, \beta)$, there exists a value $\bar{\alpha} \in (\alpha, \beta)$ such that $f'(t) > 0$ for all $t \in [\bar{\alpha}, \beta)$.

Consider the function $\phi(t) = \frac{(f'(t))^2}{f''(t)}$, $t \in [\bar{\alpha}, \beta)$. Then, since $f'(t) > 0$, $f(\cdot)$ is standard self-concordant, and $\phi(t) \leq \varkappa < 1$, we have

$$\begin{aligned} \phi'(t) &= 2f'(t) - \left(\frac{f'(t)}{f''(t)}\right)^2 f'''(t) \\ &= f'(t)\left(2 - \frac{f'(t)}{\sqrt{f''(t)}} \cdot \frac{f'''(t)}{[f''(t)]^{3/2}}\right) \geq 2(1 - \sqrt{\varkappa}) f'(t). \end{aligned}$$

Hence, for all $t \in [\bar{\alpha}, \beta)$ we obtain $\phi(t) \geq \phi(\bar{\alpha}) + 2(1 - \sqrt{\varkappa})(f(t) - f(\bar{\alpha}))$. This is a contradiction since $f$ is a barrier function and $\phi$ is bounded from above. □

**Corollary 5.4.1** *Let $F$ be a $\nu$-self-concordant barrier for $Q \subset \mathbb{E}$. Then $\nu \geq 1$.*

*Proof* Indeed, let $x \in \operatorname{int} Q$. Since $Q \subset \mathbb{E}$, there exists a nonzero direction $u \in \mathbb{E}$ such that the line $\{y = x + tu, \ t \in \mathbb{R}\}$ intersects the boundary of the set $Q$. Therefore, considering function $f(t) = F(x + tu)$, and using Lemma 5.4.1, we get the result. □

Let us prove a simple lower bound for parameters of self-concordant barriers for unbounded sets.

Let $Q$ be a closed convex set with nonempty interior. Consider $\bar{x} \in \operatorname{int} Q$. Assume that there exists a nontrivial set of *recession* directions $\{p_1, \ldots, p_k\}$ of the set $Q$:

$$\bar{x} + \alpha p_i \in Q \quad \forall \alpha \geq 0, \quad i = 1, \ldots, k.$$

**Theorem 5.4.1** *Let the positive coefficients $\{\beta_i\}_{i=1}^k$ satisfy the condition*

$$\bar{x} - \beta_i \ p_i \notin int\, Q, \quad i = 1, \ldots, k.$$

*If for some positive $\alpha_1, \ldots, \alpha_k$ we have $\bar{y} = \bar{x} - \sum\limits_{i=1}^k \alpha_i p_i \in Q$, then the parameter $\nu$ of any self-concordant barrier for the set $Q$ satisfies the inequality:*

$$\nu \geq \sum_{i=1}^k \frac{\alpha_i}{\beta_i}.$$

*Proof* Let $F$ be a $\nu$-self-concordant barrier for the set $Q$. Since $p_i$ is a recession direction, by Theorem 5.1.14 we have

$$\langle \nabla F(\bar{x}), -p_i \rangle \geq \langle \nabla^2 F(\bar{x}) p_i, p_i \rangle^{1/2} \equiv \| \ p_i \ \|_{\bar{x}} \ .$$

Note that $\bar{x} - \beta_i \ p_i \notin Q$. Therefore, in view of Theorem 5.1.5, the norm of direction $p_i$ is large enough: $\beta_i \ \| \ p_i \ \|_{\bar{x}} \geq 1$. Hence, in view of Theorem 5.3.7, we obtain

$$\begin{aligned} \nu \geq \langle \nabla F(\bar{x}), \bar{y} - \bar{x} \rangle \ &= \ \langle \nabla F(\bar{x}), -\sum_{i=1}^k \alpha_i p_i \rangle \\ &\geq \sum_{i=1}^k \alpha_i \ \| \ p_i \ \|_{\bar{x}} \geq \ \sum_{i=1}^k \frac{\alpha_i}{\beta_i}. \end{aligned}$$ □

## 5.4.2 Upper Bound: Universal Barrier and Polar Set

Let us present now an existence theorem for self-concordant barriers. Consider a closed convex set $Q$, $\operatorname{int} Q \neq \emptyset$, and assume that $Q$ contains no straight lines. Define a *polar set* of $Q$ with respect to some point $\bar{x} \in \operatorname{int} Q$ as follows:

$$P(\bar{x}) = \{s \in \mathbb{R}^n \mid \langle s, x - \bar{x} \rangle \leq 1, \ \forall x \in Q\}.$$

It can be proved that for any $x \in \operatorname{int} Q$ the set $P(x)$ is a bounded closed convex set with nonempty interior. It always contains the origin.

Define $V(x) = \operatorname{vol}_n P(x)$.

**Theorem 5.4.2** *There exist absolute constants $c_1$ and $c_2$, such that the function*

$$U(x) = c_1 \cdot \ln V(x)$$

*is a $(c_2 \cdot n)$-self-concordant barrier for $Q$.* □

We drop the proof of this statement since it is very technical. □

The function $U(\cdot)$ is called the *Universal Barrier* for the set $Q$. Note that the analytical complexity of problem (5.4.1), equipped with a universal barrier, is $O\left(\sqrt{n} \cdot \ln \frac{n}{\epsilon}\right)$ calls of oracle. Recall that such efficiency estimate is *impossible* for the methods based on a local Black-Box oracle (see Theorem 3.2.8).

The statement of Theorem 5.4.2 is mainly of theoretical interest. Indeed, in general, the value $U(x)$ cannot easily be computed. However, Theorem 5.4.2 demonstrates that self-concordant barriers, in principle, can be found for *any* convex set. Thus, the applicability of this approach is restricted only by our ability to construct a *computable* self-concordant barrier, hopefully with a small value of the parameter. The process of creating the *barrier model* of the initial problem can hardly be described in a formal way. For each particular problem, there could be many different barrier models, and we should choose the best one, taking into account the value of the parameter of the self-concordant barrier, the complexity of the computation of its gradient and Hessian, and the complexity of the solution of the corresponding Newton system.

In the remaining part of this section we will see how this can be done for some *standard* problem classes of Convex Optimization.

### 5.4.3 Linear and Quadratic Optimization

Let us start from a problem of Linear Optimization:

$$\min_{x \in \mathbb{R}^n_+} \{\langle c, x \rangle : \; Ax = b\}, \tag{5.4.2}$$

where $A$ is an $(m \times n)$-matrix, $m < n$. The basic feasible set in this problem is represented by the *positive orthant*, the set of all vectors with nonnegative coefficients in $\mathbb{R}^n$. It can be equipped with the following self-concordant barrier:

$$F(x) = -\sum_{i=1}^{n} \ln x^{(i)}, \quad \nu = n, \tag{5.4.3}$$

(see Example 5.3.1 and Theorem 5.3.2). This barrier is called the *standard logarithmic barrier* for $\mathbb{R}^n_+$.

In order to solve problem (5.4.2), we have to use a restriction of the barrier $F$ onto the affine subspace $\{x : Ax = b\}$. Since this restriction is an $n$-self-concordant barrier (see Theorem 5.3.3), the complexity bound for problem (5.4.2) is $O\left(\sqrt{n} \cdot \ln \frac{n}{\epsilon}\right)$ iterations of a path-following scheme.

Let us prove that the standard logarithmic barrier is optimal for $\mathbb{R}^n_+$.

**Lemma 5.4.2** *The parameter $\nu$ of any self-concordant barrier for $\mathbb{R}^n_+$ satisfies inequality $\nu \geq n$.*

*Proof* Let us choose

$$\bar{x} = \bar{e}_n \equiv (1, \dots, 1)^T \in \operatorname{int} \mathbb{R}^n_+,$$

$$p_i = e_i, \quad i = 1 \dots n,$$

where $e_i$ is the $i$th coordinate vector of $\mathbb{R}^n$. In this case the conditions of Theorem 5.4.1 are satisfied with $\alpha_i = \beta_i = 1$, $i = 1 \dots n$. Therefore,

$$\nu \geq \sum_{i=1}^{n} \frac{\alpha_i}{\beta_i} = n. \qquad \square$$

Note that the above lower bound is valid only for the whole set $\mathbb{R}^n_+$. The lower bound for the intersection $\{x \in \mathbb{R}^n_+ \mid Ax = b\}$ can be smaller.

Self-concordant barriers for cones usually have one important property, which is called *logarithmic homogeneity* (e.g. (5.4.3)).

**Definition 5.4.1** A function $F \in C^2(\mathbb{E})$ with $\operatorname{Dom} F = K$, where $K$ is a closed convex cone, is called logarithmically homogeneous if there exists a constant $\nu \geq 1$ such that

$$F(\tau x) = F(x) - \nu \ln \tau, \quad \forall x \in \operatorname{int} K, \ \tau > 0. \tag{5.4.4}$$

This simple property has surprisingly many interesting consequences, one of which makes the computation of the barrier parameter completely trivial.

**Lemma 5.4.3** *Let $F$ be a logarithmically homogeneous self-concordant barrier for a convex cone $K$ which contains no straight lines. Then for any $x \in \operatorname{int} K$ and $\tau > 0$ we have*

$$\nabla F(\tau x) = \frac{1}{\tau} \nabla F(x), \quad \nabla^2 F(\tau x) = \frac{1}{\tau^2} \nabla^2 F(x), \tag{5.4.5}$$

$$\langle \nabla F(x), x \rangle = -\nu, \quad \nabla^2 F(x) x = -\nabla F(x), \tag{5.4.6}$$

$$\langle \nabla^2 F(x) x, x \rangle = \nu, \quad \langle \nabla F(x), [\nabla^2 F(x)]^{-1} \nabla F(x) \rangle = \nu. \tag{5.4.7}$$

*Proof* Differentiating identity (5.4.4) in $x$, we get the first identity in (5.4.5). Differentiating the latter identity in $x$ again, we get the second relation in (5.4.5).

Differentiating identity (5.4.4) in $\tau$ and taking $\tau = 1$, we get the first identity in (5.4.6). Differentiating it in $x$, we obtain the second identity in this line.

Finally, substituting the last expression in (5.4.6) into the first one, we get the first identity in (5.4.7). Since $K$ contains no straight lines $\nabla^2 F(x)$ is non-degenerate. Therefore, $x = -[\nabla^2 F(x)]^{-1}\nabla F(x)$, and we get the last expression in (5.4.7). □

Thus, for logarithmically homogeneous barriers, the degree of homogeneity is always equal to the barrier parameter (see the second identity in (5.4.7)).

Let us look now at the quadratically constrained quadratic optimization problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \{\, & q_0(x) = \alpha_0 + \langle a_0, x\rangle + \tfrac{1}{2}\langle A_0 x, x\rangle, \\ & q_i(x) = \alpha_i + \langle a_i, x\rangle + \tfrac{1}{2}\langle A_i x, x\rangle \le \beta_i,\ i = 1\dots m\}, \end{aligned} \tag{5.4.8}$$

where $A_i$ are some positive semidefinite $(n \times n)$-matrices. Let us rewrite this problem in the standard form:

$$\min_{x\in\mathbb{R}^n, \tau\in\mathbb{R}} \{\tau :\ q_0(x) \le \tau,\ q_i(x) \le \beta_i,\ i = 1\dots m\}. \tag{5.4.9}$$

The feasible set of this problem can be equipped with the following self-concordant barrier:

$$F(x,\tau) \ = -\ln(\tau - q_0(x)) - \sum_{i=1}^{m} \ln(\beta_i - q_i(x)), \quad \nu = m+1,$$

(see Example 5.3.1, and Theorem 5.3.2). Thus, the complexity bound for problem (5.4.8) is $O\left(\sqrt{m+1}\cdot \ln\frac{m}{\epsilon}\right)$ iterations of a path-following scheme. Note that this estimate *does not depend* on $n$.

In some applications, the functional components of the problem include a nonsmooth quadratic term of the form $\| Ax - b \|$, where the norm is standard Euclidean. Let us show that we can treat such terms using an interior-point technique.

**Lemma 5.4.4** *The function*

$$F(x,t) = -\ln(t^2 - \| x \|^2)$$

*is a* 2*-self-concordant barrier for the convex cone*[5]

$$K_2 = \{(x,t) \in \mathbb{R}^{n+1} \mid t \ge \| x \|\}.$$

[5] Depending on the field, this set has different names: Lorentz cone, ice-cream cone, second-order cone.

*Proof* Let us fix a point $z = (x, t) \in \operatorname{int} K_2$ and a nonzero direction $u = (h, \tau) \in \mathbb{R}^{n+1}$. Let $\xi(\alpha) = (t + \alpha\tau)^2 - \parallel x + \alpha h \parallel^2$. We need to compare the derivatives of the function

$$\phi(\alpha) = F(z + \alpha u) = -\ln \xi(\alpha)$$

at $\alpha = 0$. Let $\phi^{(\cdot)} = \phi^{(\cdot)}(0)$, $\xi^{(\cdot)} = \xi^{(\cdot)}(0)$. Then

$$\xi' = 2(t\tau - \langle x, h\rangle), \quad \xi'' = 2(\tau^2 - \parallel h \parallel^2), \quad \xi''' = 0,$$

$$\phi' = -\tfrac{\xi'}{\xi}, \quad \phi'' = \left(\tfrac{\xi'}{\xi}\right)^2 - \tfrac{\xi''}{\xi}, \quad \phi''' = 3\tfrac{\xi'\xi''}{\xi^2} - 2\left(\tfrac{\xi'}{\xi}\right)^3.$$

Note that inequality $2\phi'' \geq (\phi')^2$ is equivalent to $(\xi')^2 \geq 2\xi\xi''$. Thus, we need to prove that for any $(h, \tau)$ we have

$$(t\tau - \langle x, h\rangle)^2 \geq (t^2 - \parallel x \parallel^2)(\tau^2 - \parallel h \parallel^2).$$

After opening the brackets and cancellation, we come to the inequality

$$\tau^2\|x\|^2 + t^2\|h\|^2 + \langle x, h\rangle^2 - 2\tau t\langle x, h\rangle \geq \|x\|^2\|h\|^2.$$

Minimizing the left-hand side in $\tau$, we get inequality

$$t^2\|h\|^2 + \langle x, h\rangle^2 - t^2\tfrac{\langle x,h\rangle^2}{\|x\|^2} \geq \|x\|^2\|h\|^2,$$
$$\Updownarrow$$
$$\|h\|^2(t^2 - \|x\|^2) \geq \langle x, h\rangle^2\left(\tfrac{t^2}{\|x\|^2} - 1\right),$$

which is valid since $t \geq \|x\|$.

Finally, since $0 \leq \frac{\xi\xi''}{(\xi')^2} \leq \frac{1}{2}$ and $[1 - \xi]^{3/2} \geq 1 - \frac{3}{2}\xi$, we get the following:

$$\tfrac{|\phi'''|}{(\phi'')^{3/2}} = 2\tfrac{|\xi'|\cdot|(\xi')^2 - \frac{3}{2}\xi\xi''|}{[(\xi')^2 - \xi\xi'']^{3/2}} \leq 2. \qquad \square$$

Let us prove that the barrier described in the above statement is optimal for the second-order cone.

**Lemma 5.4.5** *The parameter $\nu$ of any self-concordant barrier for the set $K_2$ satisfies the inequality $\nu \geq 2$.*

*Proof* Let us choose $\bar{z} = (0, 1) \in \operatorname{int} K_2$ and some $h \in \mathbb{R}^n$, $\parallel h \parallel = 1$. Define

$$p_1 = (h, 1), \quad p_2 = (-h, 1), \quad \alpha_1 = \alpha_2 = \frac{1}{2}, \quad \beta_1 = \beta_2 = \frac{1}{2}.$$

Note that for all $\gamma \geq 0$ we have $\bar{z} + \gamma p_i = (\pm\gamma h, 1 + \gamma) \in K_2$ and

$$\bar{z} - \beta_i p_i = (\pm\tfrac{1}{2}h, \tfrac{1}{2}) \notin \operatorname{int} K_2,$$

$$\bar{z} - \alpha_1 p_1 - \alpha_2 p_2 = (-\tfrac{1}{2}h + \tfrac{1}{2}h, 1 - \tfrac{1}{2} - \tfrac{1}{2}) \;=\; 0 \in K_2.$$

Therefore, the conditions of Theorem 5.4.1 are satisfied and

$$\nu \geq \tfrac{\alpha_1}{\beta_1} + \tfrac{\alpha_2}{\beta_2} \;=\; 2. \qquad \square$$

### 5.4.4 Semidefinite Optimization

In Semidefinite Optimization, the decision variables are matrices. Let

$$X = \{X^{(i,j)}\}_{i,j=1}^{n}$$

be a symmetric $n \times n$-matrix (notation: $X \in \mathbb{S}^n$). The real vector space $\mathbb{S}^n$ can be provided with the following inner product: for any $X, Y \in \mathbb{S}^n$ define

$$\langle X, Y\rangle_F = \sum_{i=1}^{n}\sum_{j=1}^{n} X^{(i,j)}Y^{(i,j)}, \qquad \| X \|_F = \langle X, X\rangle_F^{1/2}.$$

Sometimes the value $\| X \|_F$ is called the *Frobenius norm* of the matrix $X$. For symmetric matrices $X$ and $Y$, we have the following identity:

$$\begin{aligned}\langle X, Y \cdot Y\rangle_F &= \sum_{i=1}^{n}\sum_{j=1}^{n} X^{(i,j)} \sum_{k=1}^{n} Y^{(i,k)}Y^{(j,k)} \;=\; \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} X^{(i,j)}Y^{(i,k)}Y^{(j,k)} \\ &= \sum_{k=1}^{n}\sum_{j=1}^{n} Y^{(k,j)} \sum_{i=1}^{n} X^{(j,i)}Y^{(i,k)} \;=\; \sum_{k=1}^{n}\sum_{j=1}^{n} Y^{(k,j)}(XY)^{(j,k)} \\ &= \sum_{k=1}^{n} (YXY)^{(k,k)} \;=\; \text{Trace}\,(YXY) \;=\; \langle YXY, I_n\rangle_F. \end{aligned} \tag{5.4.10}$$

In Semidefinite Optimization, a nontrivial part of the constraints is formed by the *cone of positive semidefinite* $n \times n$-matrices $\mathbb{S}^N_+ \subset \mathbb{S}^n$. Recall that $X \in \mathbb{S}^n_+$ if and only if $\langle Xu, u\rangle \geq 0$ for any $u \in \mathbb{R}^n$. If $\langle Xu, u\rangle > 0$ for all nonzero $u$, we call $X$ *positive definite*. Such matrices form the interior of cone $\mathbb{S}^n_+$. Note that $\mathbb{S}^n_+$ is a closed convex set.

The general formulation of the Semidefinite Optimization problem is as follows:

$$\min_{X\in\mathbb{S}^n_+} \{\langle C, X\rangle_F : \ \langle A_i, X\rangle_F = b_i, \ i = 1\dots m\}, \tag{5.4.11}$$

where $C$ and all $A_i$ belong to $\mathbb{S}^n$. In order to apply a path-following scheme to this problem, we need a self-concordant barrier for the cone $\mathbb{S}^n_+$.

Let the matrix $X$ belong to $\operatorname{int}\mathbb{S}^n_+$. Define $F(X) = -\ln\det X$. Clearly

$$F(X) = -\sum_{i=1}^{n} \ln \lambda_i(X),$$

where $\{\lambda_i(X)\}_{i=1}^n$ is the set of eigenvalues of matrix $X$.

**Lemma 5.4.6** *Function $F$ is convex and $\nabla F(X) = -X^{-1}$. Moreover, for any direction $\Delta \in \mathbb{S}^n$, we have*

$$\begin{aligned}\langle \nabla^2 F(X)\Delta, \Delta\rangle_F &= \| X^{-1/2}\Delta X^{-1/2} \|_F^2 \ = \ \langle X^{-1}\Delta X^{-1}, \Delta\rangle_F \\ &= Trace\left([X^{-1/2}\Delta X^{-1/2}]^2\right), \\ D^3F(x)[\Delta,\Delta,\Delta] &= -2\langle I_n, [X^{-1/2}\Delta X^{-1/2}]^3\rangle_F \\ &= -2Trace\left([X^{-1/2}\Delta X^{-1/2}]^3\right).\end{aligned}$$

*Proof* Let us fix some $\Delta \in \mathbb{S}^n$ and $X \in \operatorname{int}\mathbb{S}^n_+$ such that $X + \Delta \in \mathbb{S}^n_+$. Then

$$\begin{aligned}F(X+\Delta) - F(X) &= -\ln\det(X+\Delta) - \ln\det X \\ &= -\ln\det(I_n + X^{-1/2}\Delta X^{-1/2}) \\ &\geq -\ln\left[\tfrac{1}{n}\text{Trace}\,(I_n + X^{-1/2}\Delta X^{-1/2})\right]^n \\ &= -n\ln\left[1 + \tfrac{1}{n}\langle I_n, X^{-1/2}\Delta X^{-1/2}\rangle_F\right] \\ &\geq -\langle I_n, X^{-1/2}\Delta X^{-1/2}\rangle_F \ = \ -\langle X^{-1}, \Delta\rangle_F.\end{aligned}$$

Thus, $-X^{-1} \in \partial F(X)$. Therefore, $F$ is convex (Lemma 3.1.6) and $\nabla F(x) = -X^{-1}$ (Lemma 3.1.7).

Further, consider the function $\phi(\alpha) \equiv \langle \nabla F(X + \alpha\Delta), \Delta \rangle_F$, $\alpha \in [0, 1]$. Then

$$\begin{aligned}\phi(\alpha) - \phi(0) &= \langle X^{-1} - (X + \alpha\Delta)^{-1}, \Delta \rangle_F \\ &= \langle (X + \alpha\Delta)^{-1}[(X + \alpha\Delta) - X]X^{-1}, \Delta \rangle_F \\ &= \alpha \langle (X + \alpha\Delta)^{-1}\Delta X^{-1}, \Delta \rangle_F.\end{aligned}$$

Thus, $\phi'(0) = \langle \nabla^2 F(X)\Delta, \Delta \rangle_F = \langle X^{-1}\Delta X^{-1}, \Delta \rangle_F$.

The last expression can be proved in a similar way by differentiating the function $\psi(\alpha) = \langle (X + \alpha\Delta)^{-1}\Delta(X + \alpha\Delta)^{-1}, \Delta \rangle_F$. □

**Theorem 5.4.3** *The function $F$ is an $n$-self-concordant barrier for $\mathbb{S}^n_+$.*

*Proof* Let us fix $X \in \operatorname{int} \mathbb{S}^n_+$ and $\Delta \in \mathbb{S}^n$. Define $Q = X^{-1/2}\Delta X^{-1/2}$ and $\lambda_i = \lambda_i(Q)$, $i = 1 \dots n$. Then, in view of Lemma 5.4.6, we have

$$\begin{aligned}\langle \nabla F(X), \Delta \rangle_F &= \sum_{i=1}^n \lambda_i, \\ \langle \nabla^2 F(X)\Delta, \Delta \rangle_F &= \sum_{i=1}^n \lambda_i^2, \\ D^3 F(X)[\Delta, \Delta, \Delta] &= -2 \sum_{i=1}^n \lambda_i^3.\end{aligned}$$

Using the two standard inequalities

$$\left(\sum_{i=1}^n \lambda_i\right)^2 \le n \sum_{i=1}^n \lambda_i^2, \quad \left|\sum_{i=1}^n \lambda_i^3\right| \le \left(\sum_{i=1}^n \lambda_i^2\right)^{3/2},$$

we obtain

$$\begin{aligned}\langle \nabla F(X), \Delta \rangle_F^2 &\le n \langle \nabla^2 F(X)\Delta, \Delta \rangle_F, \\ |\ D^3 F(X)[\Delta, \Delta, \Delta]\ | &\le 2\langle \nabla^2 F(X)\Delta, \Delta \rangle_F^{3/2}. \qquad \square\end{aligned}$$

Let us prove that $F(X) = -\ln \det X$ is the optimal barrier for $\mathbb{S}^n_+$.

**Lemma 5.4.7** *The parameter $\nu$ of any self-concordant barrier for the cone $\mathbb{S}^n_+$ satisfies the inequality $\nu \ge n$.*

*Proof* Let us choose $\bar{X} = I_n \in \operatorname{int} \mathbb{S}^n_+$ and directions $P_i = e_i e_i^T$, $i = 1 \dots n$, where $e_i$ is the $i$th coordinate vector of $\mathbb{R}^n$. Note that for any $\gamma \ge 0$ we have

$I_n + \gamma P_i \in \operatorname{int} \mathbb{S}^n_+$. Moreover,

$$I_n - e_i e_i^T \notin \operatorname{int} \mathbb{S}^n_+, \quad I_n - \sum_{i=1}^{n} e_i e_i^T = 0 \in \mathbb{S}^n_+.$$

Therefore conditions of Theorem 5.4.1 are satisfied with $\alpha_i = \beta_i = 1$, $i = 1 \dots n$, and we obtain $\nu \geq \sum\limits_{i=1}^{n} \frac{\alpha_i}{\beta_i} = n$. □

As in Linear Optimization problem (5.4.2), in problem (5.4.11) we need to use the restriction of $F$ onto the affine subspace

$$\mathscr{L} = \{X : \langle A_i, X \rangle_F = b_i, \ i = 1 \dots m\}.$$

This restriction is an $n$-self-concordant barrier in view of Theorem 5.3.3. Thus, the complexity bound of the problem (5.4.11) is $O\left(\sqrt{n} \cdot \ln \frac{n}{\epsilon}\right)$ iterations of a path-following scheme. Note that this estimate is very encouraging since the dimension of the problem (5.4.11) is $\frac{1}{2} n(n+1)$.

Let us estimate the arithmetical cost of each iteration of a path-following scheme (5.3.29) as applied to the problem (5.4.11). Note that we work with a restriction of the barrier $F$ to the set $\mathscr{L}$. In view of Lemma 5.4.6, each Newton step consists in solving the following problem:

$$\min_{\Delta} \{ \langle U, \Delta \rangle_F + \tfrac{1}{2} \langle X^{-1} \Delta X^{-1}, \Delta \rangle_F : \ \langle A_i, \Delta \rangle_F = 0, \ i = 1 \dots m \},$$

where $X \succ 0$ belongs to $\mathscr{L}$ and $U$ is a combination of the cost matrix $C$ and the gradient $\nabla F(X)$. In accordance with the statement (3.1.59), the solution of this problem can be found from the following system of linear equations:

$$\begin{aligned} U + X^{-1} \Delta X^{-1} &= \sum_{j=1}^{m} \lambda^{(j)} A_j, \\ \langle A_i, \Delta \rangle_F &= 0, \quad i = 1 \dots m. \end{aligned} \tag{5.4.12}$$

From the first equation in (5.4.12) we get

$$\Delta = X \left[ -U + \sum_{j=1}^{m} \lambda^{(j)} A_j \right] X. \tag{5.4.13}$$

Substituting this expression into the second equation in (5.4.12), we get the linear system

$$\sum_{j=1}^{m} \lambda^{(j)} \langle A_i, X A_j X \rangle_F = \langle A_i, X U X \rangle_F, \quad i = 1 \dots m, \tag{5.4.14}$$

which can be written in matrix form as $S\lambda = d$ with

$$S^{(i,j)} = \langle A_i, XA_jX\rangle_F, \quad d^{(j)} = \langle U, XA_jX\rangle_F, \quad i, j = 1\dots n.$$

Thus, a straightforward strategy of solving system (5.4.12) consists in the following steps.

- Compute the matrices $XA_jX$, $j = 1\dots m$. Cost: $O(mn^3)$ operations.
- Compute the elements of $S$ and $d$. Cost: $O(m^2n^2)$ operations.
- Compute $\lambda = S^{-1}d$. Cost: $O(m^3)$ operations.
- Compute $\Delta$ by (5.4.13). Cost: $O(mn^2)$ operations.

Taking into account that $m \leq \frac{n(n+1)}{2}$ we conclude that the complexity of one Newton step does not exceed

$$\boxed{O(n^2(m+n)m) \text{ arithmetic operations.}} \qquad (5.4.15)$$

However, if the matrices $A_j$ possess a certain structure, then this estimate can be significantly improved. For example, if all $A_j$ are of rank 1:

$$A_j = a_ja_j^T, \quad a_j \in \mathbb{R}^n, \quad j = 1\dots m,$$

then the computation of the Newton step can be done in

$$\boxed{O((m+n)^3) \text{ arithmetic operations.}} \qquad (5.4.16)$$

We leave the justification of this claim as an exercise for the reader.

To conclude this section, note that in many important applications we can use the barrier $-\ln\det(\cdot)$ to treat some functions of eigenvalues. Consider, for example, a matrix $\mathscr{A}(x) \in \mathbb{S}^n$ which depends linearly on $x$. Then the convex region

$$\{(x,t) \mid \max_{1\leq i\leq n} \lambda_i(\mathscr{A}(x)) \leq t\}$$

can be described by a self-concordant barrier

$$F(x,t) = -\ln\det(tI_n - \mathscr{A}(x)).$$

The value of the parameter of this barrier is equal to $n$.

## 5.4.5 *Extremal Ellipsoids*

In some applications, we are interested in approximating different sets by ellipsoids. Let us consider the most important examples.

### 5.4.5.1 Circumscribed Ellipsoid

Given a set of points $a_1, \dots, a_m \in \mathbb{R}^n$, find an ellipsoid $W$ with the minimal volume which contains all points $\{a_i\}$.

Let us pose this problem in a formal way. First of all, note that any bounded ellipsoid $W \subset \mathbb{R}^n$ can be represented as

$$W = \{x \in \mathbb{R}^n \mid x = H^{-1}(v+u), \ \| u \| \leq 1\},$$

where $H \in \operatorname{int} \mathbb{S}^n_+$, $v \in \mathbb{R}^n$, and the norm is standard Euclidean. Then the inclusion $a \in W$ is equivalent to the inequality $\| Ha - v \| \leq 1$. Note also that

$$\operatorname{vol}_n W = \operatorname{vol}_n B_2(0,1) \cdot \det H^{-1} \ = \ \tfrac{\operatorname{vol}_n B_2(0,1)}{\det H}.$$

Thus, our problem is as follows:

$$\min_{\substack{H \in \mathbb{S}^n_+, \\ v \in \mathbb{R}^n, \tau \in \mathbb{R}}} \{\tau : \ -\ln\det H \leq \tau, \ \| Ha_i - v \| \leq 1, \ i = 1 \dots m\}. \tag{5.4.17}$$

In order to solve this problem by an interior-point scheme, we need to find a self-concordant barrier for the feasible set. In view of Theorems 5.4.3 and 5.3.5, we know self-concordant barriers for all components. Indeed, we can use the following barrier:

$$F(H, v, \tau) = -\ln\det H - \ln(\tau + \ln\det H) - \sum_{i=1}^{m} \ln(1 - \| Ha_i - v \|^2),$$

$$\nu = m + n + 1.$$

The corresponding complexity bound is $O\left(\sqrt{m+n+1} \cdot \ln \frac{m+n}{\epsilon}\right)$ iterations of a path-following scheme.

#### 5.4.5.2 Inscribed Ellipsoid with Fixed Center

Let $Q$ be a convex polytope defined by a set of linear inequalities:

$$Q = \{x \in \mathbb{R}^n \mid \langle a_i, x\rangle \le b_i,\ i = 1 \dots m\},$$

and let $v \in \operatorname{int} Q$. Find an ellipsoid $W \subset Q$ with the biggest volume which is centered at $v$.

Let us fix some $H \in \operatorname{int} \mathbb{S}^n_+$. We can represent the ellipsoid $W$ as

$$W = \{x \in \mathbb{R}^n \mid \langle H^{-1}(x - v), x - v\rangle \le 1\}.$$

We need the following simple result.

**Lemma 5.4.8** *Let $\langle a, v\rangle < b$. The inequality $\langle a, x\rangle \le b$ is valid for all $x \in W$ if and only if*

$$\langle Ha, a\rangle \le (b - \langle a, v\rangle)^2.$$

*Proof* In view of Lemma 3.1.20, we have

$$\max_u \{\langle a, u\rangle \mid \langle H^{-1}u, u\rangle \le 1\} = \langle Ha, a\rangle^{1/2}.$$

Therefore, we need to ensure

$$\begin{aligned} \max_{x \in W} \langle a, x\rangle &= \max_{x \in W} [\langle a, x - v\rangle + \langle a, v\rangle] \\ &= \langle a, v\rangle + \max_x \{\langle a, u\rangle \mid \langle H^{-1}u, u\rangle \le 1\} \\ &= \langle a, v\rangle + \langle Ha, a\rangle^{1/2} \le b. \end{aligned}$$

This proves our statement since $\langle a, v\rangle < b$. □

Note that $\operatorname{vol}_n W = \operatorname{vol}_n B_2(0, 1)[\det H]^{1/2}$. Hence, our problem is as follows:

$$\min_{H \in \mathbb{S}^n_+, \tau \in \mathbb{R}} \{\tau :\ -\ln \det H \le \tau,\ \langle Ha_i, a_i\rangle \le (b_i - \langle a_i, v\rangle)^2,\ i = 1 \dots m\}. \tag{5.4.18}$$

In view of Theorems 5.4.3 and 5.3.5, we can use the following self-concordant barrier:

$$F(H,\tau) = -\ln\det H - \ln(\tau + \ln\det H) - \sum_{i=1}^{m} \ln[(b_i - \langle a_i, v\rangle)^2 - \langle Ha_i, a_i\rangle],$$

with barrier parameter $\nu = m + n + 1$. The complexity bound of the corresponding path-following scheme is

$$O\left(\sqrt{m+n+1} \cdot \ln \frac{m+n}{\epsilon}\right)$$

iterations.

#### 5.4.5.3 Inscribed Ellipsoid with Free Center

Let $Q$ be a convex polytope defined by a set of linear inequalities:

$$Q = \{x \in \mathbb{R}^n \mid \langle a_i, x\rangle \le b_i,\ i = 1 \dots m\},$$

and let $\operatorname{int} Q \neq \emptyset$. Find an ellipsoid $W$ with the biggest volume which is contained in $Q$.

Let $G \in \operatorname{int} \mathbb{S}^n_+$ and $v \in \operatorname{int} Q$. We can represent $W$ as follows:

$$W = \{x \in \mathbb{R}^n \mid \| G^{-1}(x - v) \| \le 1\}$$

$$\equiv \{x \in \mathbb{R}^n \mid \langle G^{-2}(x - v), x - v\rangle \le 1\}.$$

In view of Lemma 5.4.8, inequality $\langle a, x\rangle \le b$ is valid for any $x \in W$ if and only if

$$\| Ga \|^2 \equiv \langle G^2 a, a\rangle \le (b - \langle a, v\rangle)^2.$$

This gives us a convex feasible set for parameters $(G, v)$:

$$\| Ga \| \le b - \langle a, v\rangle.$$

Note that $\mathrm{vol}_n\, W = \mathrm{vol}_n\, B_2(0,1) \det G$. Therefore, our problem can be written as follows:

$$\min_{\substack{G \in \mathbb{S}^n_+,\\ v \in \mathbb{R}^n, \tau \in \mathbb{R}}} \{\tau :\ -\ln\det G \le \tau,\ \| Ga_i \| \le b_i - \langle a_i, v\rangle,\ i = 1 \dots m\}. \tag{5.4.19}$$

In view of Theorems 5.4.3, 5.3.5 and Lemma 5.4.4, we can use the following self-concordant barrier:

$$F(G, v, \tau) = -\ln\det G - \ln(\tau + \ln\det G) - \sum_{i=1}^{m} \ln[(b_i - \langle a_i, v\rangle)^2 - \parallel Ga_i \parallel^2]$$

with barrier parameter $\nu = 2m + n + 1$. The corresponding efficiency estimate is $O\left(\sqrt{2m+n+1} \cdot \ln \frac{m+n}{\epsilon}\right)$ iterations of a path-following scheme.

### 5.4.6 Constructing Self-concordant Barriers for Convex Sets

In this section we develop a general framework for constructing self-concordant barriers for convex cones. First of all, let us define the objects we are working with. They are related to three different real vector spaces, $\mathbb{E}_1$, $\mathbb{E}_2$, and $\mathbb{E}_3$.

Consider a function $\xi(\cdot) : \mathbb{E}_1 \to \mathbb{E}_2$ defined on a closed convex set $Q_1 \subset \mathbb{E}_1$. Assume that $\xi$ is three times continuously differentiable and *concave* with respect to a closed convex cone $K \subset E_2$:

$$-D^2\xi(x)[h,h] \in K \quad \forall x \in \text{int}\, Q_1, \;\; h \in \mathbb{E}_1. \tag{5.4.20}$$

It is convenient to write this inclusion as $D^2\xi(x)[h,h] \preceq_K 0$.

**Definition 5.4.2** Let $F(\cdot)$ be a $\nu$-self-concordant barrier for $Q_1$ and $\beta \geq 1$. We say that a function $\xi$ is $\beta$-compatible with $F$ if for all $x \in \text{int}\, Q_1$ and $h \in \mathbb{E}_1$ we have

$$D^3\xi(x)[h,h,h] \preceq_K -3\beta \cdot D^2\xi(x)[h,h] \cdot \langle \nabla^2 F(x)h, h\rangle^{1/2}. \tag{5.4.21}$$

Alternating the sign of direction $h$ in (5.4.21), we get the following equivalent condition:

$$-D^3\xi(x)[h,h,h] \preceq_K -3\beta \cdot D^2\xi(x)[h,h] \cdot \langle \nabla^2 F(x)h, h\rangle^{1/2}. \tag{5.4.22}$$

Note that the set of $\beta$-compatible functions is a convex cone: if functions $\xi_1$ and $\xi_2$ are $\beta$-compatible with barrier $F$, then the sum $\alpha_1\xi_1 + \alpha_2\xi_2$, with arbitrary $\alpha_1, \alpha_2 > 0$, is also $\beta$-compatible with $F$.

Let us construct a self-concordant barrier for a *composition* of the set

$$\mathscr{S}_1 = \{(x, y) \in Q_1 \times \mathbb{E}_2 : \; \xi(x) \succeq_K y\}$$

and a convex set $Q_2 \subset \mathbb{E}_2 \times \mathbb{E}_3$. That is

$$\mathscr{Q} = \{(x, z) \in Q_1 \times \mathbb{E}_3 : \; \exists y, \;\; \xi(x) \succeq_K y, \;\; (y, z) \in Q_2\}.$$

The necessity of such a structure is clear from the following example.

*Example 5.4.1* Let us fix some $\alpha \in (0,1)$. Consider the following *power cone*

$$K_\alpha = \left\{ (x^{(1)}, x^{(2)}, z) \in \mathbb{R}^2_+ \times \mathbb{R} : \; (x^{(1)})^\alpha \cdot (x^{(2)})^{1-\alpha} \geq |z| \right\}.$$

For our representation, we need the following objects:

$$\begin{aligned} &\mathbb{E}_1 = \mathbb{R}^2, \quad Q_1 = \mathbb{R}^2_+, \quad F(x) = -\ln x^{(1)} - \ln x^{(2)}, \quad \nu = 2, \\ &\mathbb{E}_2 = \mathbb{R}, \quad \xi(x) = (x^{(1)})^\alpha \cdot (x^{(2)})^{1-\alpha}, \quad K = \mathbb{R}_+ \subset \mathbb{E}_2, \\ &\mathbb{E}_3 = \mathbb{R}, \quad Q_2 = \{(y,z) \in \mathbb{E}_2 \times \mathbb{E}_3 : \; y \geq |z|\}. \qquad \square \end{aligned}$$

In our construction, we also need a $\mu$-self-concordant barrier $\Phi(y,z)$ for the set $Q_2$. We assume that all directions from the cone $K_0 \overset{\mathrm{def}}{=} K \times \{0\} \subset \mathbb{E}_2 \times \mathbb{E}_3$ are recession directions of the set $Q_2$. Consequently, for any $s \in K$ and $(y,z) \in \operatorname{int} Q_2$ we have

$$\langle \nabla_y \Phi(y,z), s \rangle \; = \; \langle \nabla \Phi(y,z), (s,0) \rangle \overset{(5.3.13)}{\leq} 0. \tag{5.4.23}$$

Consider the barrier

$$\Psi(x,z) = \Phi(\xi(x), z) + \beta^3 F(x).$$

Let us fix a point $(x,z) \in \operatorname{int} \mathscr{Q}$ and choose an arbitrary direction $d = (h,v) \in E_1 \times E_3$. Define

$$\xi' = D\xi(x)[h], \quad \xi'' = D^2\xi(x)[h,h], \quad \xi''' = D^3\xi(x)[h,h,h], \quad l = (\xi', v).$$

Let $\psi(x,z) = \Phi(\xi(x), z)$. Consider the following directional derivatives:

$$\Delta_1 \overset{\mathrm{def}}{=} D\psi(x,z)[d] = \langle \nabla_y \Phi(\xi(x),z), \xi' \rangle + \langle \nabla_z \Phi(\xi(x),z), v \rangle = \langle \nabla \Phi(\xi(x),z), l \rangle.$$

Note that $l \equiv l(x)$. Therefore $l' \overset{\mathrm{def}}{=} Dl(x)[d] = (\xi'', 0) \overset{(5.4.20)}{\in} -K_0$. Thus, we can continue:

$$\begin{aligned} \Delta_2 \overset{\mathrm{def}}{=} D^2\psi(x,z)[d,d] \; &= \; \langle \nabla^2 \Phi(\xi(x),z) l, l \rangle + \langle \nabla \Phi(\xi(x),z), l' \rangle \\ &= \langle \nabla^2 \Phi(\xi(x),z) l, l \rangle + \langle \nabla_y \Phi(\xi(x),z), \xi'' \rangle \overset{\mathrm{def}}{=} \sigma_1 + \sigma_2. \end{aligned} \tag{5.4.24}$$

Since $-l'$ is a recession direction of $Q_2$, by (5.3.13) we have $\sigma_2 \geq 0$. Finally,

$$\begin{aligned} \Delta_3 &\overset{\text{def}}{=} D^3\psi(x,z)[d,d,d] \\ &= D^3\Phi(\xi(x),z)[l,l,l] + 3\langle \nabla^2\Phi(\xi(x),z)l,l'\rangle + \langle \nabla_y\Phi(\xi(x),z),\xi'''\rangle. \end{aligned} \tag{5.4.25}$$

Again, since $-l'$ is a recession direction of $Q_2$,

$$\begin{aligned} \langle \nabla^2\Phi(\xi(x),z)l,l'\rangle &\leq \langle \nabla^2\Phi(\xi(x),z)l,l\rangle^{1/2} \cdot \langle \nabla^2\Phi(\xi(x),z)l',l'\rangle^{1/2} \\ &\overset{(5.3.13)}{\leq} \langle \nabla^2\Phi(\xi(x),z)l,l\rangle^{1/2} \cdot \langle -\nabla\Phi(\xi(x),z),-l'\rangle = \sigma_1^{1/2}\sigma_2. \end{aligned}$$

Further, let $\sigma_3 = \langle \nabla^2 F(x)h,h\rangle$. Since $\xi$ is $\beta$-compatible with $F$ (see (5.4.22)), we have

$$\langle -\nabla_y\Phi(\xi(x),z),-\xi'''\rangle \overset{(5.4.23)}{\leq} 3\beta\langle -\nabla_y\Phi(\xi(x),z),-\xi''\rangle \cdot \sigma_3^{1/2} = 3\beta \cdot \sigma_2 \cdot \sigma_3^{1/2}.$$

Thus, substituting these inequalities into (5.4.25) and using (5.1.4), we obtain

$$\Delta_3 \leq 2\sigma_1^{3/2} + 3\sigma_1^{1/2}\sigma_2 + 3\beta \cdot \sigma_2 \cdot \sigma_3^{1/2}.$$

Consider now $D_k$, $k = 1\ldots 3$, the directional derivatives of the function $\Psi$. Note that

$$D_2 = \Delta_2 + \beta^3\sigma_3 = \sigma_1 + \sigma_2 + \beta^3\sigma_3 \geq \sigma_1 + \sigma_2 + \beta^2\sigma_3. \tag{5.4.26}$$

Therefore,

$$\begin{aligned} D_3 &= \Delta_3 + \beta^3 D^3F(x)[h,h,h] \overset{(5.1.4)}{\leq} \Delta_3 + 2\beta^3\sigma_3^{3/2} \\ &\leq 2\sigma_1^{3/2} + 3\sigma_1^{1/2}\sigma_2 + 3\beta \cdot \sigma_2 \cdot \sigma_3^{1/2} + 2\beta^3\sigma_3^{3/2} \\ &= (\sigma_1^{1/2} + \beta\sigma_3^{1/2})(2\sigma_1 - 2\beta\sigma_1^{1/2}\sigma_3^{1/2} + 2\beta^2\sigma_3 + 3\sigma_2) \\ &\overset{(5.4.26)}{\leq} (\sigma_1^{1/2} + \beta\sigma_3^{1/2})(3D_2 - (\sigma_1^{1/2} + \beta\sigma_3^{1/2})^2) \leq 2D_2^{3/2}. \end{aligned}$$

Thus, we come to the following statement.

**Theorem 5.4.4** *Let the function $\xi(\cdot) : E_1 \to E_2$ satisfy the following conditions.*

- *It is concave with respect to a convex cone $K \subset E_2$.*
- *It is $\beta$-compatible with self-concordant barrier $F(\cdot)$ for a set $Q \subseteq dom\, \xi$.*

*Assume in addition that $\Phi(\cdot, \cdot)$ is a $\mu$-self-concordant barrier for a closed convex set $Q_2 \subset E_2 \times E_3$, and the cone $K \times \{0\} \subset E_2 \times E_3$ contains only the recession directions of the set $Q_2$. Then the function*

$$\Psi(x, z) = \Phi(\xi(x), z) + \beta^3 F(x) \tag{5.4.27}$$

*is a self-concordant barrier for the set $\{(x, z) \in Q \times \mathbb{E}_3 : \exists y,\ \xi(x) \succeq_K y,\ (y, z) \in Q_2\}$ with barrier parameter $\hat{\nu} = \mu + \beta^3 \nu$.*

*Proof* We need to justify only the value of the barrier parameter $\hat{\nu}$. Indeed,

$$\begin{aligned} D_1 &= \langle \nabla\Phi(\xi(x), z), l\rangle + \beta^3 \langle \nabla F(x), h\rangle \ \le\ \sqrt{\nu} \cdot \sigma_1^{1/2} + \beta^3 \sqrt{\mu} \cdot \sigma_3^{1/2} \\ &\le \max_{\sigma_1, \sigma_3 \ge 0} \{\sqrt{\nu} \cdot \sigma_1^{1/2} + \beta^3 \cdot \sqrt{\mu}\sigma_3^{1/2} :\ \sigma_1 + \beta^3 \sigma_3 \overset{(5.4.26)}{\le} D_2\} \\ &= \sqrt{\hat{\nu}} \cdot D_2^{1/2}. \end{aligned}$$

It remains to use definition (5.3.6). □

Note that in construction (5.4.27) the function $\xi$ must be compatible only with the barrier $F$. The function $\Phi$ can be an arbitrary self-concordant barrier for the set $Q_2$.

## 5.4.7 Examples of Self-concordant Barriers

Despite its complicated formulation, Theorem 5.4.4 is very convenient for constructing a good self-concordant barrier for convex cones. Let us confirm this claim with several examples.

**1. The power cone and epigraph of the $p$-norm.** Let us fix some $\alpha \in (0, 1)$. To the description of the representation of the power cone

$$K_\alpha = \left\{(x^{(1)}, x^{(2)}, z) \in R^2_+ \times R :\ (x^{(1)})^\alpha \cdot (x^{(2)})^{1-\alpha} \ge |z|\right\},$$

given in Example 5.4.1, we need to add only a definition of the barrier function for the set $Q_2$. In view of Lemma 5.4.4, we can take

$$\Phi(y, z) = -\ln(y^2 - z^2),$$

with barrier parameter $\mu = 2$. Thus, all conditions of Theorem 5.4.4 are clearly satisfied except $\beta$-compatibility.

Let us prove that the function $\xi(x) = (x^{(1)})^\alpha \cdot (x^{(2)})^{1-\alpha}$ is $\beta$-comptible with barrier $F(x) = -\ln x^{(1)} - \ln x^{(2)}$. Let us choose a direction $h \in \mathbb{R}^2$ and $x \in \operatorname{int} \mathbb{R}^2_+$.

Define

$$\delta_1 = \tfrac{h^{(1)}}{x^{(1)}}, \quad \delta_2 = \tfrac{h^{(2)}}{x^{(2)}}, \quad \sigma = \delta_1^2 + \delta_2^2.$$

Let us compute the directional derivatives:

$$\begin{aligned} D\xi(x)[h] &= \left[\tfrac{\alpha h^{(1)}}{x^{(1)}} + \tfrac{(1-\alpha)h^{(2)}}{x^{(2)}}\right] \cdot \xi(x) \;=\; [\alpha\delta_1 + (1-\alpha)\delta_2] \cdot \xi(x), \\ D^2\xi(x)[h,h] &= -[\alpha\delta_1^2 + (1-\alpha)\delta_2^2] \cdot \xi(x) + [\alpha\delta_1 + (1-\alpha)\delta_2] \cdot D\xi(x)[h] \\ &= -\alpha(1-\alpha)(\delta_1 - \delta_2)^2 \cdot \xi(x), \\ D^3\xi(x)[h,h,h] &= 2\alpha(1-\alpha)(\delta_1-\delta_2)\cdot(\delta_1^2 - \delta_2^2)\cdot\xi(x) \\ &\quad -\alpha(1-\alpha)(\delta_1-\delta_2)^2 \cdot D\xi(x)[h] \\ &= \xi(x)\cdot\alpha(1-\alpha)(\delta_1-\delta_2)^2\cdot[2\delta_1 + 2\delta_2 - \alpha\delta_1 - (1-\alpha)\delta_2] \\ &= -D^2\xi(x)[h,h]\cdot[(2-\alpha)\delta_1 + (1+\alpha)\delta_2]. \end{aligned}$$

Since $(2-\alpha)\delta_1 + (1+\alpha)\delta_2 \leq [(2-\alpha)^2 + (1+\alpha)^2]^{1/2}\sigma^{1/2} < 3\sigma^{1/2}$, we conclude that $\xi$ is 1-compatible with $F$. Therefore, in view of Theorem 5.4.4, function

$$\Psi_P(x,z) = -\ln\left((x^{(1)})^{2\alpha}\cdot(x^{(2)})^{2(1-\alpha)} - z^2\right) - \ln x^{(1)} - \ln x^{(2)} \tag{5.4.28}$$

is a 4-self-concordant barrier for cone $K_\alpha$.

A similar structure can be used to construct a self-concordant barrier for the cone

$$K_\alpha^+ = \left\{(x^{(1)}, x^{(2)}, z) \in \mathbb{R}_+^2 \times \mathbb{R} : \; (x^{(1)})^\alpha \cdot (x^{(2)})^{1-\alpha} \geq z\right\}.$$

In this case, we can choose $\Phi(y,z) = \ln(y-z)$ with parameter $\mu = 1$. Thus, by Theorem 5.4.4, we get the following 3-self-concordant barrier:

$$\Psi_P^+(x,z) = -\ln\left((x^{(1)})^\alpha\cdot(x^{(2)})^{(1-\alpha)} - z\right) - \ln x^{(1)} - \ln x^{(2)}. \tag{5.4.29}$$

Let us show that this barrier has the best possible value of parameter.

**Lemma 5.4.9** *Any $\nu$-self-concordant barrier for the cone $K_\alpha^+$ has $\nu \geq 3$.*

*Proof* Note that the cone $K_\alpha^+$ has three recession directions:

$$p_1 = (1,0,0)^T, \quad p_2 \;=\; (0,1,0)^T, \quad p_3 \;=\; (0,0,-1)^T.$$

Let us choose a parameter $\tau > 0$ and define $\bar{x} = (1, 1, -\tau)^T$. Note that

$$\bar{x} - p_1 \notin \operatorname{int} K_\alpha^+, \quad \bar{x} - p_2 \notin \operatorname{int} K_\alpha^+, \quad \bar{x} - (1+\tau)p_3 \in \partial K_\alpha^+.$$

On the other hand, $\bar{x} - p_1 - p_2 - \tau p_3 = 0 \in K_\alpha^+$. Thus, to apply Theorem 5.4.1, we can choose

$$\alpha_1 = \alpha_2 = 1, \quad \alpha_3 = \tau, \quad \beta_1 = \beta_2 = 1, \quad \beta_3 = 1 + \tau.$$

Hence, $\nu \geq \sum\limits_{i=1}^{3} \frac{\alpha_i}{\beta_i} = 2 + \frac{\tau}{1+\tau}$. It remains to compute the limit as $\tau \to +\infty$. □

Note that the barrier $\Psi_P(x, z)$ can be used to construct $4n$-self-concordant barrier for the epigraph of an $\ell_p$-norm in $\mathbb{R}^n$:

$$\mathscr{K}_p = \left\{(\tau, z) \in \mathbb{R} \times \mathbb{R}^n : \tau \geq \|z\|_{(p)}\right\}, \quad 1 \leq p \leq \infty,$$

where $\|z\|_{(p)} = \left[\sum\limits_{i=1}^{n} |z^{(i)}|^p\right]^{1/p}$. Let us assume that $\alpha \overset{\text{def}}{=} \frac{1}{p} \in (0, 1)$. Then, it is easy to prove that the point $(\tau, z)$ belongs to $\mathscr{K}_p$ if and only if there exists an $x \in \mathbb{R}^n_+$ satisfying the conditions

$$\begin{aligned} (x^{(i)})^\alpha \cdot \tau^{1-\alpha} &\geq |z^{(i)}|, \quad i = 1, \ldots, n, \\ \sum_{i=1}^{n} x^{(i)} &= \tau. \end{aligned} \tag{5.4.30}$$

Thus, a self-concordant barrier for the cone $\mathscr{K}_p$ can be implemented by restricting the $(4n)$-self-concordant barrier

$$\Psi_\alpha(\tau, x, z) = -\sum_{i=1}^{n} \left[\ln\left((x^{(i)})^{2\alpha} \cdot \tau^{2(1-\alpha)} - (z^{(i)})^2\right) + \ln x^{(i)} + \ln \tau\right] \tag{5.4.31}$$

onto the hyperplane $\sum\limits_{i=1}^{n} x^{(i)} = \tau$.

**2. The conic hull of the epigraph of the entropy function**. We need to describe the conic hull of the following set:

$$\left\{(x^{(1)}, z) : z \geq x^{(1)} \ln x^{(1)}, \ x^{(1)} > 0\right\}.$$

Introducing a projective variable $x^{(2)} > 0$, we obtain the cone

$$\mathscr{Q} = \left\{(x^{(1)}, x^{(2)}, z) : z \geq x^{(1)} \cdot [\ \ln x^{(1)} - \ln x^{(2)}\ ], \ x^{(1)}, x^{(2)} > 0\right\}. \tag{5.4.32}$$

Let us represent it in the format of Theorem 5.4.4:

$$\mathbb{E}_1 = \mathbb{R}^2, \quad Q_1 = R^2_+, \quad F(x) = -\ln x^{(1)} - \ln x^{(2)}, \quad \nu = 2,$$

$$\mathbb{E}_2 = \mathbb{R}, \quad \xi(x) = -x^{(1)} \cdot [\,\ln x^{(1)} - \ln x^{(2)}\,], \quad K = \mathbb{R}_+,$$

$$\mathbb{E}_3 = \mathbb{R}, \quad Q_2 = \{(y,z) :\ y + z \geq 0\}, \quad \Phi(y,z) = -\ln(y+z), \quad \mu = 1.$$

Let us show that $\xi$ is 1-compatible with $F$. We use the notation of the previous example.

$$D\xi(x)[h] = \delta_1 \cdot \xi(x) - x^{(1)} \cdot [\delta_1 - \delta_2].$$

$$\begin{aligned} D^2\xi(x)[h,h] &= -\delta_1^2 \cdot \xi(x) + \delta_1 \cdot D\xi(x)[h] - h^{(1)} \cdot [\delta_1 - \delta_2] + x^{(1)} \cdot [\delta_1^2 - \delta_2^2] \\ &= x^{(1)} \cdot [-2\delta_1(\delta_1 - \delta_2) + \delta_1^2 - \delta_2^2] \;=\; -x^{(1)} \cdot (\delta_1 - \delta_2)^2. \end{aligned}$$

$$\begin{aligned} D^3\xi(x)[h,h,h] &= -h^{(1)} \cdot (\delta_1 - \delta_2)^2 + 2x^{(1)} \cdot (\delta_1 - \delta_2) \cdot (\delta_1^2 - \delta_2^2) \\ &= x^{(1)}(\delta_1 - \delta_2)^2 \cdot [-\delta_1 + 2(\delta_1 + \delta_2)] \\ &= -D^2\xi(x)[h,h] \cdot [\delta_1 + 2\delta_2]. \end{aligned}$$

Since $\delta_1 + 2\delta_2 \leq \sqrt{5} \cdot \sigma^{1/2} < 3\sigma^{1/2}$, we conclude that $\xi$ is 1-compatible with $F$. Therefore, in view of Theorem 5.4.4 the function

$$\Psi_E(x,z) = -\ln\left(z - x^{(1)} \cdot \ln \tfrac{x^{(1)}}{x^{(2)}}\right) - \ln x^{(1)} - \ln x^{(2)} \tag{5.4.33}$$

is a 3-self-concordant barrier for the cone $\mathscr{Q}$. It is interesting that the same barrier can also describe the epigraph of logarithmic and exponent functions. Indeed,

$$\mathscr{Q} \bigcap \{x : x^{(1)} = 1\} = \{(x^{(2)}, z) :\ z \geq -\ln x^{(2)}\} \;=\; \{(x^{(2)}, z) :\ x^{(2)} \geq e^{-z}\}.$$

Let us show that we can use the 3-self-concordant barrier

$$\begin{aligned} \psi_E(x,y,\tau) \;&=\; -\ln\left(\tau \ln \tfrac{y}{\tau} - x\right) - \ln y - \ln \tau, \\ (x,y,\tau) \in \operatorname{int}\mathscr{E} &\overset{\text{def}}{=} \left\{y \geq \tau e^{x/\tau},\ \tau > 0\right\} \subset \mathbb{R}^3, \end{aligned} \tag{5.4.34}$$

in more complicated situations. Consider the conic hull of the epigraph of the following function:

$$f_n(x) \overset{\text{def}}{=} \ln\left(\sum_{i=1}^{n} e^{x^{(i)}}\right), \quad x \in \mathbb{R}^n,$$
$$Q \overset{\text{def}}{=} \left\{(x,t,\tau) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} : \ t \ge \tau f_n\left(\tfrac{x}{\tau}\right), \ \tau > 0\right\}. \tag{5.4.35}$$

Clearly $(x, t, \tau) \in Q$ if and only if

$$f_n\left(\tfrac{1}{\tau}(x - t \cdot \bar{e}_n)\right) \le 1,$$

where $\bar{e}_n \in \mathbb{R}^n$ is the vector of all ones. Therefore, we can model $Q$ as a projection of the following cone:

$$\hat{Q} = \left\{(x,y,t,\tau) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} : y^{(i)} \ge \tau e^{(x^{(i)}-t)/\tau}, \ i = 1,\dots,n, \right.$$
$$\left. \sum_{i=1}^{n} y^{(i)} \ = \ \tau \right\}.$$

This cone admits a $3n$-self-concordant barrier, obtained as a restriction of the function

$$\Psi_L(x,y,t,\tau) = -\sum_{i=1}^{n}\left[\ln\left(t + \tau \ln y^{(i)} - x^{(i)} - \tau \ln \tau\right) + \ln y^{(i)} + \ln \tau\right], \tag{5.4.36}$$

onto the hyperplane $\sum\limits_{i=1}^{n} y^{(i)} \ = \ \tau$.

**3. The geometric mean.** Let $x \in \mathbb{R}^n_+$ and $a \in \Delta_n \overset{\text{def}}{=} \left\{ y \in \mathbb{R}^n_+ : \ \sum\limits_{i=1}^{n} y^{(i)} = 1 \right\}$. Without loss of generality, we can consider $a$ with positive components. Define

$$\xi(x) = x^a \ \overset{\text{def}}{=} \ \prod_{i=1}^{n} (x^{(i)})^{a^{(i)}}.$$

Let us write down the directional derivatives of this function along some $h \in \mathbb{R}^n$. Define

$$\delta_x^{(i)}(h) = \tfrac{h^{(i)}}{x^{(i)}}, \ i = 1,\dots,n,$$

$$\delta_x(h) = \left(\delta_x^{(1)}(h), \dots, \delta_x^{(n)}(h)\right)^T,$$

$$F(x) = -\sum_{i=1}^{n} \ln x^{(i)}.$$

Clearly, $\|h\|_x \overset{\rm def}{=} \langle F''(x)h, h\rangle^{1/2} = \|\delta_x(h)\|$, where the norm is standard Euclidean. Note that

$$D(\ln \xi(x))[h] = \tfrac{1}{\xi(x)} D\xi(x)[h] \;=\; \langle a, \delta_x(h)\rangle.$$

Thus, $D\xi(x)[h] \;=\; \xi(x) \cdot \langle a, \delta_x(h)\rangle$. Denoting by $[x]^k \in \mathbb{R}^n$ a component-wise power of a vector $x \in \mathbb{R}^n$, we obtain:

$$D^2\xi(x)[h,h] = \xi(x) \cdot \langle a, \delta_x(h)\rangle^2 - \xi(x) \cdot \langle a, [\delta_x(h)]^2\rangle$$

$$= -\xi(x) \cdot \langle a, [\delta_x(h) - \langle a, \delta_x(h)\rangle \cdot \bar{e}_n]^2\rangle \;\overset{\rm def}{=}\; -\xi(x) \cdot S_2.$$

Further, defining $\xi = \xi(x)$ and $\delta = \delta_x(h)$, we obtain:

$$D^3\xi(x)[h,h,h] = \xi\langle a, \delta\rangle^3 + 2\xi\langle a, \delta\rangle\langle a, -[\delta]^2\rangle - \xi\langle a, \delta\rangle\langle a, [\delta]^2\rangle - \xi\langle a, -2[\delta]^3\rangle$$

$$= \xi\left(\langle a, \delta\rangle^3 - 3\langle a, \delta\rangle\langle a, [\delta]^2\rangle + 2\langle a, [\delta]^3\rangle\right).$$

Define

$$S_3 = \langle a, [\delta - \langle a, \delta\rangle\bar{e}_n]^3\rangle \;=\; \langle a, [\delta]^3 - 3\langle a, \delta\rangle[\delta]^2 + 3\langle a, \delta\rangle^2\delta - \langle a, \delta\rangle^3\bar{e}_n\rangle$$

$$= \langle a, [\delta]^3\rangle - 3\langle a, \delta\rangle\langle a, [\delta]^2\rangle + 2\langle a, \delta\rangle^3.$$

Then, in this new notation we have

$$D^3\xi(x)[h,h,h] = \xi\Big(\langle a, \delta\rangle^3 - 3\langle a, \delta\rangle\langle a, [\delta]^2\rangle$$

$$+2\left[S_3 + 3\langle a, \delta\rangle\langle a, [\delta]^2\rangle - 2\langle a, \delta\rangle^3\rangle\right]\Big)$$

$$= \xi\left(2S_3 + 3\langle a, \delta\rangle\langle a, [\delta]^2\rangle - 3\langle a, \delta\rangle^3\right) \;=\; \xi(2S_3 + 3\langle a, \delta\rangle S_2).$$

Therefore,

$$D^3\xi(x)[h,h,h] \le \xi S_2\left(3\langle a, \delta\rangle + 2\max_{1\le i\le n}[\delta^{(i)} - \langle a, \delta\rangle]\right)$$

$$\le \xi S_2\left(\langle a, \delta\rangle + 2\max_{1\le i\le n}|\delta^{(i)}|\right)$$

$$\le -3D^2\xi(x)[h,h] \cdot \langle F''(x)\delta, \delta\rangle^{1/2}.$$

Thus, we have proved that $\xi$ is 1-compatible with $F$. This means that the function

$$\Psi(x,t) = -\ln(\xi(x) - t) + F(x), \quad x > 0 \in R^n, \tag{5.4.37}$$

is an $(n+1)$-self-concordant barrier for the hypograph of the function $\xi$. Moreover, since the set of $\beta$-compatible functions is a convex cone, any sum

$$\xi(x) = \sum_{k=1}^{m} \alpha_k x^{a_k}, \tag{5.4.38}$$

with $\alpha_k > 0$, and $a_k \in \Delta_n$, $k = 1, \ldots, m$, is 1-compatible with $F$. Hence, for such functions formula (5.4.37) is also applicable and the parameter of this barrier remains equal to $n + 1$.

Note that the functions in the form (5.4.38) sometimes arise in optimization problems related to polynomials. Indeed, assume we need to solve the problem

$$\max_y \left\{ p(y) = \sum_{k=1}^{m} \alpha_k y^{b_k} : \ y \geq 0, \ \|y\|_{(d)} \leq 1 \right\},$$

where all $b_k$ belong to $d \cdot \Delta_n$ and $\|y\|_{(d)} = \left[ \sum\limits_{i=1}^{n} (y^{(i)})^d \right]^{1/d}$. Then for new variables $y^{(i)} = \left[x^{(i)}\right]^{1/d}$, $i = 1, \ldots, n$, our problem becomes convex with a concave objective $\xi(\cdot)$ given by (5.4.38).

**4. The hypograph of the exponent of the self-concordant barrier.** Let $F(\cdot)$ be $\nu$-self-concordant barrier for the set Dom $F$. Let us fix $p \geq \nu$ and consider the function $\xi_p(x) = \exp\left\{-\frac{1}{p}F(x)\right\}$. As we have proved in Lemma 5.3.1, this function is concave on dom $F$. Consider the following set:

$$\mathscr{H}_p = \left\{(x,t) \in \operatorname{dom} F \times \mathbb{R} : \ \xi_p(x) \geq t\right\}.$$

Let us construct a self-concordant barrier for this set.

In our framework, $Q_1 = \operatorname{Dom} F$, $Q_2 = \{(y,t) \in \mathbb{R}^2 : \ y \geq t\}$, $K = \mathbb{R}_+$, and $\Phi(y,t) = -\ln(y - t)$ with $\mu = 1$. Let us prove that $\xi_p(x)$ is concave with respect to $K$, and it is $\beta$-compatible with $F$.

Let us fix $x \in \operatorname{dom} F$ and an arbitrary direction $h \in \mathbb{E}$. Then

$$\begin{aligned}
\xi' &\overset{\text{def}}{=} D\xi_p(x)[h] \ = \ -\tfrac{1}{p}\langle \nabla F(x), h\rangle \xi_p(x), \\
\xi'' &\overset{\text{def}}{=} D^2F(x)[h,h] \ = \ \tfrac{1}{p^2}\langle \nabla F(x), h\rangle^2 \xi_p(x) - \tfrac{1}{p}\langle \nabla^2 F(x)h, h\rangle \xi_p(x), \\
\xi''' &\overset{\text{def}}{=} D^3F(x)[h,h,h] \ = \ -\tfrac{1}{p^3}\langle \nabla F(x), h\rangle^3 \xi_p(x) \\
&\qquad + \tfrac{3}{p^2}\langle \nabla F(x), h\rangle \cdot \langle \nabla^2 F(x)h, h\rangle \xi_p(x) - \tfrac{1}{p} D^3F(x)[h,h,h]\xi_p(x).
\end{aligned}$$

As we have already seen, in view of (5.3.6), we have $\xi'' \leq 0$. This means that it is concave with respect to $K$.

Let $\xi = \xi_p(x)$, $D_1 = \langle \nabla F(x), h \rangle$, $D_2 = \langle \nabla^2 F(x) h, h \rangle^{1/2}$, and $\tau = \frac{\xi}{p} D_2^2$. Then

$$\begin{aligned} \xi'' &= \tfrac{\xi}{p^2} D_1^2 - \tau \leq 0, \\ \xi''' &\overset{(5.1.4)}{\leq} \tfrac{2\xi}{p} D_2^3 + \tfrac{3\xi}{p^2} D_1 D_2^2 - \tfrac{\xi}{p^3} D_1^3 = 2\tau D_2 + \tfrac{1}{p} D_1 \left( 3\tau - \tfrac{\xi}{p^2} D_1^2 \right) \\ &= 2\tau D_2 + \tfrac{1}{p} D_1 \left( 2\tau - \xi'' \right) \overset{(5.3.6)}{\leq} 2\tau D_2 + \tfrac{\sqrt{\nu}}{p} D_2 \left( 2\tau - \xi'' \right). \end{aligned}$$

Note that $\xi'' + \tau = \frac{\xi}{p^2} D_1^2 \overset{(5.3.6)}{\leq} \frac{\xi \nu}{p^2} D_2^2 = \frac{\nu}{p} \tau$. Thus, $\tau \leq \frac{p}{p-\nu}(-\xi'')$, and therefore

$$\xi''' \leq D_2 \left( 2(1 + \tfrac{\sqrt{\nu}}{p})\tau + \tfrac{\sqrt{\nu}}{p}(-\xi'') \right) \leq D_2 \left( \tfrac{2}{\sqrt{p} - \sqrt{\nu}} + \tfrac{\sqrt{\nu}}{p} \right) (-\xi'').$$

This means that for $p \geq (1 + \sqrt{\nu})^2$ the function $\xi_p(x)$ is 1-compatible with $F$ and by Theorem 5.4.4 we get a $(\nu + 1)$-self-concordant barrier

$$\Psi_H(x, t) = -\ln \left( \exp \left\{ -\tfrac{1}{p} F(x) \right\} - t \right) + F(x) \tag{5.4.39}$$

for the set $\mathscr{H}_p$.

**5. The matrix epigraph of the inverse matrix.** Consider the following set

$$\mathscr{I}_n = \{ (X, Y) \in \mathbb{S}^n_+ \times \mathbb{S}^n_+ : \; X^{-1} \preceq Y \}.$$

In order to construct a barrier for this set, consider the mapping $\xi(X) = -X^{-1}$. It is defined on the set of positive definite matrices, for which we know a $\nu$-self-concordant barrier $F(X) = -\ln \det X$ with the barrier parameter $\nu = n$ (see Theorem 5.4.3). Let us show that $\xi$ is 1-compatible with $F$.

Indeed, let us fix an arbitrary direction $H \in \mathbb{S}^n$. By the same reasoning as in Lemma 5.4.6, we can prove that

$$\begin{aligned} D\xi(X)[H] &= X^{-1} H X^{-1}, \\ D^2\xi(X)[H, H] &= -2X^{-1} H X^{-1} H X^{-1} \in -S^n_+, \\ D^3\xi(X)[H, H, H] &= 6X^{-1} H X^{-1} H X^{-1} H X^{-1}. \end{aligned}$$

Let $A = X^{-1/2}HX^{-1/2}$ and $\rho = \max\limits_{1\le i\le n} |\lambda_i(A)|$. Then, in view of Lemma 5.4.6,

$$\langle \nabla^2 F(X)H, H\rangle = \|A\|_F^2 \;=\; \sum_{i=1}^n \lambda_i^2(A) \;\ge\; \rho^2.$$

On the other hand,

$$\begin{aligned} D^3\xi(X)[H,H,H] &= 6X^{-1/2}A^3X^{-1/2} \;\preceq\; 6\rho X^{-1/2}A^2X^{-1/2} \\ &\preceq 6\langle \nabla^2 F(X)H, H\rangle^{1/2} X^{-1/2}A^2X^{-1/2} \\ &= 3\langle \nabla^2 F(X)H, H\rangle^{1/2} D^2F(X)[H,H]. \end{aligned}$$

Thus, condition (5.4.21) is satisfied with $\beta = 1$. Hence, by Theorem 5.4.4 the function

$$F(X,Y) = -\ln\det(Y - X^{-1}) - \ln\det X \tag{5.4.40}$$

is a $\nu$-self-concordant barrier for $\mathscr{I}_n$ with $\nu = 2n$.

**Lemma 5.4.10** *Any self-concordant barrier for the set $\mathscr{I}_n$ has parameter $\nu \ge 2n$.*

*Proof* Let us choose $\gamma > 1$ and consider matrices $\bar X = \bar Y = \gamma I_n$. Clearly the point $(\bar X, \bar Y)$ belongs to int $\mathscr{I}_n$. Note that for positive definite matrices, relation $Y \succeq X^{-1}$ holds if and only if $X \succeq Y^{-1}$. Therefore, all directions

$$p_i = (e_ie_i^T, 0), \quad q_i \;=\; (0, e_ie_i^T), \quad i = 1,\dots,n,$$

are recession directions of the set $\mathscr{I}_n$. It is easy to check that for $\beta = \gamma - \frac{1}{\gamma}$ we get

$$(\bar X, \bar Y) - \beta p_i \in \partial\mathscr{I}_n, \quad (\bar X, \bar Y) - \beta q_i \;\in \partial\mathscr{I}_n, \quad i = 1,\dots,n.$$

On the other hand, for $\alpha = \gamma - 1$, we have $\bar Y - \alpha\sum\limits_{i=1}^n e_ie_i^T = I_n = (\bar X - \alpha\sum\limits_{i=1}^n e_ie_i^T)^{-1}$. Therefore, in the conditions of Theorem 5.4.1 we can get all $\alpha_i = \alpha$ and all $\beta_i = \beta$. Thus, we obtain $\nu \ge 2n\frac{\alpha}{\beta} = \frac{2n\gamma}{1+\gamma}$. Since $\gamma$ can be arbitrarily big, we come to the bound $\nu \ge 2n$. □

### 5.4.8 Separable Optimization

In problems of Separable Optimization all nonlinear terms in functional components are represented by univariate functions. A general formulation of such a problem is as follows:

$$
\begin{aligned}
\min_{x \in \mathbb{R}^n} \Big\{ \, q_0(x) &= \sum_{j=1}^{m_0} \alpha_{0,j} f_{0,j}(\langle a_{0,j}, x\rangle + b_{0,j}), \\
q_i(x) &= \sum_{j=1}^{m_i} \alpha_{i,j} f_{i,j}(\langle a_{i,j}, x\rangle + b_{i,j}) \le \beta_i, \; i = 1 \dots m \Big\},
\end{aligned}
\tag{5.4.41}
$$

where $\alpha_{i,j}$ are some positive coefficients, $a_{i,j} \in \mathbb{R}^n$ and $f_{i,j}(\cdot)$ are convex functions of one variable. Let us rewrite this problem in the standard form:

$$
\begin{aligned}
\min_{x \in \mathbb{R}^n, \tau \in \mathbb{R}^{m+1}, t \in \mathbb{R}^M} \Big\{ \, \tau_0 : \; & \sum_{j=1}^{m_i} \alpha_{i,j} t_{i,j} \le \tau_i, \; i = 0 \dots m, \; \tau_i \le \beta_i, \; i = 1 \dots m, \\
& f_{i,j}(\langle a_{i,j}, x\rangle + b_{i,j}) \le t_{i,j}, \; j = 1 \dots m_i, \; i = 0 \dots m, \; \Big\},
\end{aligned}
\tag{5.4.42}
$$

where $M = \sum_{i=0}^{m} m_i$. Thus, in order to construct a self-concordant barrier for the feasible set of this problem, we need barriers for epigraphs of univariate convex functions $f_{i,j}$. Let us point out such barriers for several important examples.

#### 5.4.8.1 Logarithm and Exponent

By fixing the first coordinate in the barrier (5.4.33), we obtain the barrier function $F_1(x, t) = -\ln x - \ln(\ln x + t)$, which is a 3-self-concordant barrier for the set

$$
Q_1 = \{(x, t) \in \mathbb{R}^2 \mid \; x > 0, \; t \ge -\ln x\}.
$$

Similarly, we obtain the function $F_2(x, t) = -\ln t - \ln(\ln t - x)$ as a 3-self-concordant barrier for the set

$$
Q_2 = \{(x, t) \in \mathbb{R}^2 \mid \; t \ge e^x\}.
$$

#### 5.4.8.2 Entropy Function

By fixing the second coordinate in the barrier (5.4.33), we obtain the barrier function $F_3(x, t) = -\ln x - \ln(t - x \ln x)$, which is a 3-self-concordant barrier for the set

$$
Q_3 = \{(x, t) \in \mathbb{R}^2 \mid \; x \ge 0, \; t \ge x \ln x\}.
$$

#### 5.4.8.3 Increasing Power Functions

Let $p \geq 1$ and define $\alpha = \frac{1}{p}$. By fixing the second variable in barrier (5.4.28), $x^{(2)} = 1$, we get function $F_4(x,t) = -\ln t - \ln(t^{2/p} - x^2)$, which is a 4-self-concordant barrier for the set

$$Q_4 = \{(x,t) \in \mathbb{R}^2 \mid \ t \geq \mid x \mid^p\}, \quad p \geq 1.$$

If $p < 1$, then a similar operation with the barrier (5.4.29) gives us the function $F_5(x,t) = -\ln t - \ln(t^p - x)$, which is a 3-self-concordant barrier for the set

$$Q_5 = \{(x,t) \in \mathbb{R}^2 \mid \ t \geq 0, \ t^p \geq x\}, \quad 0 < p \leq 1.$$

#### 5.4.8.4 Decreasing Power Functions

Let $p > 0$. Define $\alpha = \frac{p}{p+1}$. Then by fixing $z = 1$ in the barrier (5.4.29), we get the function $F_6(x,t) = -\ln x - \ln t - \ln(x^\alpha t^{1-\alpha} - 1)$, which is a 3-self-concordant barrier for the set

$$Q_6 = \left\{(x,t) \in \mathbb{R}^2 \mid \ x > 0, \ t \geq \tfrac{1}{x^p}\right\}.$$

Let us conclude our discussion with two examples.

#### 5.4.8.5 Geometric Optimization

The initial formulation of such problems is as follows:

$$\begin{aligned} \min_{x \in \mathbb{R}^n_{++}} \Big\{ \ q_0(x) &= \sum_{j=1}^{m_0} \alpha_{0,j} \prod_{j=1}^{n} (x^{(j)})^{\sigma_{0,j}^{(j)}}, \\ q_i(x) &= \sum_{j=1}^{m_i} \alpha_{i,j} \prod_{j=1}^{n} (x^{(j)})^{\sigma_{i,j}^{(j)}} \leq 1, \ i = 1 \ldots m \Big\}, \end{aligned} \tag{5.4.43}$$

where $\mathbb{R}^N_{++}$ is the interior of the positive orthant, and $\alpha_{i,j}$ are some positive coefficients. Note that the problem (5.4.43) is not convex.

Let us introduce vectors $a_{i,j} = (\sigma_{i,j}^{(1)}, \ldots, \sigma_{i,j}^{(n)}) \in \mathbb{R}^n$, and change variables:

$$x^{(i)} = e^{y^{(i)}}, \quad i = 1, \ldots, n.$$

Then problem (5.4.43) can be written in a *convex* form.

$$\min_{y\in\mathbb{R}^n}\left\{\sum_{j=1}^{m_0}\alpha_{0,j}\exp(\langle a_{0,j},y\rangle):\ \sum_{j=1}^{m_i}\alpha_{i,j}\exp(\langle a_{i,j},y\rangle)\leq 1,\ i=1\dots m\right\}. \tag{5.4.44}$$

Let $M=\sum\limits_{i=0}^{m}m_i$. The complexity of solving (5.4.44) by a path-following scheme is $O\left(M^{1/2}\cdot\ln\frac{M}{\epsilon}\right)$ iterations.

#### 5.4.8.6 Approximation in an $\ell_p$-Norm

The simplest problem of this type is as follows:

$$\min_{x\in\mathbb{R}^n}\left\{\sum_{i=1}^{m}\mid\langle a_i,x\rangle-b^{(i)}\mid^p:\ \alpha\leq x\leq\beta\right\}, \tag{5.4.45}$$

where $p\geq 1$ and $\alpha,\beta\in\mathbb{R}^n$. Clearly, we can rewrite this problem in an equivalent standard form:

$$\begin{aligned}\min_{x\in\mathbb{R}^n,\tau\in\mathbb{R}^{m+1}}&\Big\{\tau^{(0)}:\mid\langle a_i,x\rangle-b^{(i)}\mid^p\leq\tau^{(i)},\ i=1\dots m,\\ &\sum_{i=1}^{m}\tau^{(i)}\leq\tau^{(0)},\ \alpha\leq x\leq\beta\Big\}.\end{aligned} \tag{5.4.46}$$

The complexity bound of this problem is $O\left(\sqrt{m+n}\cdot\ln\frac{m+n}{\epsilon}\right)$ iterations of a path-following scheme.

We have discussed the performance of Interior-Point Methods for several *pure* optimization problems. However, it is important that we can apply these methods to *mixed* problems. For example, in problems (5.4.11) or (5.4.45) we can also treat the quadratic constraints. To do this, we need to construct a corresponding self-concordant barrier. Such barriers are known for all important functional components arising in practical applications.

### *5.4.9 Choice of Minimization Scheme*

We have seen that the majority of convex optimization problems can be solved by Interior-Point Methods. However, the same problems can also be solved by methods of Nonsmooth Optimization. In general, we cannot say which approach is better, since the answer depends on the individual structure of a particular problem. However, the complexity estimates for optimization schemes are often helpful in making the choice. Let us consider a simple example.

Assume we are going to solve a problem of finding the best approximation in an $\ell_p$-norm:

$$\min_{x \in \mathbb{R}^n} \left\{ \sum_{i=1}^{m} \mid \langle a_i, x \rangle - b^{(i)} \mid^p : \ \alpha \leq x \leq \beta \right\}, \tag{5.4.47}$$

where $p \geq 1$. We have two available numerical methods:

- The Ellipsoid Method (Sect. 3.2.8).
- The Interior-Point Path-Following Scheme.

Which of them should we use? The answer can be derived from the complexity analysis of the corresponding schemes.

Firstly, let us estimate the performance of the Ellipsoid Method as applied to problem (5.4.47).

**Complexity of the Ellipsoid Method**

| | |
|---|---|
| Number of iterations: | $O\left(n^2 \ln \frac{1}{\epsilon}\right)$, |
| Complexity of the oracle: | $O(mn)$ operations, |
| Complexity of the iteration: | $O(n^2)$ operations. |
| **Total complexity**: | $O\left(n^3 (m+n) \ln \frac{1}{\epsilon}\right)$ operations. |

The analysis of the Path-Following Method is more involved. First of all, we should form a barrier model of the problem:

$$\begin{aligned} \min_{x \in \mathbb{R}^n, \tau \in \mathbb{R}^m, \xi \in \mathbb{R}} & \left\{ \xi \ : \ |\langle a_i, x \rangle - b^{(i)}|^p \leq \tau^{(i)}, \ i = 1 \dots m, \right. \\ & \left. \sum_{i=1}^{m} \tau^{(i)} \leq \xi, \ \alpha \leq x \leq \beta \right\}, \\ F(x, \tau, \xi)) &= \sum_{i=1}^{m} f(\langle a_i, x \rangle - b^{(i)}, \tau^{(i)}) - \ln(\xi - \sum_{i=1}^{m} \tau^{(i)}) \\ & \quad - \sum_{i=1}^{n} [\ln(x^{(i)} - \alpha^{(i)}) + \ln(\beta^{(i)} - x^{(i)})], \end{aligned} \tag{5.4.48}$$

where $f(y, t) = -\ln t - \ln(t^{2/p} - y^2)$.

We have seen that the parameter of barrier $F(x,\tau,\xi)$ is $\nu = 4m+n+1$. Therefore, the Path-Following Scheme needs $O\left(\sqrt{4m+n+1}\cdot \ln\frac{m+n}{\epsilon}\right)$ iterations at most.

At each iteration of this method, we need to compute the gradient and the Hessian of barrier $F(x,\tau,\xi)$. Define

$$g_1(y,t) = \nabla_y f(y,t), \quad g_2(y,t) = f'_t(y,t).$$

Then

$$\nabla_x F(x,\tau,\xi) = \sum_{i=1}^m g_1(\langle a_i,x\rangle - b^{(i)}, \tau^{(i)})a_i - \sum_{i=1}^n \left[\tfrac{1}{x^{(i)}-\alpha^{(i)}} - \tfrac{1}{\beta^{(i)}-x^{(i)}}\right] e_i,$$

$$F'_{\tau^{(i)}}(x,\tau,\xi) = g_2(\langle a_i,x\rangle - b^{(i)}, \tau^{(i)}) + \left[\xi - \sum_{j=1}^m \tau^{(j)}\right]^{-1},$$

$$F'_\xi(x,\tau,\xi) = -\left[\xi - \sum_{i=1}^m \tau^{(i)}\right]^{-1}.$$

Further, defining

$$h_{11}(y,t) = \nabla^2_{yy}F(y,t), \quad h_{12}(y,t) = \nabla^2_{yt}F(y,t), \quad h_{22}(y,t) = F''_{tt}(y,t),$$

we obtain

$$\begin{aligned}\nabla^2_{xx}F(x,\tau,\xi) &= \sum_{i=1}^m h_{11}(\langle a_i,x\rangle - b^{(i)}, \tau^{(i)})a_i a_i^T \\ &\quad +\text{diag}\left[\tfrac{1}{(x^{(i)}-\alpha^{(i)})^2} + \tfrac{1}{(\beta^{(i)}-x^{(i)})^2}\right],\end{aligned}$$

$$\nabla^2_{\tau^{(i)}x}F(x,\tau,\xi) = h_{12}(\langle a_i,x\rangle - b^{(i)}, \tau^{(i)})a_i,$$

$$F''_{\tau^{(i)},\tau^{(i)}}(x,\tau,\xi) = h_{22}(\langle a_i,x\rangle - b^{(i)}, \tau^{(i)}) + \left(\xi - \sum_{i=1}^m \tau^{(i)}\right)^{-2},$$

$$F''_{\tau^{(i)},\tau^{(j)}}(x,\tau,\xi) = \left(\xi - \sum_{i=1}^m \tau^{(i)}\right)^{-2}, \ i \neq j,$$

$$\nabla^2_{x,\xi}F(x,\tau,\xi) = 0, \quad F''_{\tau^{(i)},\xi}(x,\tau,\xi) \ = \ -\left(\xi - \sum_{i=1}^m \tau^{(i)}\right)^{-2},$$

$$F''_{\xi,\xi}(x,\tau,\xi) = \left(\xi - \sum_{i=1}^m \tau^{(i)}\right)^{-2}.$$

Thus, the complexity of the second-order oracle in the Path-Following Scheme is $O(mn^2)$ arithmetic operations.

Let us estimate now the complexity of each iteration. The main source of computations at each iteration is the solution of the Newton system. Let

$$\varkappa = \left(\xi - \sum_{i=1}^{m} \tau^{(i)}\right)^{-2}, \quad s_i = \langle a_i, x\rangle - b^{(i)}, \ i = 1 \dots n,$$

and

$$\Lambda_0 = \mathrm{diag}\left[\tfrac{1}{(x^{(i)}-\alpha^{(i)})^2} + \tfrac{1}{(\beta^{(i)}-x^{(i)})^2}\right]_{i=1}^{n} \ \Lambda_1 = \mathrm{diag}\,(h_{11}(s_i, \tau^{(i)}))_{i=1}^{m},$$

$$\Lambda_2 = \mathrm{diag}\,(h_{12}(s_i, \tau^{(i)}))_{i=1}^{m}, \qquad D = \mathrm{diag}\,(h_{22}(s_i, \tau^{(i)}))_{i=1}^{m}.$$

Then, using the notation $A = (a_1, \dots, a_m)$, $\bar{e}_m = (1, \dots, 1) \in \mathbb{R}^m$, the Newton system can be written in the following form:

$$\begin{aligned} [A(\Lambda_0 + \Lambda_1)A^T]\Delta x + A\Lambda_2\Delta\tau &= \nabla_x F(x, \tau, \xi), \\ \Lambda_2 A^T \Delta x + [D + \varkappa I_m]\Delta\tau + \varkappa\bar{e}_m\Delta\xi &= F'_\tau(x, \tau, \xi), \\ \varkappa\langle\bar{e}_m, \Delta\tau\rangle + \varkappa\Delta\xi &= F'_\xi(x, \tau, \xi) + t, \end{aligned} \tag{5.4.49}$$

where $t$ is the penalty parameter. From the second equation in (5.4.49), we obtain

$$\Delta\tau = [D + \varkappa I_m]^{-1}(F'_\tau(x, \tau, \xi) - \Lambda_2 A^T \Delta x - \varkappa\bar{e}_m\Delta\xi).$$

Substituting $\Delta\tau$ into the first equation in (5.4.49), we have

$$\begin{aligned} \Delta x = [A(\Lambda_0 + \Lambda_1 - \Lambda_2^2[D + \varkappa I_m]^{-1})A^T]^{-1}\{\nabla_x F(x, \tau, \xi) \\ -A\Lambda_2[D + \varkappa I_m]^{-1}(F'_\tau(x, \tau, \xi) - \varkappa\bar{e}_m\Delta\xi)\}. \end{aligned}$$

Using these relations, we can find $\Delta\xi$ from the last equation in (5.4.49).

Thus, the Newton system (5.4.49) can be solved in $O(n^3 + mn^2)$ operations. This implies that the total complexity of the Path-Following Scheme can be estimated as

$$O\left(n^2(m+n)^{3/2} \cdot \ln \tfrac{m+n}{\epsilon}\right)$$

arithmetic operations. Comparing this estimate with the bound for the Ellipsoid Method, we conclude that the Interior-Point Method is more efficient if $m$ is not too big, namely, if $m \le O(n^2)$.

Of course, this analysis is valid only if the methods behave in accordance with their worst-case complexity bounds. For the Ellipsoid Method this is indeed true. However, Interior-Point Path-Following Schemes can be accelerated by long-step strategies. The explanation of these abilities requires the introduction of a primal-dual setting of the optimization problems, posed in a conic form. Because of the volume constraints, we have decided not to touch on this deep theory in the present book.

# Chapter 6
# The Primal-Dual Model of an Objective Function

In the previous chapters, we have proved that in the Black-Box framework the non-smooth optimization problems are much more difficult than the smooth ones. However, very often we know the explicit structure of the functional components. In this chapter we show how this knowledge can be used to accelerate the minimization methods and to extract a useful information about the dual counterpart of the problem. The main acceleration idea is based on the approximation of a nondifferentiable function by a differentiable one. We develop a technique for creating computable smoothed versions of non-differentiable functions and minimize them by Fast Gradient Methods. 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. This technique can be used either in the primal form, or in the symmetric primal-dual form. We include in this chapter an example of application of this approach to the problem of Semidefinite Optimization. The chapter is concluded by analysis of performance of the Conditional Gradient method, which is based only on solving at each iteration an auxiliary problem of minimization of a linear function. We show that this method can also reconstruct the primal-dual solution of the problem. A similar idea is used in the second-order Trust Region Method with contraction, the first method of this type with provable global worst-case performance guarantees.

## 6.1 Smoothing for an Explicit Model of an Objective Function

(The minimax model of non-differentiable objective functions; The Fast Gradient Method for arbitrary norms and composite objective function; Application examples: minimax strategies for matrix games, the continuous location problem, variational inequalities with linear operator, minimization of piece-wise linear functions; Implementation issues.)

Y. Nesterov, *Lectures on Convex Optimization*, Springer Optimization
and Its Applications 137, https://doi.org/10.1007/978-3-319-91578-4_6

### 6.1.1 Smooth Approximations of Non-differentiable Functions

As we have seen in Chap. 3, subgradient methods solve the problem of Nonsmooth Convex Optimization in

$$O\left(\frac{1}{\epsilon^2}\right) \tag{6.1.1}$$

calls of the oracle, where $\epsilon$ is the desired absolute accuracy of finding the approximate solution in the function value. Moreover, we have already seen that the efficiency bound of the simplest Subgradient Method *cannot* be improved uniformly in the dimension of the space of variables (see Sect. 3.2). Of course, this statement is valid only for a Black-Box model of the objective function. However, the proof is constructive: it can be shown that the simplest problems like

$$\min_{x \in \mathbb{R}^n} \left\{ \gamma \max_{1 \le i \le k} x^{(i)} + \frac{\mu}{2} \|x\|^2 \right\}, \quad 1 \le k \le n,$$

where the norm is standard Euclidean, are difficult for all numerical schemes. The extremal simplicity of these functions possibly explains a common pessimistic belief that the actual worst-case complexity bound for finding an $\epsilon$-approximation of the minimal value of a piece-wise linear function by gradient schemes is indeed given by (6.1.1).

In fact, this is not absolutely true. In practice, we almost never meet a pure Black-Box model. We always know something about the structure of the underlying objects (we have already discussed this in Sect. 5.1.1), and the proper use of this structure can and does help in constructing more efficient schemes.

In this section, we discuss one such possibility based on constructing a smooth approximation of a nonsmooth function. Let us look at the following situation. Consider a function $f$ which is convex on $\mathbb{E}$. Assume that $f$ satisfies the following growth condition:

$$f(x) \le f(0) + L\|x\|, \quad \forall x \in \mathbb{R}^n, \tag{6.1.2}$$

where the Euclidean norm $\|x\| = \langle Bx, x \rangle^{1/2}$ is defined by a self-adjoint positive definite linear operator $B : \mathbb{E} \to \mathbb{E}^*$. Define the *Fenchel conjugate* of the function $f$ as follows:

$$f_*(s) = \sup_{x \in \mathbb{E}} [\langle s, x \rangle - f(x)], \quad s \in \mathbb{E}^*. \tag{6.1.3}$$

Clearly, this function is closed and convex in view of Theorem 3.1.8. Its domain is not empty since by Theorem 3.1.20

$$\operatorname{dom} f_* \supseteq \partial f(x), \quad \forall x \in \mathbb{E}.$$

At the same time, dom $f_*$ is bounded:

$$\|s\| \overset{(6.1.2)}{\leq} L \quad \forall s \in \text{dom}\, f_*. \tag{6.1.4}$$

Note that for all $x \in \mathbb{E}$ and $g \in \partial f(x)$, we have

$$f(x) + f_*(g) = \langle g, x \rangle. \tag{6.1.5}$$

Hence, for any $s \in \text{dom}\, f_*$ this implies that

$$f_*(s) \overset{(6.1.3)}{\geq} \langle s, x \rangle - f(x) \overset{(6.1.5)}{=} f_*(g) + \langle s - g, x \rangle.$$

In other words, if $g \in \partial f(x)$, then $x \in \partial f_*(g)$.

Let us prove the following relation (compare with general Theorem 3.1.16).

**Lemma 6.1.1** *For all $x \in \mathbb{R}^n$, we have*

$$f(x) = \max_{s \in \text{dom}\, f_*} [\langle s, x \rangle - f_*(s)].$$

*Proof* Indeed, for any $s \in \text{dom}\, f_*$, we have $\langle s, x \rangle - f_*(s) \overset{(6.1.3)}{\leq} f(x)$, and, in view of (6.1.5), equality is achieved for $s \in \partial f(x)$. □

Let us now look at the following smooth approximation of function $f$:

$$f_\mu(x) = \max_{s \in \text{dom}\, f_*} \left\{ \langle s, x \rangle - f_*(s) - \tfrac{1}{2}\mu(\|s\|^*)^2 \right\}, \tag{6.1.6}$$

where $\mu \geq 0$ is a smoothing parameter and the dual norm is defined as $\|s\|^* = \langle s, B^{-1}s \rangle^{1/2}$. In view of Lemma 6.1.1, we have

$$f(x) \geq f_\mu(x) \overset{(6.1.4)}{\geq} f(x) - \tfrac{1}{2}\mu L^2, \quad \forall x \in \mathbb{E}. \tag{6.1.7}$$

On the other hand, it appears that the function $f_\mu$ has a Lipschitz continuous gradient.

**Lemma 6.1.2** *The function $f_\mu$ is differentiable on $\mathbb{E}$, and for any points $x_1$ and $x_2 \in \mathbb{E}$ we have*

$$\|\nabla f_\mu(x_1) - \nabla f_\mu(x_2)\|^* \leq \tfrac{1}{\mu}\|x_1 - x_2\|. \tag{6.1.8}$$

*Proof* Consider two points $x_1$ and $x_2$ from $\mathbb{E}$. Let $s_i^*$, $i = 1, 2$ be the optimal solutions of the corresponding optimization problems in (6.1.6). They are uniquely defined since the objective function in definition (6.1.6) is strongly concave.

Note that by Theorem 3.1.14, $s_i^* \in \partial f_\mu(x_i)$, $i = 1, 2$. On the other hand, by the first-order optimality condition of Theorem 3.1.20, there exist vectors $\tilde{x}_i \in \partial f_*(s_i^*)$ such that

$$\langle s - s_i^*, x_i - \tilde{x}_i - \mu B^{-1} s_i^* \rangle \le 0, \quad \forall s \in \operatorname{dom} f_*, \quad i = 1, 2.$$

Taking in this inequality $s = s_{3-i}^*$ and adding two copies of it with $i = 1, 2$, we get

$$\mu(\|s_1^* - s_2^*\|^*)^2 \le \langle s_1^* - s_2^*, x_1 - \tilde{x}_1 - (x_2 - \tilde{x}_2)\rangle \overset{(3.1.24)}{\le} \langle s_1^* - s_2^*, x_1 - x_2\rangle$$

$$\le \|s_1^* - s_2^*\|^* \cdot \|x_1 - x_2\|.$$

Thus, $\|s_1^* - s_2^*\|^* \le \frac{1}{\mu}\|x_1 - x_2\|$. Now, applying Lemma 3.1.10, we get $\nabla f_\mu(x_i) = s_i^*$, $i = 1, 2$. □

Of course the smooth approximation (6.1.6) of the function $f$ is not very practical since its internal minimization problem includes a potentially complicated function $f_*$. However, it already gives us some hints. Indeed, if we choose $\mu \approx \epsilon$, then the Lipschitz constant $L_\mu$ for the gradient of $f_\mu$ will be $O(\frac{1}{\epsilon})$. Therefore, Fast Gradient Methods (e.g. (2.2.20)) can find an $\epsilon$-approximation of function $f$ (this is $f_\mu$) in $O\left(\sqrt{\frac{L_\mu}{\epsilon}}\right) \approx O(\frac{1}{\epsilon})$ calls of an oracle.

It remains to find a systematic and computationally inexpensive way of approximating the initial non-smooth objective function by a function with a Lipschitz continuous gradient. This can be done by exploiting a special max-representation of the objective function, which we introduce in Sect. 6.1.2.

For our goals, it is convenient to use the following notation. We often work with two finite-dimensional real vector spaces $\mathbb{E}_1$ and $\mathbb{E}_2$. In these spaces, we use the corresponding scalar products and general norms

$$\langle s, x\rangle_{E_i}, \quad \|x\|_{\mathbb{E}_i}, \quad \|s\|_{\mathbb{E}_i}^*, \quad x \in \mathbb{E}_i, \quad s \in \mathbb{E}_i^*, \quad i = 1, 2,$$

which are not necessarily Euclidean. A *norm* of a linear operator $A : \mathbb{E}_1 \to \mathbb{E}_2^*$ is defined in the standard way:

$$\|A\|_{1,2} = \max_{x,u}\{\langle Ax, u\rangle_{\mathbb{E}_2} : \|x\|_{\mathbb{E}_1} = 1, \ \|u\|_{\mathbb{E}_2} = 1\}.$$

Clearly,

$$\|A\|_{1,2} = \|A^*\|_{2,1} = \max_x\{\|Ax\|_{\mathbb{E}_2}^* : \|x\|_{\mathbb{E}_1} = 1\}$$

$$= \max_u\{\|A^*u\|_{\mathbb{E}_1}^* : \|u\|_{\mathbb{E}_2} = 1\}.$$

Hence, for any $x \in \mathbb{E}_1$ and $u \in \mathbb{E}_2$ we have

$$\|Ax\|^*_{\mathbb{E}_2} \leq \|A\|_{1,2} \cdot \|x\|_{\mathbb{E}_1}, \quad \|A^*u\|^*_{\mathbb{E}_1} \leq \|A\|_{1,2} \cdot \|u\|_{\mathbb{E}_2}. \tag{6.1.9}$$

### 6.1.2 The Minimax Model of an Objective Function

In this section, our main problem of interest is as follows:

$$\text{Find } f^* = \min_x \{ f(x) : \; x \in Q_1 \}, \tag{6.1.10}$$

where $Q_1$ is a bounded closed convex set in a finite-dimensional real vector space $E_1$, and $f(\cdot)$ is a continuous convex function on $Q_1$. We do not assume $f$ to be differentiable.

Quite often, the *structure* of the objective function in (6.1.10) is given explicitly. Let us assume that this structure can be described by the following *model*:

$$f(x) = \hat{f}(x) + \max_u \{ \langle Ax, u \rangle_{\mathbb{E}_2} - \hat{\phi}(u) : \; u \in Q_2 \}, \tag{6.1.11}$$

where the function $\hat{f}(\cdot)$ is continuous and convex on $Q_1$, $Q_2$ is a bounded closed convex set in a finite-dimensional real vector space $E_2$, $\hat{\phi}(\cdot)$ is a continuous convex function on $Q_2$, and the linear operator $A$ maps $E_1$ to $E_2^*$. In this case, problem (6.1.10) can be written in an *adjoint* form. Indeed,

$$\begin{aligned} f^* &= \min_{x \in Q_1} \max_{u \in Q_2} \{ \hat{f}(x) + \langle Ax, u \rangle_{\mathbb{E}_2} - \hat{\phi}(u) \} \\ &\overset{(1.3.6)}{\geq} \max_{u \in Q_2} \min_{x \in Q_1} \{ \hat{f}(x) + \langle Ax, u \rangle_{\mathbb{E}_2} - \hat{\phi}(u) \}. \end{aligned}$$

Thus, the adjoint problem can be stated as follows:

$$\begin{aligned} f_* &= \max_{u \in Q_2} \phi(u), \\ \phi(u) &= -\hat{\phi}(u) + \min_{x \in Q_1} \{ \langle Ax, u \rangle_{\mathbb{E}_2} + \hat{f}(x) \}. \end{aligned} \tag{6.1.12}$$

However, the complexity of this problem is not completely identical to that of (6.1.10). Indeed, in the primal problem (6.1.10), we implicitly assume that the function $\hat{\phi}(\cdot)$ and set $Q_2$ are so simple that the solution of the optimization problem in (6.1.11) can be found in a closed form. This assumption may be not valid for the objects defining the function $\phi(\cdot)$.

Note that usually, for a convex function $f$, representation (6.1.11) is *not* uniquely defined. If we decide to use, for example, the Fenchel dual of $f$,

$$\hat{\phi}(u) \equiv f_*(u) = \max_x \{\langle u, x\rangle_{\mathbb{E}_1} - f(x) : \; x \in \mathbb{E}_1\}, \quad Q_2 \equiv \mathbb{E}_2 = \mathbb{E}_1^*,$$

then we can take $\hat{f}(x) \equiv 0$, and $A$ is equal to $I_n$, the identity operator. However, in this case the function $\hat{\phi}(\cdot)$ may be too complicated for our goals. Intuitively, it is clear that the bigger the dimension of the space $\mathbb{E}_2$ is, the simpler is the structure of the adjoint object defined by the function $\hat{\phi}(\cdot)$ and the set $Q_2$. Let us demonstrate this with an example.

*Example 6.1.1* Consider $f(x) = \max\limits_{1\le j\le m} |\langle a_j, x\rangle_{\mathbb{E}_1} - b^{(j)}|$. Let us choose $A = I_n$, $\mathbb{E}_2 = \mathbb{E}_1^* = \mathbb{R}^n$, and

$$\begin{aligned}
\hat{\phi}(u) = f_*(u) &= \max_x \left\{ \langle u, x\rangle_{\mathbb{E}_1} - \max_{1\le j\le m} |\langle a_j, x\rangle_{\mathbb{E}_1} - b^{(j)}| \right\} \\
&= \max_x \min_{s\in\mathbb{R}^m} \left\{ \langle u, x\rangle_{\mathbb{E}_1} - \sum_{j=1}^m s^{(j)}[\langle a_j, x\rangle_{\mathbb{E}_1} - b^{(j)}] : \; \sum_{j=1}^m |s^{(j)}| \le 1 \right\} \\
&= \min_{s\in\mathbb{R}^m} \left\{ \langle b, s\rangle_{\mathbb{E}_2} : \; As = u, \; \sum_{j=1}^m |s^{(j)}| \le 1 \right\}.
\end{aligned}$$

It is clear that the structure of such a function can be very complicated.

Let us look at another possibility. Note that

$$\begin{aligned}
f(x) &= \max_{1\le j\le m} |\langle a_j, x\rangle_{\mathbb{E}_1} - b^{(j)}| \\
&= \max_{u\in\mathbb{R}^m} \left\{ \sum_{j=1}^m u^{(j)}[\langle a_j, x\rangle_{\mathbb{E}_1} - b^{(j)}] : \; \sum_{j=1}^m |u^{(j)}| \le 1 \right\}.
\end{aligned}$$

In this case $\mathbb{E}_2 = \mathbb{R}^m$, $\hat{\phi}(u) = \langle b, u\rangle_{\mathbb{E}_2}$ and $Q_2 = \left\{ u \in \mathbb{R}^m : \; \sum\limits_{j=1}^m |u^{(j)}| \le 1 \right\}$.

Finally, we can also represent $f(x)$ as follows:

$$f(x) = \max_{u=(u_1,u_2)\in\mathbb{R}^{2m}_+} \left\{ \sum_{j=1}^m (u_1^{(j)} - u_2^{(j)}) \cdot [\langle a_j, x\rangle_{\mathbb{E}_1} - b^{(j)}] : \sum_{j=1}^m (u_1^{(j)} + u_2^{(j)}) = 1 \right\}.$$

In this case $\mathbb{E}_2 = \mathbb{R}^{2m}$, $\hat{\phi}(u)$ is a linear function and $Q_2$ is a simplex. In Sect. 6.1.4.4 we will see that this representation is the easiest one. □

Let us show that the knowledge of structure (6.1.11) can help in solving both problems (6.1.10) and (6.1.12). We are going to use this structure to construct a smooth approximation of the objective function in (6.1.10).

Consider a differentiable *prox-function* $d_2(\cdot)$ of the set $Q_2$. This means that $d_2(\cdot)$ is strongly convex on $Q_2$ with convexity parameter one. Denote by

$$u_0 = \arg\min_u \{d_2(u) : \ u \in Q_2\}$$

its *prox-center*. Without loss of generality, we assume that $d_2(u_0) = 0$. Thus, for any $u \in Q_2$ we have

$$d_2(u) \overset{(2.2.40)}{\geq} \frac{1}{2}\|u - u_0\|^2_{\mathbb{E}_2}. \tag{6.1.13}$$

Let $\mu$ be a positive *smoothing* parameter. Consider the following function:

$$f_\mu(x) = \max_u \{\langle Ax, u\rangle_{\mathbb{E}_2} - \hat{\phi}(u) - \mu d_2(u) : \ u \in Q_2\}. \tag{6.1.14}$$

Denote by $u_\mu(x)$ the optimal solution of the above problem. Since the function $d_2(\cdot)$ is strongly convex, this solution is unique.

**Theorem 6.1.1** *The function $f_\mu$ is well defined and continuously differentiable at any $x \in \mathbb{E}_1$. Moreover, this function is convex and its gradient*

$$\nabla f_\mu(x) = A^* u_\mu(x) \tag{6.1.15}$$

*is Lipschitz continuous with constant*

$$L_\mu = \tfrac{1}{\mu}\|A\|^2_{1,2}.$$

*Proof* Indeed the function $f_\mu(\cdot)$ is convex as a maximum of functions which are linear in $x$, and $A^* u_\mu(x) \in \partial f_\mu(x)$ (see Lemma 3.1.14). Let us prove now the existence and Lipschitz continuity of its gradient.

Consider two points $x_1$ and $x_2$ from $\mathbb{E}_1$. From the first-order optimality conditions (3.1.56), we have

$$\langle Ax_i - g_i - \mu \nabla d_2(u_\mu(x_i)), u_\mu(x_{3-i}) - u_\mu(x_i)\rangle_{\mathbb{E}_2} \leq 0$$

for some $g_i \in \partial\hat{\phi}(u_\mu(x_i))$, $i = 1, 2$. Adding these inequalities, we get

$$\begin{aligned}
\mu\|u_\mu(x_1) - u_\mu(x_2)\|^2_{\mathbb{E}_2} &\overset{(2.1.22)}{\leq} \mu\langle \nabla d_2(u_\mu(x_1)) - \nabla d_2(u_\mu(x_2)), u_\mu(x_1) - u_\mu(x_2)\rangle_{\mathbb{E}_2} \\
&\leq \langle A(x_1 - x_2) - (g_1 - g_2), u_\mu(x_1) - u_\mu(x_2)\rangle_{\mathbb{E}_2}
\end{aligned}$$

$$\overset{(3.1.24)}{\leq} \langle A(x_1 - x_2), u_\mu(x_1) - u_\mu(x_2)\rangle_{\mathbb{E}_2}$$

$$\leq \|A\|_{1,2} \cdot \|x_1 - x_2\|_{\mathbb{E}_1} \cdot \|u_\mu(x_1) - u_\mu(x_2)\|_{\mathbb{E}_2}.$$

Thus, in view of (6.1.9), we have

$$\begin{aligned}\|A^* u_\mu(x_1) - A^* u_\mu(x_2))\|^*_{\mathbb{E}_1} &\leq \|A\|_{1,2} \cdot \|u_\mu(x_1) - u_\mu(x_2)\|^2_{\mathbb{E}_2} \\ &\leq \tfrac{1}{\mu}\|A\|^2_{1,2} \cdot \|x_1 - x_2\|_{\mathbb{E}_1}.\end{aligned}$$

It remains to use Lemma 3.1.10. □

Let $D_2 = \max\limits_{u \in Q_2} d_2(u)$ and $f_0(x) = \max\limits_{u \in Q_2}\{\langle Ax, u\rangle_{\mathbb{E}_2} - \hat{\phi}(u)\}$. Then, for any $x \in \mathbb{E}_1$ we have

$$f_0(x) \overset{(6.1.14)}{\geq} f_\mu(x) \overset{(6.1.14)}{\geq} f_0(x) - \mu D_2. \qquad (6.1.16)$$

Thus, for $\mu > 0$ the function $f_\mu$ can be seen as a uniform $\mu$-approximation of the objective function $f_0$ with Lipschitz constant for the gradient of the order $O(\frac{1}{\mu})$.

### 6.1.3 The Fast Gradient Method for Composite Minimization

Let $f(\cdot)$ be a convex differentiable function defined on a closed convex set $Q \subseteq E$. Assume that the gradient of this function is Lipschitz continuous:

$$\|\nabla f(x) - \nabla f(y)\|^* \leq L\|x - y\|, \quad \forall x, y \in Q.$$

Denote by $d(\cdot)$ a differentiable *prox-function* of the set $Q$. Assume that $d(\cdot)$ is strongly convex on $Q$ with convexity parameter one. Let $x_0$ be the *$d$-center* of $Q$:

$$x_0 = \arg\min_{x \in Q} d(x).$$

Without loss of generality, assume that $d(x_0) = 0$. Thus, for any $x \in Q$ we have

$$d(x) \overset{(2.2.40)}{\geq} \tfrac{1}{2}\|x - x_0\|^2. \qquad (6.1.17)$$

In this section, we present a fast gradient method for solving the following *composite* optimization problem:

$$\min_x \left\{ \tilde{f}(x) \overset{\text{def}}{=} f(x) + \Psi(x) : \; x \in Q \right\}, \qquad (6.1.18)$$

where $\Psi(\cdot)$ is an arbitrary *simple* closed convex function defined on $Q$. Our main assumption is that the auxiliary minimization problem of the form

$$\min_{x\in Q}\{\langle s, x\rangle + \alpha d(x) + \beta\Psi(x)\}, \quad \alpha, \beta \geq 0,$$

is easily solvable. For simplicity, we assume that the constant $L > 0$ is known.

**Method of Similar Triangles**

**0.** Choose $x_0 \in Q$. Set $v_0 = x_0$ and $\phi_0(x) = Ld(x)$.

**1.** ***k*th iteration ($k \geq 0$).**

(a) Define $y_k = \frac{k}{k+2}x_k + \frac{2}{k+2}v_k$.

(b) Set $\phi_{k+1}(x) = \phi_k(x) + \frac{k+1}{2}[f(y_k) + \langle\nabla f(y_k), x - y_k\rangle + \Psi(x)]$.

(c) Compute $v_{k+1} = \min\limits_{x\in Q}\phi_{k+1}(x)$.

(d) Define $x_{k+1} = \frac{k}{k+2}x_k + \frac{2}{k+2}v_{k+1}$.

(6.1.19)

In this scheme, we generate two sequences of feasible points $\{x_k\}_{k=0}^{\infty}$ and $\{y_k\}_{k=0}^{\infty}$, and a sequence of estimating functions $\{\phi_k(x)\}_{k=0}^{\infty}$. At each iteration of this method, all "events" happen in the two-dimensional plane defined by the triangle

$$\{x_k, v_k, v_{k+1}\}.$$

Note that this triangle is similar to the resulting triangle $\{x_k, y_k, x_{k+1}\}$, defining the new point of the sequence $\{x_k\}_{k=0}^{\infty}$, for which we are able to establish the rate of convergence.

**Theorem 6.1.2** *Let the sequences $\{x_k\}_{k=0}^{\infty}$, $\{y_k\}_{k=0}^{\infty}$, and $\{v_k\}_{k=0}^{\infty}$ be generated by method (6.1.19). Then, for any $k \geq 0$ and $x \in Q$ we have*

$$\begin{aligned}&\tfrac{k(k+1)}{4}\tilde{f}(x_k) + \tfrac{L}{2}\|v_k - x\|^2\\ &\leq \phi_k(x) = Ld(x) + \sum_{i=0}^{k-1}\tfrac{i+1}{2}[f(y_i) + \langle\nabla f(y_i), x - y_i\rangle] + \tfrac{k(k+1)}{4}\Psi(x).\end{aligned} \tag{6.1.20}$$

*Therefore, for any $k \geq 1$, we get*

$$\tilde{f}(x_k) - \tilde{f}(x^*) + \tfrac{2L}{k(k+1)}\|v_k - x^*\|^2 \leq \tfrac{4Ld(x^*)}{k(k+1)}, \tag{6.1.21}$$

*where $x^*$ is an optimal solution to problem (6.1.18).*

*Proof* For $k \geq 0$, let

$$a_k = \tfrac{k}{2}, \quad A_k = \sum_{i=0}^{k} a_i = \tfrac{k(k+1)}{4}, \quad \tau_k = \tfrac{a_{k+1}}{A_{k+1}}.$$

Then the rules of method (6.1.19) can be written as follows:

$$y_k = (1-\tau_k)x_k + \tau_k v_k, \quad x_{k+1} = (1-\tau_k)x_k + \tau_k v_{k+1}. \tag{6.1.22}$$

Let us prove that

$$A_k \tilde{f}(x_k) \leq \phi_k^* \overset{\text{def}}{=} \min_{x \in Q} \phi_k = \phi_k(v_k), \quad k \geq 0. \tag{6.1.23}$$

Since $A_0 = 0$, this inequality is valid for $k = 0$. Assume that it is true for some $k \geq 0$. Since all functions $\phi_k$ are strongly convex with convexity parameter $L$, we have

$$\begin{aligned}
\phi_{k+1}^* &= \phi_k(v_{k+1}) + a_{k+1}[f(y_k) + \langle \nabla f(y_k), v_{k+1} - y_k\rangle + \Psi(v_{k+1})] \\
&\overset{(2.2.40)}{\geq} \phi_k^* + \tfrac{L}{2}\|v_{k+1} - v_k\|^2 \\
&\quad + a_{k+1}[f(y_k) + \langle \nabla f(y_k), v_{k+1} - y_k\rangle + \Psi(v_{k+1})] \\
&\overset{(6.1.23)}{\geq} A_k[f(x_k) + \Psi(x_k)] + \tfrac{L}{2}\|v_{k+1} - v_k\|^2 \\
&\quad + a_{k+1}[f(y_k) + \langle \nabla f(y_k), v_{k+1} - y_k\rangle + \Psi(v_{k+1})] \\
&\overset{(2.1.2)}{\geq} A_{k+1} f(y_k) + \langle \nabla f(y_k), A_k(x_k - y_k) + a_{k+1}(v_{k+1} - y_k)\rangle \\
&\quad + \tfrac{L}{2}\|v_{k+1} - v_k\|^2 + A_k\Psi(x_k) + a_{k+1}\Psi(v_{k+1}).
\end{aligned}$$

By the rules of the method, $A_k(x_k - y_k) + a_{k+1}(v_{k+1} - y_k) \overset{(6.1.22)}{=} a_{k+1}(v_{k+1} - v_k)$ and $A_k\Psi(x_k) + a_{k+1}\Psi(v_{k+1}) \geq A_{k+1}\Psi(x_{k+1})$. Therefore,

$$\begin{aligned} \phi_{k+1}^* &\geq A_{k+1}f(y_k) + a_{k+1}\langle \nabla f(y_k), v_{k+1} - v_k\rangle + \tfrac{L}{2}\|v_{k+1} - v_k\|^2 \\ &\quad + A_{k+1}\Psi(x_{k+1}) \\ &\overset{(6.1.22)}{=} A_{k+1}[f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k\rangle + \tfrac{LA_{k+1}}{2a_{k+1}^2}\|x_{k+1} - y_k\|^2 \\ &\quad + \Psi(x_{k+1})]\,. \end{aligned}$$

Since $\frac{A_{k+1}}{a_{k+1}^2} = \frac{(k+1)(k+2)}{4} \cdot \frac{4}{(k+1)^2} > 1$, we get $\phi_{k+1}^* \overset{(2.1.9)}{\geq} A_{k+1}f(x_{k+1})$. By strong convexity of the function $\phi_k$, we have

$$\phi_k(x) \overset{(2.2.40)}{\geq} \phi_k^* + \tfrac{L}{2}\|x - v_k\|^2 \overset{(6.1.23)}{\geq} A_k\tilde{f}(x_k) + \tfrac{L}{2}\|x - v_k\|^2,$$

and this is inequality (6.1.20). Finally, inequality (6.1.21) follows from (6.1.20) in view of the convexity of the function $f$. □

*Remark 6.1.1* Note that method (6.1.19) generates bounded sequences of points. Indeed, by the rules of this method we have

$$x_k, y_k \in \operatorname{Conv}\{v_0, \dots, v_k\}, \quad k \geq 0.$$

On the other hand, from inequality (6.1.21), it follows that

$$\|v_k - x^*\|^2 \leq 2d(x^*). \tag{6.1.24}$$

In the Euclidean case, $d(x) = \frac{1}{2}\|x - x_0\|^2$, and we get

$$\|v_k - x^*\| \leq \|x_0 - x^*\|, \quad k \geq 0. \tag{6.1.25}$$

### 6.1.4 Application Examples

Let us put the results of the previous sections together. Assume that the function $\hat{f}(\cdot)$ in (6.1.11) is differentiable and its gradient is Lipschitz-continuous with some constant $M \geq 0$. Then the smoothing technique as applied to problem (6.1.10) provides us with the following objective function:

$$\bar{f}_\mu(x) = \hat{f}(x) + f_\mu(x) \quad \to \quad \min : \; x \in Q_1. \tag{6.1.26}$$

In view of Theorem 6.1.1, the gradient of this function is Lipschitz continuous with the constant

$$L_\mu = M + \tfrac{1}{\mu}\|A\|_{1,2}^2.$$

Let us choose some prox-function $d_1(\cdot)$ for the set $Q_1$ with convexity parameter equal to one. Recall that the set $Q_1$ is assumed to be bounded:

$$\max_{x \in Q_1} d_1(x) \le D_1.$$

**Theorem 6.1.3** *Let us apply method (6.1.19) to problem (6.1.26) with the following value of the smoothness parameter:*

$$\mu = \mu(N) = \tfrac{2\|A\|_{1,2}}{\sqrt{N(N+1)}} \cdot \sqrt{\tfrac{D_1}{D_2}}.$$

*Then after $N$ iterations we can generate approximate solutions to problems (6.1.10) and (6.1.12), namely,*

$$\hat{x} = x_N \in Q_1, \quad \hat{u} = \sum_{i=0}^{N-1} \tfrac{2(i+1)}{(N+1)(N+2)} u_\mu(y_i) \in Q_2, \tag{6.1.27}$$

*which satisfy the following inequality:*

$$0 \le f(\hat{x}) - \phi(\hat{u}) \le \tfrac{4\|A\|_{1,2}}{\sqrt{N(N+1)}} \cdot \sqrt{D_1 D_2} + \tfrac{4MD_1}{N(N+1)}. \tag{6.1.28}$$

*Thus, the complexity of finding an $\epsilon$-solution to problems (6.1.10), (6.1.12) by the smoothing technique does not exceed*

$$4\|A\|_{1,2}\sqrt{D_1 D_2} \cdot \tfrac{1}{\epsilon} + 2\sqrt{\tfrac{MD_1}{\epsilon}} \tag{6.1.29}$$

*iterations of method (6.1.19).*

*Proof* Let us fix an arbitrary $\mu > 0$. In view of Theorem 6.1.2, after $N$ iterations of method (2.2.63) we can deliver a point $\hat{x} = x_N$ such that

$$\bar{f}_\mu(\hat{x}) \le \tfrac{4L_\mu D_1}{N(N+1)} + \min_{x \in Q_1} \sum_{i=0}^{N-1} \tfrac{2(i+1)}{N(N+1)}[\bar{f}_\mu(y_i) + \langle \nabla \bar{f}_\mu(x_i), x - y_i\rangle_{\mathbb{E}_1}]. \tag{6.1.30}$$

Note that

$$\begin{aligned} f_\mu(y) &= \max_u \{\langle Ay, u\rangle_{\mathbb{E}_2} - \hat{\phi}(u) - \mu d_2(u) : \ u \in Q_2\} \\ &= \langle Ay, u_\mu(y)\rangle_{\mathbb{E}_2} - \hat{\phi}(u_\mu(y)) - \mu d_2(u_\mu(y)), \\ \langle \nabla f_\mu(y), y\rangle_{\mathbb{E}_1} &= \langle A^* u_\mu(y), y\rangle_{\mathbb{E}_1}. \end{aligned}$$

Therefore, for $i = 0, \dots, N-1$ we have

$$f_\mu(y_i) - \langle \nabla f_\mu(y_i), y_i\rangle_{\mathbb{E}_1} \ = \ -\hat{\phi}(u_\mu(y_i)) - \mu d_2(u_\mu(y_i)). \tag{6.1.31}$$

Thus, in view of (6.1.15) and (6.1.31) we obtain

$$\begin{aligned} &\sum_{i=0}^{N-1} (i+1)[\bar{f}_\mu(y_i) + \langle \nabla \bar{f}_\mu(y_i), x - y_i\rangle_{\mathbb{E}_1}] \\ \overset{(2.1.2)}{\leq} &\sum_{i=0}^{N-1} (i+1)[f_\mu(y_i) - \langle \nabla f_\mu(y_i), y_i\rangle_{\mathbb{E}_1}] + \tfrac{1}{2}N(N+1)(\hat{f}(x) + \langle A^*\hat{u}, x\rangle_{\mathbb{E}_1}) \\ \leq & \ -\sum_{i=0}^{N-1} (i+1)\hat{\phi}(u_\mu(y_i)) + \tfrac{1}{2}N(N+1)(\hat{f}(x) + \langle A^*\hat{u}, x\rangle_{\mathbb{E}_1}) \\ \leq & \ \tfrac{1}{2}N(N+1)[-\hat{\phi}(\hat{u}) + \hat{f}(x) + \langle Ax, \hat{u}\rangle_{\mathbb{E}_2}]. \end{aligned}$$

Hence, using (6.1.30), (6.1.12) and (6.1.16), we get the following bound:

$$\tfrac{4L_\mu D_1}{N(N+1)} \geq \bar{f}_\mu(\hat{x}) - \phi(\hat{u}) \geq f(\hat{x}) - \phi(\hat{u}) - \mu D_2.$$

This is

$$0 \leq f(\hat{x}) - \phi(\hat{u}) \leq \mu D_2 + \tfrac{4\|A\|_{1,2}^2 D_1}{\mu N(N+1)} + \tfrac{4MD_1}{N(N+1)}. \tag{6.1.32}$$

Minimizing the right-hand side of this inequality in $\mu$, we get inequality (6.1.28). □

Note that the efficiency estimate (6.1.29) is much better than the standard bound $O\left(\frac{1}{\epsilon^2}\right)$. In accordance with the above theorem, for $M = 0$ the optimal dependence of the parameters $\mu$, $L_\mu$ and $N$ in $\epsilon$ is as follows:

$$\sqrt{N(N+1)} \geq 4\|A\|_{1,2}\sqrt{D_1 D_2} \cdot \tfrac{1}{\epsilon}, \quad \mu \ = \ \tfrac{\epsilon}{2D_2}, \quad L_\mu \ = \ D_2 \cdot \tfrac{\|A\|_{1,2}^2}{\epsilon}. \tag{6.1.33}$$

*Remark 6.1.2* Inequality (6.1.28) shows that the pair of adjoint problems (6.1.10) and (6.1.12) has no *duality gap*:

$$f^* = f_*. \tag{6.1.34}$$

Let us now look at some examples.

#### 6.1.4.1 Minimax Strategies for Matrix Games

Denote by $\Delta_n$ the standard simplex in $\mathbb{R}^n$:

$$\Delta_n = \left\{ x \in \mathbb{R}^n_+ : \sum_{i=1}^n x^{(i)} = 1 \right\}.$$

Let $A : \mathbb{R}^n \to \mathbb{R}^m$, $\mathbb{E}_1 = \mathbb{R}^n$, and $\mathbb{E}_2 = \mathbb{R}^m$. Consider the following saddle point problem:

$$\min_{x\in\Delta_n} \max_{u\in\Delta_m} \{\langle Ax, u\rangle_{\mathbb{E}_2} + \langle c, x\rangle_{\mathbb{E}_1} + \langle b, u\rangle_{\mathbb{E}_2}\}. \tag{6.1.35}$$

From the viewpoint of players, this problem can be seen as a pair of non-smooth minimization problems:

$$\begin{aligned} \min_{x\in\Delta_n} f(x),\ f(x) &= \langle c, x\rangle_{\mathbb{E}_1} + \max_{1\le j\le m} [\langle a_j, x\rangle_{\mathbb{E}_1} + b^{(j)}], \\ \max_{u\in\Delta_m} \phi(u),\ \phi(u) &= \langle b, u\rangle_{\mathbb{E}_2} + \min_{1\le i\le n} [\langle \hat{a}_i, u\rangle_{\mathbb{E}_2} + c^{(i)}], \end{aligned} \tag{6.1.36}$$

where $a_j$ are the rows and $\hat{a}_i$ are the columns of matrix $A$. In order to solve this pair of problems using the smoothing approach, we need to find a reasonable prox-function for the simplex. Let us compare two possibilities.

**1. Euclidean Distance** Let us choose

$$\|x\|_{\mathbb{E}_1} = \left[\sum_{i=1}^n (x^{(i)})^2\right]^{1/2}, \quad d_1(x) = \tfrac{1}{2}\sum_{i=1}^n (x^{(i)} - \tfrac{1}{n})^2,$$

$$\|u\|_{\mathbb{E}_2} = \left[\sum_{j=1}^m (u^{(j)})^2\right]^{1/2}, \ d_2(x) = \tfrac{1}{2}\sum_{j=1}^m (u^{(j)} - \tfrac{1}{m})^2.$$

Then $D_1 = 1 - \frac{1}{n} < 1$, $D_2 = 1 - \frac{1}{m} < 1$ and

$$\|A\|_{1,2} = \max_u \{\|Ax\|_2^* : \|x\|_{\mathbb{E}_1} = 1\} = \lambda_{\max}^{1/2}(A^TA).$$

Thus, in our case the estimate (6.1.28) for the result (6.1.27) can be specified as follows:

$$0 \le f(\hat{x}) - \phi(\hat{u}) \le \frac{4\lambda_{\max}^{1/2}(A^T A)}{\sqrt{N(N+1)}}. \tag{6.1.37}$$

**2. Entropy Distance** Let us choose

$$\|x\|_{\mathbb{E}_1} = \sum_{i=1}^{n} |x^{(i)}|, \;\; d_1(x) = \ln n + \sum_{i=1}^{n} x^{(i)} \ln x^{(i)},$$

$$\|u\|_{\mathbb{E}_2} = \sum_{j=1}^{m} |u^{(j)}|, \; d_2(u) = \ln m + \sum_{j=1}^{m} u^{(j)} \ln u^{(j)}.$$

Functions $d_1$ and $d_2$ are called the *entropy functions*.

**Lemma 6.1.3** *The above prox-functions are strongly convex in an $\ell_1$-norm with convexity parameter one and $D_1 = \ln n$, $D_2 = \ln m$.*

*Proof* Note that the function $d_1$ is twice continuously differentiable in the interior of simplex $\Delta_n$, and

$$\langle \nabla^2 d_1(x) h, h \rangle = \sum_{i=1}^{n} \frac{(h^{(i)})^2}{x^{(i)}}.$$

Thus, in view of Theorem 2.1.11 strong convexity of $d_1$ is a consequence of the following variant of Cauchy–Schwarz inequality,

$$\left( \sum_{i=1}^{n} |h^{(i)}| \right)^2 \le \left( \sum_{i=1}^{n} x^{(i)} \right) \cdot \left( \sum_{i=1}^{n} \frac{(h^{(i)})^2}{x^{(i)}} \right),$$

which is valid for all positive vectors $x \in \mathbb{R}^n$. Since $d_1(\cdot)$ is a convex symmetric function of the arguments, its minimum is attained at the center of the simplex, the point $x_0 = \frac{1}{n}\bar{e}_n$. Clearly, $d_1(x_0) = 0$. On the other hand, its maximum is attained at one of the vertices of the simplex (see Corollary 3.1.2).

The reasoning for $d_2(\cdot)$ is similar. □

Note also that now we get the following norm of the operator $A$:

$$\|A\|_{1,2} = \max_{x} \{ \max_{1 \le j \le m} |\langle a_j, x \rangle| : \; \|x\|_{\mathbb{E}_1} \le 1 \} = \max_{i,j} |A^{(i,j)}|$$

(see Corollary 3.1.2). Thus, if we apply the entropy distance, the estimate (6.1.28) can be written as follows:

$$0 \le f(\hat{x}) - \phi(\hat{u}) \le \frac{4\sqrt{\ln n \ln m}}{\sqrt{N(N+1)}} \cdot \max_{i,j} |A^{(i,j)}|. \tag{6.1.38}$$

Note that typically the estimate (6.1.38) is much better than its Euclidean variant (6.1.37).

Let us write down explicitly the smooth approximation for the objective function in the first problem of (6.1.36) using the entropy distance. By definition,

$$\bar{f}_\mu(x) = \langle c, x\rangle_{\mathbb{E}_1} + \max_{u\in\Delta_m}\left\{\sum_{j=1}^m u^{(j)}[\langle a_j, x\rangle + b^{(j)}] - \mu\sum_{j=1}^m u^{(j)}\ln u^{(j)} - \mu\ln m\right\}.$$

Let us apply the following result.

**Lemma 6.1.4** *The solution of the problem*

$$\text{Find } \phi_*(s) = \max_{u\in\Delta_m}\left\{\sum_{j=1}^m u^{(j)}s^{(j)} - \mu\sum_{j=1}^m u^{(j)}\ln u^{(j)}\right\} \tag{6.1.39}$$

*is given by the vector $u_\mu(s)\in\Delta_m$ with the following entries*

$$u_\mu^{(j)}(s) = \frac{e^{s^{(j)}/\mu}}{\sum\limits_{i=1}^m e^{s^{(i)}/\mu}}, \quad j = 1,\dots,m. \tag{6.1.40}$$

*Therefore,* $\phi_*(s) = \mu\ln\left(\sum\limits_{i=1}^m e^{s^{(i)}/\mu}\right)$.

*Proof* Note that the gradient of the objective function in problem (6.1.39) goes to infinity as the argument approaches the boundary of the domain. Therefore, the first order necessary and sufficient optimality conditions for this problem are as follows (see (3.1.59)):

$$s^{(j)} - \mu(1+\ln u^{(j)}) = \lambda,\ \ j = 1,\dots,m,$$

$$\sum_{j=1}^m u^{(j)} = 1.$$

Clearly, they are satisfied by (6.1.40) with $\lambda = \mu\ln\left(\sum\limits_{l=1}^m e^{s^{(l)}/\mu}\right) - \mu$. □

Using the result of Lemma 6.1.4, we conclude that in our case the problem (6.1.26) is as follows:

$$\min_{x\in\Delta_n}\left\{\bar{f}_\mu(x) = \langle c, x\rangle_{\mathbb{E}_1} + \mu\ln\left(\frac{1}{m}\sum_{j=1}^m e^{[\langle a_j,x\rangle + b^{(j)}]/\mu}\right)\right\}.$$

Note that the complexity of the oracle for this problem is basically the same as that of the initial problem (6.1.36).

#### 6.1.4.2 The Continuous Location Problem

Consider the following *location* problem. There are $p$ cities with population $m_j$, which are located at points $c_j \in \mathbb{R}^n$, $j = 1, \dots, p$. We want to construct a service center at some position $x \in \mathbb{R}^n \equiv \mathbb{E}_1$, which minimizes the total social distance $f(x)$ to the center. On the other hand, this center must be constructed not too far from the origin.

Mathematically, the above problem can be posed as follows

$$\text{Find } f^* = \min_x \left\{ f(x) = \sum_{j=1}^{p} m_j \|x - c_j\|_{\mathbb{E}_1} : \|x\|_{\mathbb{E}_1} \le \bar{r} \right\}. \tag{6.1.41}$$

In accordance to its interpretation, it is natural to choose

$$\|x\|_{\mathbb{E}_1} = \left[ \sum_{i=1}^{n} (x^{(i)})^2 \right]^{1/2}, \quad d_1(x) = \tfrac{1}{2}\|x\|^2_{\mathbb{E}_1}.$$

Then $D_1 = \frac{1}{2}\bar{r}^2$.

Further, the structure of the adjoint space $\mathbb{E}_2$ is quite clear:

$$\mathbb{E}_2 = (\mathbb{E}_1^*)^p, \quad Q_2 = \left\{ u = (u_1, \dots, u_p) \in \mathbb{E}_2 : \|u_j\|^*_{\mathbb{E}_1} \le 1, \ j = 1, \dots, p \right\}.$$

Let us choose

$$\|u\|_{\mathbb{E}_2} = \left[ \sum_{j=1}^{p} m_j (\|u_j\|^*_{\mathbb{E}_1})^2 \right]^{1/2}, \quad d_2(u) = \tfrac{1}{2}\|u\|^2_{\mathbb{E}_2}.$$

Then $D_2 = \frac{1}{2}P$ with $P \equiv \sum\limits_{j=1}^{p} m_j$. Note that the value $P$ may be interpreted as the total size of the population.

It remains to compute the norm of the operator $A$:

$$\begin{aligned} \|A\|_{1,2} &= \max_{x,u} \left\{ \sum_{j=1}^{p} m_j \langle u_j, x \rangle_{\mathbb{E}_1} : \sum_{j=1}^{p} m_j (\|u_j\|^*_{\mathbb{E}_1})^2 = 1, \ \|x\|_{\mathbb{E}_1} = 1 \right\} \\ &= \max_{r_j} \left\{ \sum_{j=1}^{p} m_j r_j : \sum_{j=1}^{p} m_j r_j^2 = 1 \right\} = P^{1/2} \end{aligned}$$

(see Lemma 3.1.20).

Putting the computed values into the estimate (6.1.28), we get the following rate of convergence:

$$f(\hat{x}) - f^* \leq \frac{2P\bar{r}}{\sqrt{N(N+1)}}. \tag{6.1.42}$$

Note that the value $\tilde{f}(x) = \frac{1}{P} f(x)$ corresponds to the average individual expenses generated by the location $x$. Therefore,

$$\tilde{f}(\hat{x}) - \tilde{f}^* \leq \frac{2\bar{r}}{\sqrt{N(N+1)}}.$$

It is interesting that the right-hand side of this inequality is independent of any dimension. At the same time, it is clear that the reasonable accuracy for the approximate solution of our problem should not be too high. Given the low complexity of each iteration in the scheme (6.1.19), the total efficiency of the proposed technique looks quite promising.

To conclude with the location problem, let us write down explicitly a smooth approximation of the objective function.

$$\begin{aligned} f_\mu(x) &= \max_u \left\{ \sum_{j=1}^{p} m_j \langle u_j, x - c_j \rangle_{\mathbb{E}_1} - \mu d_2(u) : \; u \in Q_2 \right\} \\ &= \max_u \left\{ \sum_{j=1}^{p} m_j \left( \langle u_j, x - c_j \rangle_{\mathbb{E}_1} - \tfrac{1}{2}\mu (\|u_j\|^*_{\mathbb{E}_1})^2 \right) : \; \|u_j\|^*_{\mathbb{E}_1} \leq 1, \right. \\ &\quad j = 1, \dots, p \} \\ &= \sum_{j=1}^{p} m_j \psi_\mu(\|x - c_j\|_{\mathbb{E}_1}), \end{aligned}$$

where the function $\psi_\mu(\tau)$, $\tau \geq 0$, is defined as follows:

$$\psi_\mu(\tau) = \max_{\gamma \in [0,1]} \{\gamma\tau - \frac{1}{2}\mu\gamma^2\} = \begin{cases} \frac{\tau^2}{2\mu}, & 0 \leq \tau \leq \mu, \\ \tau - \frac{\mu}{2}, \; \mu \leq \tau. \end{cases} \tag{6.1.43}$$

This is the so-called the *Huber loss function*.

### 6.1.4.3 Variational Inequalities with a Linear Operator

Consider a linear operator $B(w) = Bw + c \colon \mathbb{E} \to \mathbb{E}^*$, which is *monotone*:

$$\langle Bh, h \rangle \geq 0 \quad \forall h \in \mathbb{E}.$$

Let $Q$ be a bounded closed convex set in $\mathbb{E}$. Then we can pose the following *variational inequality* problem:

$$\text{Find } w^* \in Q: \quad \langle B(w^*), w - w^* \rangle \geq 0 \quad \forall w \in Q. \tag{6.1.44}$$

Note that we can always rewrite problem (6.1.44) as an optimization problem. Indeed, define

$$\psi(w) = \max_v \{\langle B(v), w - v\rangle : \ v \in Q\}.$$

In view of Theorem 3.1.8, $\psi(w)$ is a convex function. Let us show that the problem

$$\min_w \{\psi(w) : \ w \in Q\} \tag{6.1.45}$$

is equivalent to (6.1.44).

**Lemma 6.1.5** *A point $w^*$ is a solution to (6.1.45) if and only if it solves variational inequality (6.1.44). Moreover, for such $w^*$ we have $\psi(w^*) = 0$.*

*Proof* Indeed, at any $w \in Q$ the function $\psi$ is non-negative. If $w^*$ is a solution to (6.1.44), then for any $v \in Q$ we have

$$\langle B(v), v - w^*\rangle \geq \langle B(w^*), v - w^*\rangle \geq 0.$$

Hence, $\psi(w^*) = 0$ and $w^* \in \operatorname{Arg}\min\limits_{w \in Q} \psi(w)$.

Now, consider some $w^* \in Q$ with $\psi(w^*) = 0$. Then for any $v \in Q$ we have

$$\langle B(v), v - w^*\rangle \geq 0.$$

Suppose there exists some $v_1 \in Q$ such that $\langle B(w^*), v_1 - w^*\rangle < 0$. Consider the points

$$v_\alpha = w^* + \alpha(v_1 - w^*), \quad \alpha \in [0, 1].$$

Then

$$\begin{aligned} 0 \leq \langle B(v_\alpha), v_\alpha - w^*\rangle &= \alpha\langle B(v_\alpha), v_1 - w^*\rangle \\ &= \alpha\langle B(w^*), v_1 - w^*\rangle + \alpha^2\langle B \cdot (v_1 - w^*), v_1 - w^*\rangle. \end{aligned}$$

Hence, for $\alpha$ small enough we get a contradiction. □

There are two possibilities for representing the problem (6.1.44), (6.1.45) in the form (6.1.10), (6.1.11).

**1. Primal Form** We take $\mathbb{E}_1 = \mathbb{E}_2 = \mathbb{E}$, $Q_1 = Q_2 = Q$, $d_1(x) = d_2(x) = d(x)$, $A = B$, and

$$\hat{f}(x) = \langle b, x \rangle_{\mathbb{E}_1}, \quad \hat{\phi}(u) = \langle b, u \rangle_{\mathbb{E}_1} + \langle Bu, u \rangle_{\mathbb{E}_1}.$$

Note that the quadratic function $\hat{\phi}(u)$ is convex. To compute the value and the gradient of the function $f_\mu(x)$, we need to solve the following problem:

$$\max_{u \in Q} \{ \langle Bx, u \rangle_{\mathbb{E}_1} - \mu d(u) - \langle b, u \rangle_{\mathbb{E}_1} - \langle Bu, u \rangle_{\mathbb{E}_1} \}. \tag{6.1.46}$$

Since in our case $M = 0$, from Theorem 6.1.3 we get the following estimate for the complexity of problem (6.1.44):

$$\tfrac{4 D_1 \| B \|_{1,2}}{\epsilon}. \tag{6.1.47}$$

However, because of the presence of a non-trivial quadratic function in (6.1.46), the oracle for the function $\hat{f}$ can be quite expensive. We can avoid that in the dual variant of this problem.

**2. Dual Form** Consider the dual variant of problem (6.1.45):

$$\min_{w \in Q} \max_{v \in Q} \langle B(v), w - v \rangle = \max_{v \in Q} \min_{w \in Q} \langle B(v), w - v \rangle = - \min_{v \in Q} \max_{w \in Q} \langle B(v), v - w \rangle.$$

Thus, we can take $\mathbb{E}_1 = \mathbb{E}_2 = \mathbb{E}$, $Q_1 = Q_2 = Q$, $d_1(x) = d_2(x) = d(x)$, $A = B$, and

$$\hat{f}(x) = \langle b, x \rangle_{\mathbb{E}_1} + \langle Bx, x \rangle_{\mathbb{E}_1}, \quad \hat{\phi}(u) = \langle b, u \rangle_{\mathbb{E}_1}.$$

Now the computation of the function value $f_\mu(x)$ becomes much simpler:

$$f_\mu(x) = \max_u \{ \langle Bx, u \rangle_{\mathbb{E}_1} - \mu d(u) - \langle b, u \rangle_{\mathbb{E}_1} : u \in Q \}.$$

Note that we pay quite a moderate cost for this. Indeed, now $M$ becomes equal to $\| B \|_{1,2}$. Hence, the complexity estimate (6.1.47) increases up to the following level:

$$\tfrac{4 D_1 \| B \|_{1,2}}{\epsilon} + \sqrt{\tfrac{D_1 \| B \|_{1,2}}{\epsilon}}.$$

In the important particular case of skew-symmetry of the operator $B$, that is $B + B^* = 0$, the primal and dual variant have a similar complexity.

#### 6.1.4.4 Piece-Wise Linear Optimization

**1. Maximum of Absolute Values** Consider the following problem:

$$\min_{x \in Q_1} \left\{ f(x) = \max_{1 \le j \le m} |\langle a_j, x \rangle_{\mathbb{E}_1} - b^{(j)}| \right\}. \tag{6.1.48}$$

For simplicity, let us choose

$$\|x\|_{\mathbb{E}_1} = \left[ \sum_{i=1}^n (x^{(i)})^2 \right]^{1/2}, \quad d_1(x) = \tfrac{1}{2}\|x\|^2.$$

Denote by $A$ the matrix with rows $a_j$, $j = 1, \ldots, m$. It is convenient to choose

$$\mathbb{E}_2 = \mathbb{R}^{2m}, \quad \|u\|_{\mathbb{E}_2} = \sum_{j=1}^{2m} |u^{(j)}|, \quad d_2(u) = \ln(2m) + \sum_{j=1}^{2m} u^{(j)} \ln u^{(j)}.$$

Then

$$f(x) = \max_u \{ \langle \hat{A}x, u \rangle_{\mathbb{E}_2} - \langle \hat{b}, u \rangle_{\mathbb{E}_2} : u \in \Delta_{2m} \},$$

where $\hat{A} = \begin{pmatrix} A \\ -A \end{pmatrix}$ and $\hat{b} = \begin{pmatrix} b \\ -b \end{pmatrix}$. Thus, $D_2 = \ln(2m)$, and

$$D_1 = \frac{1}{2}\bar{r}^2, \quad \bar{r} = \max_x \{ \|x\|_{\mathbb{E}_1} : x \in Q_1 \}.$$

It remains to compute the norm of the operator $\hat{A}$:

$$\|\hat{A}\|_{1,2} = \max_{x,u} \{ \langle \hat{A}x, u \rangle_{\mathbb{E}_2} : \|x\|_{\mathbb{E}_1} = 1, \|u\|_{\mathbb{E}_2} = 1 \}$$

$$= \max_x \{ \max_{1 \le j \le m} |\langle a_j, x \rangle_{\mathbb{E}_1}| : \|x\|_{\mathbb{E}_1} = 1 \} = \max_{1 \le j \le m} \|a_j\|_1^*.$$

Putting all the computed values into the estimate (6.1.29), we see that the problem (6.1.48) can be solved in

$$2\sqrt{2}\, \bar{r} \max_{1 \le j \le m} \|a_j\|_1^* \sqrt{\ln(2m)} \cdot \tfrac{1}{\epsilon}$$

iterations of scheme (6.1.19). The standard subgradient schemes in this situation can count only on an

$$O\left( \left[ \bar{r} \max_{1 \le j \le m} \|a_j\|_1^* \cdot \tfrac{1}{\epsilon} \right]^2 \right)$$

upper bound for the number of iterations.

Finally, the smooth version of the objective function in (6.1.48) is as follows:

$$\bar{f}_\mu(x) = \mu \ln \left( \frac{1}{m} \sum_{j=1}^{m} \xi \left( \frac{1}{\mu} [\langle a_j, x \rangle + b^{(j)}] \right) \right)$$

with $\xi(\tau) = \frac{1}{2}[e^\tau + e^{-\tau}]$. We leave the justification of this expression as an exercise for the reader.

**2. Sum of Absolute Values** Consider now the problem

$$\min_{x \in Q_1} \left\{ f(x) = \sum_{j=1}^{m} |\langle a_j, x \rangle_{\mathbb{E}_1} - b^{(j)}| \right\}. \tag{6.1.49}$$

The simplest representation of the function $f(\cdot)$ is as follows. Denote by $A$ the matrix with the rows $a_j$. Let us choose

$$\mathbb{E}_2 = \mathbb{R}^m, \quad Q_2 = \{u \in \mathbb{R}^m : |u^{(j)}| \leq 1, \; j = 1, \dots, m\},$$

$$d_2(u) = \tfrac{1}{2}\|u\|^2_{\mathbb{E}_2} = \tfrac{1}{2} \sum_{j=1}^{m} \|a_j\|^*_{\mathbb{E}_1} \cdot (u^{(j)})^2.$$

Then the smooth version of the objective function is as follows:

$$\begin{aligned} f_\mu(x) &= \max_u \{\langle Ax - b, u \rangle_{\mathbb{E}_2} - \mu d_2(u) : \; u \in Q_2\} \\ &= \sum_{j=1}^{m} \|a_j\|^*_{\mathbb{E}_1} \cdot \psi_\mu \left( \frac{|\langle a_j, x \rangle_{\mathbb{E}_1} - b^{(j)}|}{\|a_j\|^*_{\mathbb{E}_1}} \right), \end{aligned}$$

where the function $\psi_\mu(\tau)$ is defined by (6.1.43). Note that

$$\begin{aligned} \|A\|_{1,2} &= \max_{x,u} \left\{ \sum_{j=1}^{m} u^{(j)} \langle a_j, x \rangle_{\mathbb{E}_1} : \; \|x\|_{\mathbb{E}_1} \leq 1, \; \|u\|_{\mathbb{E}_2} \leq 1 \right\} \\ &\leq \max_u \left\{ \sum_{j=1}^{m} \|a_j\|^*_{\mathbb{E}_1} \cdot |u^{(j)}| : \; \sum_{j=1}^{m} \|a_j\|^*_{\mathbb{E}_1} \cdot (u^{(j)})^2 \leq 1 \right\} \\ &= D^{1/2} \; \equiv \; \left[ \sum_{j=1}^{m} \|a_j\|^*_{\mathbb{E}_1} \right]^{1/2}. \end{aligned}$$

On the other hand, $D_2 = \frac{1}{2}D$. Therefore from Theorem 6.1.3 we get the following complexity bound:

$$\frac{2}{\epsilon} \cdot \sqrt{2D_1} \cdot \sum_{j=1}^{m} \|a_j\|^*_{\mathbb{E}_1}$$

iterations of method (6.1.19).

## 6.1.5 Implementation Issues

### 6.1.5.1 Computational Complexity

Let us discuss the computational complexity of the method (6.1.19) as applied to the function $\bar{f}_\mu(\cdot)$. The main computations are performed at Steps (b) and (c) of the algorithm.

**Step (b). Call of Oracle** At this step we need to compute the solution of the following maximization problem:

$$\max_{u \in Q_2} \{\langle Ay_k, u\rangle_{\mathbb{E}_2} - \hat{\phi}(u) - \mu d_2(u) : \ u \in Q_2\}.$$

Note that from the origin of this problem we know that this computation for $\mu = 0$ can be done in a closed form. Thus, we can expect that with a properly chosen prox-function, computation of the smoothed version is not too difficult. In Sect. 6.1.4 we have seen three examples which confirm this belief.

**Step (c). Computation of $v_{k+1}$** This computation consists in solving the following problem:

$$\min_{x \in Q_1} \{d_1(x) + \langle s, x\rangle_{\mathbb{E}_1}\}$$

for some fixed $s \in \mathbb{E}_1^*$. If the set $Q_1$ and the prox-function $d_1(\cdot)$ are simple enough, this computation can be done in a closed form (see Sect. 6.1.4). For some sets we need to solve an auxiliary equation with one variable.

### 6.1.5.2 Computational Stability

Our approach is based on the smoothing of non-differentiable functions. In accordance with (6.1.33), the value of the smoothness parameter $\mu$ must be of the order of $\epsilon$. This may cause some numerical troubles in computing the function $\bar{f}_\mu(x)$ and its gradient. Among examples of Sect. 6.1.4, only a smooth variant of the objective function in Sect. 6.1.4.2 does not involve dangerous operations; all others need a careful implementation.

In both Sects. 6.1.4.1 and 6.1.4.4 we need a stable technique for computing the values and derivatives of the function

$$\eta(u) = \mu \ln \left( \sum_{j=1}^{m} e^{u^{(j)}/\mu} \right) \tag{6.1.50}$$

with very small values of parameter $\mu$. This can be done in the following way. Let

$$\bar{u} = \max_{1 \le j \le m} u^{(j)}, \quad v^{(j)} = u^{(j)} - \bar{u}, \; j = 1, \dots, m.$$

Then

$$\eta(u) = \bar{u} + \eta(v).$$

Note that all components of the vector $v$ are non-negative and one of them is zero. Therefore, the value $\eta(v)$ can be computed quite accurately. The same technique can be used to compute the gradient since $\nabla\eta(u) = \nabla\eta(v)$.

## 6.2 An Excessive Gap Technique for Non-smooth Convex Minimization

(Primal-dual problem structure; An excessive gap condition; Gradient mapping; Convergence analysis; Minimizing strongly convex functions.)

### 6.2.1 Primal-Dual Problem Structure

In this section, we give some extensions of the results presented in Sect. 6.1, where it was shown that some structured non-smooth optimization problems can be solved in $O(\frac{1}{\epsilon})$ iterations of a gradient-type scheme with $\epsilon$ being the desired accuracy of the solution. This complexity is much better than the theoretical lower complexity bound $O(\frac{1}{\epsilon^2})$ for Black-Box methods (see Sect. 3.2). This improvement, of course, is possible because of certain relaxations of the standard Black Box assumption. Instead, it was assumed that our problem has an explicit and quite simple minimax structure. However, the approach discussed in Sect. 6.1 has a certain drawback. Namely, the number of steps of the optimization scheme must be fixed in advance. It is chosen in accordance with the worst case complexity analysis and desired accuracy. Let us try to be more flexible.

Consider the same optimization problems as before:

$$\text{Find } f^* = \min_{x \in Q_1} f(x), \tag{6.2.1}$$

where $Q_1$ is a bounded closed convex set in a finite-dimensional real vector space $\mathbb{E}_1$, and $f$ is a continuous convex function on $Q_1$. We do not assume $f$ to be differentiable. Let the structure of the objective function be described by the following *model*:

$$f(x) = \hat{f}(x) + \max_{u \in Q_2} \{\langle Ax, u\rangle_{\mathbb{E}_2} - \hat{\phi}(u)\}, \tag{6.2.2}$$

where the function $\hat{f}$ is continuous and convex on $Q_1$, $Q_2$ is a closed convex bounded set in a finite-dimensional real vector space $\mathbb{E}_2$, $\hat{\phi}(\cdot)$ is a continuous convex function on $Q_2$, and the linear operator $A$ maps $\mathbb{E}_1$ to $\mathbb{E}_2^*$. In this case, problem (6.2.1) can be written in an *adjoint* form:

$$\begin{aligned} f_* &= \max_{u \in Q_2} \ \phi(u), \\ \phi(u) &= -\hat{\phi}(u) + \min_{x \in Q_1} \{\langle Ax, u\rangle_{\mathbb{E}_2} + \hat{f}(x)\}, \end{aligned} \tag{6.2.3}$$

which has zero duality gap (see (6.1.34)).

We assume that this representation is completely similar to (6.2.1) in the following sense. All methods described in this section are implementable only if the optimization problems involved in the definitions of functions $f$ and $\phi$ can be solved in a closed form. So, we assume that the structure of all objects in $\hat{f}$, $\hat{\phi}$, $Q_1$ and $Q_2$ is simple enough. We also assume that functions $\hat{f}$ and $\hat{\phi}$ have Lipschitz continuous gradients with Lipschitz constants $L_1(\hat{f})$ and $L_2(\hat{\phi})$ respectively.

Let us show that the knowledge of structure (6.2.2) can help in solving problems (6.2.1) and (6.2.3). Consider a *prox-function* $d_2(\cdot)$ of the set $Q_2$. This means that $d_2$ is continuous and strongly convex on $Q_2$ with a strong convexity parameter equal to one. Denote by

$$u_0 = \arg \min_{u \in Q_2} \ d_2(u)$$

the *prox-center* of the function $d_2$. Without loss of generality we assume that $d_2(u_0) = 0$. Thus, in view of (4.2.18), for any $u \in Q_2$ we have

$$d_2(u) \geq \frac{1}{2}\|u - u_0\|_2^2. \tag{6.2.4}$$

Let $\mu_2$ be a positive *smoothing* parameter. Consider the following function:

$$f_{\mu_2}(x) = \hat{f}(x) + \max_{u \in Q_2} \{\langle Ax, u\rangle_{\mathbb{E}_2} - \hat{\phi}(u) - \mu_2 d_2(u)\}. \tag{6.2.5}$$

Denote by $u_{\mu_2}(x)$ the optimal solution of this problem. Since the function $d_2$ is strongly convex, this solution is unique. In accordance with Danskin's theorem, the gradient of $f_{\mu_2}$ is well defined as

$$\nabla f_{\mu_2}(x) = \nabla \hat{f}(x) + A^* u_{\mu_2}(x). \tag{6.2.6}$$

Moreover, this gradient is Lipschitz-continuous with constant

$$L_1(f_{\mu_2}) = L_1(\hat{f}) + \tfrac{1}{\mu_2} \|A\|_{1,2}^2 \tag{6.2.7}$$

(see Theorem 6.1.1).

Similarly, let us consider a prox-function $d_1(\cdot)$ of the set $Q_1$, which has convexity parameter equal to one, and the prox-center $x_0$ with $d_1(x_0) = 0$. By (4.2.18), for any $x \in Q_1$ we have

$$d_1(x) \geq \frac{1}{2} \|x - x_0\|_1^2. \tag{6.2.8}$$

Let $\mu_1$ be a positive smoothing parameter. Consider

$$\phi_{\mu_1}(u) = -\hat{\phi}(u) + \min_{x \in Q_1} \{\langle Ax, u\rangle_{\mathbb{E}_2} + \hat{f}(x) + \mu_1 d_1(x)\}. \tag{6.2.9}$$

Since the second term in the above definition is a minimum of linear functions, $\phi_{\mu_1}(u)$ is concave. Denote by $x_{\mu_1}(u)$ the unique optimal solution of the above problem. In accordance with Theorem 6.1.1, the gradient

$$\nabla \phi_{\mu_1}(u) = -\nabla \hat{\phi}(u) + Ax_{\mu_1}(u) \tag{6.2.10}$$

is Lipschitz-continuous with constant

$$L_2(\phi_{\mu_1}) = L_2(\hat{\phi}) + \tfrac{1}{\mu_1} \|A\|_{1,2}^2. \tag{6.2.11}$$

### 6.2.2 *An Excessive Gap Condition*

In view of Theorem 1.3.1, for any $x \in Q_1$ and $u \in Q_2$ we have

$$\phi(u) \leq f(x), \tag{6.2.12}$$

and our assumptions guarantee no duality gap for problems (6.2.1) and (6.2.3). However, $f_{\mu_2}(x) \leq f(x)$ and $\phi(u) \leq \phi_{\mu_1}(u)$. This opens a possibility to satisfy the following *excessive gap condition*:

$$\boxed{f_{\mu_2}(\bar{x}) \;\leq\; \phi_{\mu_1}(\bar{u})} \tag{6.2.13}$$

for certain $\bar{x} \in Q_1$ and $\bar{u} \in Q_2$. Let us show that condition (6.2.13) provides us with an upper bound on the quality of the primal-dual pair $(\bar{x}, \bar{u})$.

**Lemma 6.2.1** *Let $\bar{x} \in Q_1$ and $\bar{u} \in Q_2$ satisfy (6.2.13). Then*

$$\begin{aligned} 0 &\le \max\{f(\bar{x}) - f^*, f^* - \phi(\bar{u})\} \\ &\le f(\bar{x}) - \phi(\bar{u}) \ \le\ \mu_1 D_1 + \mu_2 D_2, \end{aligned} \tag{6.2.14}$$

*where $D_1 = \max\limits_{x \in Q_1} d_1(x)$, and $D_2 = \max\limits_{u \in Q_2} d_2(u)$.*

*Proof* Indeed, for any $\bar{x} \in Q_1$, $\bar{u} \in Q_2$ we have

$$f(\bar{x}) - \mu_2 D_2 \le f_{\mu_2}(\bar{x}) \overset{(6.2.13)}{\le} \phi_{\mu_1}(\bar{u}) \ \le\ \phi(\bar{u}) + \mu_1 D_1.$$

It remains to apply inequality (6.2.12). □

Our goal is to justify a process for recursively updating the pair $(\bar{x}, \bar{u})$, which maintains inequality (6.2.13) as $\mu_1$ and $\mu_2$ go to zero. Before we start our analysis, let us prove a useful inequality.

**Lemma 6.2.2** *For any $x$ and $\hat{x}$ from $Q_1$ we have:*

$$f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), x - \hat{x} \rangle_{\mathbb{E}_1} \le \hat{f}(x) + \langle Ax, u_{\mu_2}(\hat{x}) \rangle_{\mathbb{E}_2} - \hat{\phi}(u_{\mu_2}(\hat{x})). \tag{6.2.15}$$

*Proof* Let us take arbitrary $x$ and $\hat{x}$ from $Q_1$. Let $\hat{u} = u_{\mu_2}(\hat{x})$. Then

$$\begin{aligned} f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), x - \bar{y} \rangle_{\mathbb{E}_1} &\overset{(6.2.5),(6.2.6)}{=} \hat{f}(\hat{x}) + \langle A\bar{y}, \hat{u} \rangle_{\mathbb{E}_2} - \hat{\phi}(\hat{u}) - \mu_2 d_2(\hat{u}) \\ &\quad + \langle \nabla \hat{f}(\hat{x}) + A^* \hat{u}, x - \hat{x} \rangle_{\mathbb{E}_1} \\ &\overset{(2.1.2)}{\le} \hat{f}(x) + \langle Ax, \hat{u} \rangle_{\mathbb{E}_2} - \hat{\phi}(\hat{u}). \end{aligned}$$

□

Let us justify the possibility of satisfying the excessive gap condition (6.2.13) at some starting primal-dual pair.

**Lemma 6.2.3** *Let us choose an arbitrary $\mu_2 > 0$ and set*

$$\begin{aligned} \bar{x} &= \arg\min_{x \in Q_1} \{\langle \nabla f_{\mu_2}(x_0)), x - x_0 \rangle_{\mathbb{E}_1} + L_1(f_{\mu_2}) d_1(x)\}, \\ \bar{u} &= u_{\mu_2}(x_0). \end{aligned} \tag{6.2.16}$$

*Then the excessive gap condition is satisfied for any $\mu_1 \ge L_1(f_{\mu_2})$.*

*Proof* Indeed, in view of (1.2.11) we have

$$\begin{aligned} f_{\mu_2}(\bar{x}) \quad &\leq \quad f_{\mu_2}(x_0) + \langle \nabla f_{\mu_2}(x_0), \bar{x} - x_0 \rangle_{\mathbb{E}_1} + \tfrac{1}{2} L_1(f_{\mu_2}) \| \bar{x} - x_0 \|_1^2 \\ &\overset{(6.2.4)}{\leq} \quad f_{\mu_2}(x_0) + \langle \nabla f_{\mu_2}(x_0), \bar{x} - x_0 \rangle_{\mathbb{E}_1} + \tfrac{1}{2} L_1(f_{\mu_2}) d_1(\bar{x}) \\ &\overset{(6.2.16)}{=} \quad f_{\mu_2}(x_0) + \min_{x \in Q_1} \{ \langle \nabla f_{\mu_2}(x_0), x - x_0 \rangle_{\mathbb{E}_1} + L_1(f_{\mu_2}) d_1(x) \} \\ &\overset{(6.2.15)}{\leq} \quad \min_{x \in Q_1} \left\{ \hat{f}(x) + \langle Ax, u_{\mu_2}(x_0) \rangle_{\mathbb{E}_2} - \hat{\phi}(u_{\mu_2}(x_0)) + L_1(f_{\mu_2}) d_1(x) \right\} \\ &\overset{(6.2.9)}{=} \quad \phi_{L_1(f_{\mu_2})}(\bar{u}) \ \leq \ \phi_{\mu_1}(\bar{u}). \qquad \square \end{aligned}$$

Thus, condition (6.2.13) can be satisfied for some primal-dual pair. Let us show how we can update the points $\bar{x}$ and $\bar{u}$ in order to keep it valid for smaller values of $\mu_1$ and $\mu_2$. In view of the symmetry of the situation, at the first step of the process we can try to decrease only $\mu_1$, keeping $\mu_2$ unchanged. After that, at the second step, we update $\mu_2$ and keep $\mu_1$ constant, and so on. The main advantage of such a switching strategy is that we need to find a justification only for the first step. The proof for the second one will be symmetric.

**Theorem 6.2.1** *Let points $\bar{x} \in Q_1$ and $\bar{u} \in Q_2$ satisfy the excessive gap condition (6.2.13) for some positive $\mu_1$ and $\mu_2$. Let us fix $\tau \in (0, 1)$ and choose $\mu_1^+ = (1 - \tau)\mu_1$,*

$$\begin{aligned} \hat{x} \ &= (1 - \tau)\bar{x} + \tau x_{\mu_1}(\bar{u}), \\ \bar{u}_+ &= (1 - \tau)\bar{u} + \tau u_{\mu_2}(\hat{x}), \\ \bar{x}_+ &= (1 - \tau)\bar{x} + \tau x_{\mu_1^+}(\bar{u}_+). \end{aligned} \tag{6.2.17}$$

*Then the pair $(\bar{x}_+, \bar{u}_+)$ satisfies condition (6.2.13) with smoothing parameters $\mu_1^+$ and $\mu_2$ provided that $\tau$ satisfies the following relation:*

$$\boxed{\tfrac{\tau^2}{1-\tau} \leq \tfrac{\mu_1}{L_1(f_{\mu_2})}} \tag{6.2.18}$$

*Proof* Let $\hat{u} = u_{\mu_2}(\hat{x})$, $x_1 = x_{\mu_1}(\bar{u})$, and $\tilde{x}_+ = x_{\mu_1^+}(\bar{u}_+)$. Since $\hat{\phi}$ is convex, in view of the operation in (6.2.17), we have $\hat{\phi}(\bar{u}_+) \le (1-\tau)\hat{\phi}(\bar{u}) + \tau\hat{\phi}(\hat{u})$. Therefore,

$$\begin{aligned}
\phi_{\mu_1^+}(\bar{u}_+) &= (1-\tau)\mu_1 d_1(\tilde{x}_+) + \langle A\tilde{x}_+, (1-\tau)\bar{u} + \tau\hat{u}\rangle_{\mathbb{E}_2} + \hat{f}(\tilde{x}_+) - \hat{\phi}(\bar{u}_+) \\
&\ge (1-\tau)[\mu_1 d_1(\tilde{x}_+) + \langle A\tilde{x}_+, \bar{u}\rangle_{\mathbb{E}_2} + \hat{f}(\tilde{x}_+) - \hat{\phi}(\bar{u})] \\
&\quad + \tau[\hat{f}(\tilde{x}_+) + \langle A\tilde{x}_+, \hat{u}\rangle_{\mathbb{E}_2} - \hat{\phi}(\hat{u})] \\
&\overset{(6.2.15)}{\ge} (1-\tau)[\phi_{\mu_1}(\bar{u}) + \tfrac{1}{2}\mu_1\|\tilde{x}_+ - x_1\|_1^2]_a \\
&\quad + \tau[f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), \tilde{x}_+ - \hat{x}\rangle_{\mathbb{E}_1}]_b.
\end{aligned}$$

Note that in view of condition (6.2.13) and the first line in (6.2.17) we have

$$\begin{aligned}
\phi_{\mu_1}(\bar{u}) \ge f_{\mu_2}(\bar{x}) &\ge f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), \bar{x} - \hat{x}\rangle_{\mathbb{E}_1} \\
&= f_{\mu_2}(\hat{x}) + \tau\langle \nabla f_{\mu_2}(\hat{x}), \bar{x} - x_1\rangle_{\mathbb{E}_1}.
\end{aligned}$$

Therefore, we can estimate the expression in the first brackets as follows:

$$[\,\cdot\,]_a \ge f_{\mu_2}(\hat{x}) + \tau\langle \nabla f_{\mu_2}(\hat{x}), \bar{x} - x_1\rangle_{\mathbb{E}_1} + \tfrac{1}{2}\mu_1\|\tilde{x}_+ - x_1\|_1^2.$$

In view of the first line in (6.2.15), for second brackets we have

$$[\,\cdot\,]_b = f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), \tilde{x}_+ - x_1 + (1-\tau)(x_1 - \bar{x})\rangle_{\mathbb{E}_1}.$$

Thus, taking into account that $\bar{x}_+ - \hat{x} \overset{(6.2.17)}{=} \tau(\tilde{x}_+ - x_1)$, we finish the proof as follows:

$$\begin{aligned}
\phi_{\mu_1^+}(\bar{u}_+) &\ge f_{\mu_2}(\hat{x}) + \tau\langle \nabla f_{\mu_2}(\hat{x}), \tilde{x}_+ - x_1\rangle_{\mathbb{E}_1} + \tfrac{1}{2}(1-\tau)\mu_1\|\tilde{x}_+ - x_1\|_1^2 \\
&= f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), \bar{x}_+ - \hat{x}\rangle_{\mathbb{E}_1} + \tfrac{(1-\tau)\mu_1}{2\tau^2}\|\bar{x}_+ - \hat{x}\|_1^2 \\
&\overset{(6.2.18)}{\ge} f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), \bar{x}_+ - \hat{x}\rangle_{\mathbb{E}_1} + \tfrac{1}{2}L_1(f_{\mu_2})\|\bar{x}_+ - \hat{x}\|_1^2 \\
&\overset{(1.2.11)}{\ge} f_{\mu_2}(\bar{x}_+). \qquad \square
\end{aligned}$$

### 6.2.3 Convergence Analysis

In Sect. 6.2.2, we have seen that the smoothness parameters $\mu_1$ and $\mu_2$ can be decreased by a switching strategy. Thus, in order to transform the result of Theorem 6.2.1 into an algorithmic scheme, we need to point out a strategy for updating these parameters, which is compatible with the growth condition (6.2.18). In this section, we do this for an important case $L_1(\hat{f}) = L_2(\hat{\phi}) = 0$.

It is convenient to represent the smoothness parameters as follows:

$$\mu_1 = \lambda_1 \cdot \|A\|_{1,2} \cdot \sqrt{\tfrac{D_2}{D_1}}, \quad \mu_2 = \lambda_2 \cdot \|A\|_{1,2} \cdot \sqrt{\tfrac{D_1}{D_2}}. \tag{6.2.19}$$

Then the estimate (6.2.14) for the duality gap becomes symmetric:

$$f(\bar{x}) - \phi(\bar{u}) \leq (\lambda_1 + \lambda_2) \cdot \|A\|_{1,2} \cdot \sqrt{D_1 D_2}. \tag{6.2.20}$$

Since by (6.2.7), $L_1(f_{\mu_2}) = \frac{1}{\mu_2}\|A\|^2_{1,2}$, condition (6.2.18) becomes problem independent:

$$\tfrac{\tau^2}{1-\tau} \leq \mu_1\mu_2 \cdot \tfrac{1}{\|A\|^2_{1,2}} = \lambda_1\lambda_2. \tag{6.2.21}$$

Let us write down the corresponding switching algorithmic scheme in an explicit form. It is convenient to have a permanent iteration counter. In this case, at even iterations we apply the primal update (6.2.17), and at odd iterations the corresponding dual update is used. Since at even iterations $\lambda_2$ does not change and at odd iterations $\lambda_1$ does not change it is convenient to put their new values in the same sequence $\{\alpha_k\}_{k=-1}^{\infty}$. Let us fix the following relations between the sequences:

$$\begin{aligned} k = \quad 2l \quad &: \lambda_{1,k} = \alpha_{k-1},\ \lambda_{2,k} = \alpha_k, \\ k = 2l+1 &: \lambda_{1,k} = \alpha_k, \quad \lambda_{2,k} = \alpha_{k-1}. \end{aligned} \tag{6.2.22}$$

Then the corresponding parameters $\tau_k$ (see the rule (6.2.1)) define the reduction rate of the sequence $\{\alpha_k\}_{k=-1}^{\infty}$.

**Lemma 6.2.4** *For all $k \geq 0$ we have $\alpha_{k+1} = (1-\tau_k)\alpha_{k-1}$.*

*Proof* Indeed, in accordance with (6.2.22), if $k = 2l$, then

$$\alpha_{k+1} = \lambda_{1,k+1} = (1-\tau_k)\lambda_{1,k} = (1-\tau_k)\alpha_{k-1}.$$

And if $k = 2l+1$, then $\alpha_{k+1} = \lambda_{2,k+1} = (1-\tau_k)\lambda_{2,k} = (1-\tau_k)\alpha_{k-1}$. □

**Corollary 6.2.1** *In terms of the sequence* $\{\alpha_k\}_{k=-1}^{\infty}$, *condition* (6.2.21) *is as follows:*

$$(\alpha_{k+1} - \alpha_{k-1})^2 \le \alpha_{k+1}\alpha_k\alpha_{k-1}^2, \quad k \ge 0. \tag{6.2.23}$$

*Proof* In view of (6.2.22), we always have $\lambda_{1,k}\lambda_{2,k} = \alpha_k\alpha_{k-1}$. Since $\tau_k = 1 - \frac{\alpha_{k+1}}{\alpha_{k-1}}$, we get (6.2.23). □

Clearly, condition (6.2.23) is satisfied by

$$\alpha_k = \tfrac{2}{k+2}, \quad k \ge -1. \tag{6.2.24}$$

Then

$$\tau_k = 1 - \tfrac{\alpha_{k+1}}{\alpha_{k-1}} = \tfrac{2}{k+3}, \quad k \ge 0. \tag{6.2.25}$$

Now we are ready to write down an algorithmic scheme. Let us do this for the rule (6.2.17). In this scheme, we use the sequences $\{\mu_{1,k}\}_{k=-1}^{\infty}$ and $\{\mu_{2,k}\}_{k=-1}^{\infty}$, generated in accordance with rules (6.2.19), (6.2.22) and (6.2.24).

**1. Initialization:** Choose $\bar{x}_0$ and $\bar{u}_0$ in accodance with (6.2.16) taking $\mu_1 = \mu_{1,0}$ and $\mu_2 = \mu_{2,0}$.

**2. Iterations** ($k \ge 0$)**:**

(a) Set $\tau_k = \frac{2}{k+3}$.

(b) If $k$ is even, then generate $(\bar{x}_{k+1}, \bar{u}_{k+1})$ from $(\bar{x}_k, \bar{u}_k)$ using (6.2.17).

(c) If $k$ is odd, then generate $(\bar{x}_{k+1}, \bar{u}_{k+1})$ from $(\bar{x}_k, \bar{u}_k)$ using the symmetric dual variant of (6.2.17).

(6.2.26)

**Theorem 6.2.2** *Let the sequences* $\{\bar{x}_k\}_{k=0}^{\infty}$ *and* $\{\bar{u}_k\}_{k=0}^{\infty}$ *be generated by method* (6.2.26). *Then each pair of points* $(\bar{x}_k, \bar{u}_k)$ *satisfy the excessive gap condition. Therefore,*

$$f(\bar{x}_k) - \phi(\bar{u}_k) \le \tfrac{4\|A\|_{1,2}}{k+1}\sqrt{D_1 D_2}. \tag{6.2.27}$$

*Proof* In accordance with our choice of parameters,

$$\mu_{1,0}\mu_{2,0} = \lambda_{1,0}\lambda_{2,0} \cdot \|A\|_{1,2}^2 \;=\; 2\mu_{2,0}L_1(f_{\mu_{2,0}}) \;>\; \mu_{2,0}L_1(f_{\mu_{2,0}}).$$

Hence, in view of Lemma 6.2.3 the pair $(\bar{x}_0, \bar{u}_0)$ satisfies the excessive gap condition. We have already checked that the sequence $\{\tau_k\}_{k=0}^{\infty}$ defined by (6.2.25) satisfies the conditions of Theorem 6.2.1. Therefore, excessive gap conditions will be valid for the sequences generated by (6.2.26). It remains to use inequality (6.2.20). □

### 6.2.4 Minimizing Strongly Convex Functions

Consider now the model (6.2.2), which satisfies the following assumption.

**Assumption 6.2.1** *In representation (6.2.2) the function $\hat{f}$ is strongly convex with convexity parameter $\hat{\sigma} > 0$.*

Let us prove the following variant of Danskin's theorem.

**Lemma 6.2.5** *Under Assumption 6.2.1 the function $\phi$ defined by (6.2.3) is concave and differentiable. Moreover, its gradient*

$$\nabla\phi(u) = -\nabla\hat{\phi}(u) + Ax_0(u), \tag{6.2.28}$$

*where $x_0(u)$ is defined by (6.2.9), is Lipschitz-continuous with constant*

$$L_2(\phi) = \tfrac{1}{\hat{\sigma}}\|A\|^2_{1,2} + L_2(\hat{\phi}). \tag{6.2.29}$$

*Proof* Let $\tilde{\phi}(u) = \min\limits_{x \in Q_1}\{\langle Ax, u\rangle_{\mathbb{E}_2} + \hat{f}(x)\}$. This function is concave as a minimum of linear functions. Since $\hat{f}$ is strongly convex, the solution of the latter minimization problem is unique. Therefore, $\tilde{\phi}(\cdot)$ is differentiable and $\nabla\tilde{\phi}(u) = Ax_0(u)$.

Consider two points $u_1$ and $u_2$. From the first-order optimality conditions for (6.2.3) we have

$$\langle A^*u_1 + \nabla\hat{f}(x_0(u_1)), x_0(u_2) - x_0(u_1)\rangle_{\mathbb{E}_1} \geq 0,$$

$$\langle A^*u_2 + \nabla\hat{f}(x_0(u_2)), x_0(u_1) - x_0(u_2)\rangle_{\mathbb{E}_1} \geq 0.$$

Adding these inequalities and using the strong convexity of $\hat{f}(\cdot)$, we continue as follows:

$$\begin{aligned} & \langle Ax_0(u_2) - Ax_0(u_1), u_1 - u_2\rangle_{\mathbb{E}_2} \\ \geq\ & \langle \nabla\hat{f}(x_0(u_1)) - \nabla\hat{f}(x_0(u_2)), x_0(u_1) - x_0(u_2)\rangle_{\mathbb{E}_1} \\ \overset{(2.1.22)}{\geq}\ & \hat{\sigma}\|x_0(u_1) - x_0(u_2)\|^2_{\mathbb{E}_1} \overset{(6.1.9)}{\geq} \frac{\hat{\sigma}}{\|A\|^2_{1,2}}\left(\|\nabla\tilde{\phi}(u_1) - \nabla\tilde{\phi}(u_2)\|^*_{\mathbb{E}_2}\right)^2. \end{aligned}$$

Thus, $\|\nabla\tilde{\phi}(u_1) - \nabla\tilde{\phi}(u_2)\|^*_{|E_2} \leq \frac{1}{\hat{\sigma}}\|A\|^2_{1,2} \cdot \|u_1 - u_2\|_{\mathbb{E}_2}$, and (6.2.29) follows. □

**Lemma 6.2.6** *For any $u$ and $\hat{u}$ from $Q_2$, we have:*

$$\phi(\hat{u}) + \langle \nabla\phi(\hat{u}), u - \hat{u}\rangle_{\mathbb{E}_2} \;\geq\; -\hat{\phi}(u) + \langle Ax_0(\hat{u}), u\rangle_{\mathbb{E}_2} + \hat{f}(x_0(\hat{u})). \tag{6.2.30}$$

*Proof* Let us take arbitrary $u$ and $\hat{u}$ from $Q_2$. Define $\hat{x} = x_0(\hat{u})$. Then

$$\begin{aligned} & \phi(\hat{u}) + \langle \nabla\phi(\hat{u}), u - \hat{u}\rangle_{\mathbb{E}_2} \\ = \;& -\hat{\phi}(\hat{u}) + \langle A\hat{x}, \hat{u}\rangle_{\mathbb{E}_2} + \hat{f}(\hat{x}) + \langle -\nabla\hat{\phi}(\hat{u}) + A\hat{x}, u - \hat{u}\rangle_{\mathbb{E}_2} \\ \overset{(2.1.2)}{\geq} \;& -\hat{\phi}(u) + \langle A\hat{x}, u\rangle_{\mathbb{E}_2} + \hat{f}(\hat{x}). \qquad \square \end{aligned}$$

In this section, we derive an optimization scheme from the following variant of excessive gap condition:

$$\boxed{f_{\mu_2}(\bar{x}) \;\leq\; \phi(\bar{u})} \tag{6.2.31}$$

for some $\bar{x} \in Q_1$ and $\bar{u}$ in $Q_2$.

This condition can be seen as a variant of condition (6.2.13) with $\mu_1 = 0$. However, we prefer not to use the results of the previous sections since our assumptions will be slightly different. For example, we no longer need the set $Q_1$ to be bounded.

**Lemma 6.2.7** *Let points $\bar{x}$ from $Q_1$ and $\bar{u}$ from $Q_2$ satisfy condition (6.2.31). Then*

$$0 \leq f(\bar{x}) - \phi(\bar{u}) \;\leq\; \mu_2 D_2. \tag{6.2.32}$$

*Proof* Indeed, for any $x \in Q_1$, we have $f_{\mu_2}(x) \geq f(x) - \mu_2 D_2$. □

Define the adjoint gradient mapping as follows:

$$V(u) = \arg\max_{v\in Q_2} \left\{ \langle \nabla\phi(u), v - u\rangle_{\mathbb{E}_2} - \frac{1}{2} L_2(\phi)\|v - u\|^2_{\mathbb{E}_2} \right\}. \tag{6.2.33}$$

**Lemma 6.2.8** *The excessive gap condition (6.2.31) is valid for $\mu_2 = L_2(\phi)$ and*

$$\bar{x} \;=\; x_0(u_0), \quad \bar{u} \;=\; V(u_0). \tag{6.2.34}$$

*Proof* Indeed, in view of Lemma 6.2.5 and (1.2.11), we get the following relations:

$$\begin{aligned}
\phi(V(u_0)) \quad &\geq \quad \phi(u_0) + \langle \nabla\phi(u_0), V(u_0) - u_0\rangle_{\mathbb{E}_2} - \frac{1}{2}L_2(\phi)\|V(u_0) - u_0\|_2^2 \\
&\overset{(6.2.33)}{=} \quad \max_{u\in Q_2}\left\{\phi(u_0) + \langle \nabla\phi(u_0), u - u_0\rangle_{\mathbb{E}_2} - \frac{1}{2}L_2(\phi)\|u - u_0\|_2^2\right\} \\
&\overset{(6.2.3),(6.2.28)}{=} \max_{u\in Q_2}\left\{-\hat{\phi}(u_0) + \langle Ax_0(u_0), u_0\rangle_{\mathbb{E}_2} + \hat{f}(x_0(u_0))\right. \\
&\qquad \left. +\langle Ax_0(u_0) - \nabla\hat{\phi}(u_0), u - u_0\rangle_{\mathbb{E}_2} - \frac{1}{2}\mu_2\|u - u_0\|_2^2\right\} \\
&\overset{(6.2.4)}{\geq} \quad \max_{u\in Q_2}\left\{-\hat{\phi}(u) + \hat{f}(x_0(u_0)) + \langle Ax_0(u_0), u\rangle_{\mathbb{E}_2} - \mu_2 d_2(u)\right\} \\
&\overset{(6.2.5)}{=} \quad f_{\mu_2}(x_0(u_0)).
\end{aligned}$$

□

**Theorem 6.2.3** *Let points $\bar{x} \in Q_1$ and $\bar{u} \in Q_2$ satisfy the excessive gap condition (6.2.31) for some positive $\mu_2$. Let us fix $\tau \in (0,1)$ and choose $\mu_2^+ = (1-\tau)\mu_2$,*

$$\begin{aligned}
\hat{u} \; &= (1-\tau)\bar{u} + \tau u_{\mu_2}(\bar{x}), \\
\bar{x}_+ &= (1-\tau)\bar{x} + \tau x_0(\hat{u}), \qquad (6.2.35) \\
\bar{u}_+ &= V(\hat{u}).
\end{aligned}$$

*Then the pair $(\bar{x}_+, \bar{u}_+)$ satisfies condition (6.2.31) with smoothness parameter $\mu_2^+$, provided that $\tau$ satisfies the following growth relation:*

$$\tfrac{\tau^2}{1-\tau} \leq \tfrac{\mu_2}{L_2(\phi)}. \qquad (6.2.36)$$

*Proof* Let $\hat{x} = x_0(\hat{u})$ and $u_2 = u_{\mu_2}(\bar{x})$. In view of the second rule in (6.2.35), and (6.2.5), we have:

$$\begin{aligned}
f_{\mu_2^+}(\bar{x}_+) &= \hat{f}(\bar{x}_+) + \max_{u \in Q_2} \Big\{ \langle A((1-\tau)\bar{x} + \tau\hat{x}), u\rangle_{\mathbb{E}_2} - \hat{\phi}(u) \\
&\quad -(1-\tau)\mu_2 d_2(u) \} \\
&\overset{(3.1.2)}{\le} \max_{u \in Q_2} \Big\{ (1-\tau)\Big[ \hat{f}(\bar{x}) + \langle A\bar{x}, u\rangle_{\mathbb{E}_2} - \hat{\phi}(u) - \mu_2 d_2(u) \Big] \\
&\quad +\tau[\hat{f}(\hat{x}) + \langle A\hat{x}, u\rangle_{\mathbb{E}_2} - \hat{\phi}(u)] \Big\} \\
&\overset{(4.2.18)}{\le} \max_{u \in Q_2} \Big\{ (1-\tau)\Big[ f_{\mu_2}(\bar{x}) - \tfrac{1}{2}\mu_2 \|u - u_2\|_2^2 \Big] \\
&\quad +\tau[\phi(\hat{u}) + \langle \nabla\phi(\hat{u}), u - \hat{u}\rangle_{\mathbb{E}_2}] \Big\},
\end{aligned}$$

where we used (6.2.30) in the last line. Since $\phi$ is concave, by (6.2.31) we obtain

$$\begin{aligned}
f_{\mu_2}(\bar{x}) &\le \phi(\bar{u}) \le \phi(\hat{u}) + \langle \nabla\phi(\hat{u}), \bar{u} - \hat{u}\rangle_{\mathbb{E}_2} \\
&\overset{\text{Line 1 in (6.2.35)}}{=} \phi(\hat{u}) + \tau\langle \nabla\phi(\hat{u}), \bar{u} - u_2\rangle_{\mathbb{E}_2}.
\end{aligned}$$

Hence, we can finish the proof as follows:

$$\begin{aligned}
f_{\mu_2^+}(\bar{x}_+) &\le \max_{u \in Q_2} \Big\{ \phi(\hat{u}) + \tau\langle \nabla\phi(\hat{u}), u - u_2\rangle_{\mathbb{E}_2} - \tfrac{1}{2}(1-\tau)\mu_2\|u - u_2\|_2^2 \Big\} \\
&\overset{(6.2.36)}{\le} \max_{u \in Q_2} \Big\{ \phi(\hat{u}) + \tau\langle \nabla\phi(\hat{u}), u - u_2\rangle_{\mathbb{E}_2} - \tfrac{1}{2}\tau^2 L_2(\phi)\|u - u_2\|_2^2 \Big\}.
\end{aligned}$$

Defining now $v = \bar{u} + \tau(u - \bar{u}))$ with $u \in Q_2$, we continue:

$$\begin{aligned}
f_{\mu_2^+}(\bar{x}_+) &\le \max_{v \in \bar{u} + \tau(Q_2 - \bar{u})} \Big\{ \phi(\hat{u}) + \langle \nabla\phi(\hat{u}), v - \hat{u}\rangle_{\mathbb{E}_2} - \tfrac{1}{2}L_2(\phi)\|v - \hat{u}\|_2^2 \Big\} \\
(Q_2 \text{ is convex}) &\le \max_{v \in Q_2} \Big\{ \phi(\hat{u}) + \langle \nabla\phi(\hat{u}), v - \hat{u}\rangle_{\mathbb{E}_2} - \tfrac{1}{2}L_2(\phi)\|v - \hat{u}\|_2^2 \Big\} \\
&\overset{(6.2.33)}{\le} \phi(\hat{u}) + \langle \nabla\phi(\hat{u}), \bar{u}_+ - \hat{u}\rangle_{\mathbb{E}_2} - \tfrac{1}{2}L_2(\phi)\|\bar{u}_+ - \hat{u}\|_2^2 \\
&\overset{(1.2.11)}{\le} \phi(\bar{u}_+). \qquad \square
\end{aligned}$$

Now we can justify the following minimization scheme.

```
1. Initialization:
   Set $\mu_{2,0} = 2L_2(\phi)$, $\bar{x}_0 = x_0(u_0)$ and $\bar{u}_0 = V(u_0)$.

2. For $k \geq 0$ iterate:

    Set $\tau_k = \frac{2}{k+3}$ and $\hat{u}_k = (1-\tau_k)\bar{u}_k + \tau_k u_{\mu_{2,k}}(\bar{x}_k)$.

    Update $\mu_{2,k+1} = (1-\tau_k)\mu_{2,k}$,

           $\bar{x}_{k+1} = (1-\tau_k)\bar{x}_k + \tau_k x_0(\hat{u}_k)$,

           $\bar{u}_{k+1} = V(\hat{u}_k)$.
```

(6.2.37)

**Theorem 6.2.4** *Let problem (6.2.1) satisfy Assumption 6.2.1. Then the pairs $(\bar{x}_k, \bar{u}_k)$ generated by scheme (6.2.37) satisfy the following inequality:*

$$f(\bar{x}_k) - \phi(\bar{u}_k) \leq \tfrac{4L_2(\phi)D_2}{(k+1)(k+2)}, \tag{6.2.38}$$

*where $L_2(\phi)$ is given by (6.2.29).*

*Proof* Indeed, in view of Theorem 6.2.3 and Lemma 6.2.8 we need only to justify that the sequences $\{\mu_{2,k}\}_{k=0}^{\infty}$ and $\{\tau_k\}_{k=0}^{\infty}$ satisfy relation (6.2.36). This is straightforward because of the following relation:

$$\mu_{2,k} = \tfrac{4L_2(\phi)}{(k+1)(k+2)},$$

which is valid for all $k \geq 0$. □

Let us conclude this section with an example. Consider the problem

$$f(x) = \frac{1}{2}\|x\|^2_{\mathbb{E}_1} + \max_{1\leq j\leq m} [f_j + \langle g_j, x - x_j\rangle_{\mathbb{E}_1}] \quad \to \quad \min : \; x \in \mathbb{E}_1. \tag{6.2.39}$$

Let $\mathbb{E}_1 = \mathbb{R}^n$ and choose

$$\|x\|_1^2 = \sum_{i=1}^{n} (x^{(i)})^2, \quad x \in \mathbb{E}_1.$$

Then this problem can be solved by the method (6.2.37).

Indeed, we can represent the objective function in (6.2.39) in the form (6.2.2) using the following objects:

$$\mathbb{E}_2 = \mathbb{R}^m, \quad Q_2 \;=\; \Delta_m = \{u \in \mathbb{R}^m_+ :\; \sum_{j=1}^m u^{(j)} = 1\},$$

$$\hat{f}(x) = \tfrac{1}{2}\|x\|_1^2, \quad \hat{\phi}(u) \;=\; \langle b, u\rangle_{\mathbb{E}_2}, \quad b^{(j)} \;=\; \langle g_j, x_j\rangle_{\mathbb{E}_1} - f_j,\; j = 1, \dots, m,$$

$$A^T \;=\; (g_1, \dots, g_m).$$

Thus, $\hat{\sigma} = 1$ and $L_2(\hat{\phi}) = 0$. Let us choose for $\mathbb{E}_2$ the following norm:

$$\|u\|_{\mathbb{E}_2} = \sum_{j=1}^m |u^{(j)}|.$$

Then we can use the entropy distance function,

$$d_2(u) = \ln m + \sum_{j=1}^m u^{(j)} \ln u^{(j)}, \quad u_0 = (\tfrac{1}{m}, \dots, \tfrac{1}{m}),$$

for which the convexity parameter is one and $D_2 = \ln m$. Note that in this case

$$\|A\|_{1,2} \;=\; \max_{1\le j\le m} \|g_j\|_1^*.$$

Thus, method (6.2.37) as applied to problem (6.2.39) converges with the following rate:

$$f(\bar{x}_k) - \phi(\bar{u}_k) \le \tfrac{4\ln m}{(k+1)(k+2)} \cdot \max_{1\le j\le m} \left(\|g_j\|_1^*\right)^2.$$

Let us study the complexity of method (6.2.37) for our example. At each iteration, we need to compute the following objects.

1. **Computation of** $u_{\mu_2}(\bar{x})$. This is the solution of the following problem:

$$\max_u \left\{ \sum_{j=1}^m u^{(j)} s^{(j)}(\bar{x}) - \mu_2 d_2(u) :\; u \in Q_2 \right\}$$

with $s^{(j)}(\bar{x}) = f_j + \langle g_j, \bar{x} - x_j\rangle$, $j = 1, \dots, m$. As we have seen several times, this solution can be found in a closed form:

$$u^{(j)}_{\mu_2}(\bar{x}) = e^{s^{(j)}(\bar{x})/\mu_2} \cdot \left[\sum_{l=1}^m e^{s^{(l)}(\bar{x})/\mu_2}\right]^{-1}, \quad j = 1, \dots, m.$$

2. **Computation of** $x_0(\hat{u})$. In our case, this is a solution to the problem

$$\min_x \left\{ \langle Ax, \hat{u}\rangle_{\mathbb{E}_2} + \frac{1}{2}\|x\|^2_{\mathbb{E}_1} \,:\, x \in \mathbb{E}_1 \right\}.$$

   Hence, the answer is very simple: $x_0(\hat{u}) = -A^T\hat{u}$.
3. **Computation of** $V(\hat{u})$. In our case,

$$\phi(\bar{u}) = \min_{x\in\mathbb{E}_1} \left\{ \sum_{j=1}^{m} u^{(j)}[f_j + \langle g_j, x - x_j\rangle_{\mathbb{E}_1}] + \tfrac{1}{2}\|x\|^2_{\mathbb{E}_1} \right\}$$

$$= -\langle b, u\rangle_{\mathbb{E}_2} - \tfrac{1}{2}\left(\|A^T\hat{u}\|^*_{\mathbb{E}_1}\right)^2.$$

   Thus, $\nabla\phi(\bar{u}) = -b - AA^T\hat{u}$. Now we can compute $V(\hat{u})$ by (6.2.33). It can be easily shown that the complexity of finding $V(\hat{u})$ is of the order $O(m \ln m)$, which comes from the necessity to sort the components of a vector in $\mathbb{R}^m$.

Thus, we have seen that all computations at each iteration of method (6.2.37) as applied to problem (6.2.39) are very cheap. The most expensive part of the iteration is the multiplication of matrix $A$ by a vector. In a straightforward implementation, we need three such multiplications per iteration. However, a simple modification of the order of operations can reduce this amount to two.

## 6.3 The Smoothing Technique in Semidefinite Optimization

(Smooth symmetric functions of eigenvalues; Minimizing the maximal eigenvalue of a symmetric matrix.)

### *6.3.1 Smooth Symmetric Functions of Eigenvalues*

In Sects. 6.1 and 6.2, we have shown that a proper use of the structure of nonsmooth convex optimization problems leads to very efficient gradient schemes, whose performance is significantly better than the lower complexity bounds derived from the Black Box assumptions. However, this observation leads to implementable algorithms only if we are able to form a computable smooth approximation of the objective function of our problem. In this case, applying to this approximation an optimal method (6.1.19) for minimizing smooth convex functions, we can easily obtain a good solution to our initial problem.

Our previous results are related mainly to piece-wise linear functions. In this section, we extend them to the problems of Semidefinite Optimization (SO).

For that, we introduce computable smooth approximation for one of the most important nonsmooth functions of symmetric matrices, its *maximal eigenvalue*. Our approximation is based on entropy smoothing.

In what follows, we denote by $\mathbb{M}_n$ the space of real $n \times n$-matrices, and by $\mathbb{S}_n \subset \mathbb{M}_n$ the space of symmetric matrices. A particular matrix is always denoted by a capital letter. In the spaces $\mathbb{R}^n$ and $\mathbb{M}_n$ we use the standard inner products

$$\langle x, y \rangle = \sum_{i=1}^{n} x^{(i)} y^{(i)}, \;\; x, y \in \mathbb{R}^n,$$

$$\langle X, Y \rangle_F = \sum_{i,j=1}^{n} X^{(i,j)} Y^{(i,j)}, \;\; X, Y \in \mathbb{M}_n.$$

For $X \in \mathbb{S}_n$, we denote by $\lambda(X) \in \mathbb{R}^n$ the vector of its eigenvalues. We assume that the eigenvalues are ordered in a decreasing order:

$$\lambda^{(1)}(X) \geq \lambda^{(2)}(X) \geq \cdots \geq \lambda^{(n)}(X), \quad X \in \mathbb{S}_n.$$

Thus, $\lambda_{\max}(X) = \lambda^{(1)}(X)$. The notation $D(\lambda) \in \mathbb{S}_n$ is used for a diagonal matrix with vector $\lambda \in \mathbb{R}^n$ on the main diagonal. Note that any $X \in \mathbb{S}_n$ admits an eigenvalue decomposition

$$X = U(X) D(\lambda(X)) U(X)^T$$

with $U(X)U(X)^T = I_n$, where $I_n \in \mathbb{S}_n$ is the identity matrix.

Let us mention some notations with different meanings for vectors and matrices. For a vector $\lambda \in \mathbb{R}^n$, we denote by $|\lambda| \in \mathbb{R}^n$ the vector with entries $|\lambda^{(i)}|$, $i = 1, \ldots, n$. The notation $\lambda^k \in \mathbb{R}^n$ is used for the vector with components $(\lambda^{(i)})^k$, $i = 1, \ldots, n$. However, for $X \in \mathbb{S}_n$ we define

$$|X| \overset{\text{def}}{=} U(X) D(|\lambda(X)|) U(X)^T \succeq 0,$$

and the notation $X^k$ is used for the standard matrix power. Since the power $k \geq 0$ does not change the ordering of nonnegative components, for any $X \succeq 0$ we have

$$\lambda^k(X) = \lambda(X^k). \tag{6.3.1}$$

Further, in $\mathbb{R}^n$, we use a standard notation for $\ell_p$-norms:

$$\| x \|_{(p)} = \left[ \sum_{i=1}^{n} |x^{(i)}|^p \right]^{1/p}, \quad x \in \mathbb{R}^n,$$

where $p \geq 1$, and $\|x\|_{(\infty)} = \max\limits_{1 \leq i \leq n} |x^{(i)}|$. The corresponding norms in $\mathbb{S}_n$ are introduced by

$$\|X\|_{(p)} = \|\lambda(X)\|_{(p)} = \|\lambda(|X|)\|_{(p)}, \quad X \in \mathbb{S}_n. \tag{6.3.2}$$

For $k \geq 1$, consider the following function:

$$\pi_k(X) = \langle X^k, I_n \rangle_F = \sum_{i=1}^{n} (\lambda^{(i)}(X))^k, \quad X \in \mathbb{S}_n.$$

Let us derive an upper bound for its second derivative. Note that this bound is nontrivial only for $k \geq 2$.

The derivatives of this function along a direction $H \in \mathbb{S}_n$ are defined as follows:

$$\begin{aligned} \langle \nabla \pi_k(X), H \rangle_F &= k \langle X^{k-1}, H \rangle_F, \\ \langle \nabla^2 \pi_k(X) H, H \rangle_F &= k \sum_{p=0}^{k-2} \langle X^p H X^{k-2-p}, H \rangle_F. \end{aligned} \tag{6.3.3}$$

We need the following result.

**Lemma 6.3.1** *For any $p, q \geq 0$, and $X$, $H$ from $\mathbb{S}_n$ we have*

$$\begin{aligned} \langle X^p H X^q + X^q H X^p, H \rangle_F &\leq 2 \langle |X|^{p+q}, H^2 \rangle_F \\ &\leq 2 \, \langle \lambda^{p+q}(|X|), \lambda^2(|H|) \rangle. \end{aligned} \tag{6.3.4}$$

*Proof* Indeed, let $\lambda = \lambda(X)$, $D = D(\lambda)$, $U = U(X)$ and $\hat{H} = U^T H U$. Then

$$\begin{aligned} \langle X^p H X^q + X^q H X^p, H \rangle_F &= \langle U D^p U^T H U D^q U^T + U D^q U^T H U D^p U^T, H \rangle_F \\ &= \langle D^p \hat{H} D^q + D^q \hat{H} D^p, \hat{H} \rangle_F \\ &= \sum_{i,j=1}^{n} (\hat{H}^{(i,j)})^2 \left( (\lambda^{(i)})^p (\lambda^{(j)})^q + (\lambda^{(i)})^q (\lambda^{(j)})^p \right) \\ &\leq \sum_{i,j=1}^{n} (\hat{H}^{(i,j)})^2 \left( |\lambda^{(i)}|^p |\lambda^{(j)}|^q + |\lambda^{(i)}|^q |\lambda^{(j)}|^p \right). \end{aligned}$$

Note that for arbitrary non-negative values $a$ and $b$ we always have

$$0 \leq (a^p - b^p)(a^q - b^q) = (a^{p+q} + b^{p+q}) - (a^p b^q + a^q b^p).$$

Thus, we can continue as follows:

$$\begin{aligned}\langle X^p H X^q + X^q H X^p, H\rangle_F &\le \sum_{i,j=1}^{n} (\hat{H}^{(i,j)})^2 \left(|\lambda^{(i)}|^{p+q} + |\lambda^{(j)}|^{p+q}\right)\\ &= 2\sum_{i,j=1}^{n} (\hat{H}^{(i,j)})^2 |\lambda^{(i)}|^{p+q} = 2\langle D(|\lambda|)^{p+q}\hat{H}, \hat{H}\rangle_F\\ &= 2\langle D^{p+q}(|\lambda|), \hat{H}^2\rangle_F = 2\langle |X|^{p+q}, H^2\rangle_F.\end{aligned}$$

Hence, we get the first inequality in (6.3.4). Further, by von Neumann's inequality

$$\langle |X|^{p+q}, H^2\rangle_F \;\le\; \langle \lambda(|X|^{p+q}), \lambda(H^2)\rangle \;\overset{(6.3.1)}{=}\; \langle \lambda^{p+q}(|X|), \lambda^2(|H|)\rangle,$$

and this proves the remaining part of (6.3.4). □

**Corollary 6.3.1** *For any $k \ge 2$, we have*

$$\langle \nabla^2 \pi_k(X)H, H\rangle_F \le k(k-1)\langle \lambda^{k-2}(|X|), \lambda^2(|H|)\rangle. \tag{6.3.5}$$

*Proof* For $k = 2$, the bound is trivial. For $k \ge 3$, in representation (6.3.3) we can unify the terms in the expression $\sum\limits_{p=0}^{k-2} \langle X^p H X^{k-2-p}, H\rangle_F$ in symmetric pairs

$$\langle X^p H X^{k-2-p} + X^{k-2-p} H X^p, H\rangle_F.$$

Applying inequality (6.3.4) to each pair, we get the estimate (6.3.5). □

Let $f(\cdot)$ be a function of a real variable, defined by a power series

$$f(\tau) = a_0 + \sum_{k=1}^{\infty} a_k \tau^k$$

with $a_k \ge 0$ for $k \ge 2$. We assume that its domain $\mathrm{dom}\, f = \{\tau \,:\, |\tau| < R\}$ is nonempty. For $X \in \mathbb{S}_n$, consider the following symmetric function of eigenvalues:

$$F(X) = \sum_{i=1}^{n} f(\lambda^{(i)}(X)).$$

Clearly, $\mathrm{dom}\, F = \{X \in \mathbb{S}_n \,:\, \lambda^{(1)}(X) < R,\; \lambda^{(n)}(X) > -R\}$.

**Theorem 6.3.1** *For any $X \in \mathrm{dom}\, F$ and $H \in \mathbb{S}_n$ we have*

$$\langle \nabla^2 F(X)H, H\rangle \le \sum_{i=1}^{n} \nabla^2 f(\lambda^{(i)}(|X|))(\lambda^{(i)}(|H|))^2.$$

*Proof* Indeed,

$$F(X) = n \cdot a_0 + \sum_{i=1}^{n} \sum_{k=1}^{\infty} a_k (\lambda^{(i)}(X))^k$$

$$= n \cdot a_0 + \sum_{k=1}^{\infty} a_k \sum_{i=1}^{n} (\lambda^{(i)}(X))^k \;=\; n \cdot a_0 + \sum_{k=1}^{\infty} a_k \pi_k(X).$$

Thus, in view of inequality (6.3.5),

$$\begin{aligned}
\langle \nabla^2 F(X) H, H \rangle_F &= \sum_{k=2}^{\infty} a_k \langle \nabla^2 \pi_k(X) H, H \rangle_F \\
&\leq \sum_{k=2}^{\infty} k(k-1) a_k \langle \lambda^{k-2}(|X|), \lambda^2(|H|) \rangle \\
&= \sum_{i=1}^{n} \sum_{k=2}^{\infty} k(k-1) a_k (\lambda^{(i)}(|X|))^{k-2} (\lambda^{(i)}(|H|))^2 \\
&= \sum_{i=1}^{n} \nabla^2 f(\lambda^{(i)}(|X|)) (\lambda^{(i)}(|H|))^2. \qquad \square
\end{aligned}$$

Let us consider now two important examples of symmetric functions of eigenvalues.

**1. Squared $\ell_p$-Matrix Norm**. For an integer $p \geq 1$, consider the following function:

$$F_p(X) = \tfrac{1}{2} \| \lambda(X) \|_{(2p)}^2 \;=\; \tfrac{1}{2} \langle X^{2p}, I_n \rangle_F^{1/p}, \quad X \in \mathbb{S}_n. \tag{6.3.6}$$

Thus, $F_p(X) = \frac{1}{2}(\pi_{2p}(X))^{1/p}$. Therefore, in view of (6.3.5), for any $X, H \in \mathbb{S}_n$ we have

$$\langle \nabla F_p(X), H \rangle_F = \tfrac{1}{2p} (\pi_{2p}(X))^{\frac{1}{p}-1} \langle \nabla \pi_{2p}(X), H \rangle_F,$$

$$\begin{aligned}
\langle \nabla^2 F_p(X) H, H \rangle_F &= \tfrac{1}{2p} \cdot \left( \tfrac{1}{p} - 1 \right) \cdot (\pi_{2p}(X))^{\frac{1}{p}-2} \langle \nabla \pi_{2p}(X), H \rangle_F^2 \\
&\quad + \tfrac{1}{2p} (\pi_{2p}(X))^{\frac{1}{p}-1} \langle \nabla^2 \pi_{2p}(X) H, H \rangle_F \qquad (6.3.7) \\
&\leq (2p-1)(\pi_{2p}(X))^{\frac{1}{p}-1} \langle \lambda^{2p-2}(|X|), \lambda^2(|H|) \rangle.
\end{aligned}$$

Let us apply Hölder's inequality $\langle x, y\rangle \leq \|x\|_{(\beta)}\|y\|_{(\gamma)}$ with $\beta = \frac{p}{p-1}$, $\gamma = \frac{\beta}{\beta-1} = p$, and

$$x^{(i)} = (\lambda^{(i)}(|X|))^{2p-2}, \quad y^{(i)} = (\lambda^{(i)}(|H|))^2, \quad i = 1, \ldots, n.$$

Then,

$$\begin{aligned} \langle x, y\rangle &\leq \left[\sum_{i=1}^{n}(\lambda^{(i)}(|X|))^{2p}\right]^{\frac{p-1}{p}} \cdot \left[\sum_{i=1}^{n}(\lambda^{(i)}(|H|))^{2p}\right]^{\frac{1}{p}} \\ &\overset{(6.3.2)}{=} \pi_{2p}(X)^{\frac{p-1}{p}} \cdot \|\lambda(H)\|^2_{(2p)}, \end{aligned}$$

and we can continue:

$$\langle \nabla^2 F_p(X)H, H\rangle_F \leq (2p-1)\|\lambda(H)\|^2_{(2p)} = (2p-1)\|H\|^2_{(2p)}. \tag{6.3.8}$$

**2. Entropy Smoothing of Maximal Eigenvalue**. Consider the function

$$E(X) = \ln \sum_{i=1}^{n} e^{\lambda^{(i)}(X)} \overset{\text{def}}{=} \ln F(X), \quad X \in \mathbb{S}_n. \tag{6.3.9}$$

Note that

$$\begin{aligned} \langle \nabla E(X), H\rangle_F &= \tfrac{1}{F(X)}\langle \nabla F(X), H\rangle_F, \\ \langle \nabla^2 E(X)H, H\rangle_F &= -\tfrac{1}{F^2(X)}\langle \nabla F(X), H\rangle_F^2 + \tfrac{1}{F(X)}\langle \nabla^2 F(X)H, H\rangle_F \\ &\leq \tfrac{1}{F(X)}\langle \nabla^2 F(X)H, H\rangle_F. \end{aligned}$$

Let us assume first that $X \succeq 0$. The function $F(X)$ is formed by the auxiliary function $f(\tau) = e^\tau$, which satisfies the assumptions of Theorem 6.3.1. Therefore,

$$\langle \nabla^2 E(X)H, H\rangle_F \leq \left[\sum_{i=1}^{n} e^{\lambda^{(i)}(X)}\right]^{-1} \sum_{i=1}^{n} e^{\lambda^{(i)}(X)}(\lambda^{(i)}(|H|))^2 \leq \|H\|^2_{(\infty)}. \tag{6.3.10}$$

It remains to note that $E(X + \tau I_n) = E(X) + \tau$. Hence, the Hessian $\nabla^2 E(X + \tau I_n)$ does not depend on $\tau$, and we conclude that the estimate (6.3.10) is valid for arbitrary $X \in \mathbb{S}_n$.

### 6.3.2 Minimizing the Maximal Eigenvalue of the Symmetric Matrix

Consider the following problem:

$$\text{Find } \phi^* = \min_{y \in Q} \{\phi(y) \overset{\text{def}}{=} \lambda_{\max}(C + A(y))\}, \tag{6.3.11}$$

where $Q$ is a closed convex set in $\mathbb{R}^m$ and $A(\cdot)$ is a linear operator from $\mathbb{R}^m$ to $\mathbb{S}_n$:

$$A(y) = \sum_{i=1}^{m} y^{(i)} A_i \in \mathbb{S}_n, \quad y \in \mathbb{R}^m.$$

Note that the objective function in (6.3.11) is nonsmooth. Therefore, this problem can be solved either by interior-point methods (see Chap. 5), or by general methods of nonsmooth convex optimization (see Chap. 3). However, due to the very special structure of the objective function, for problem (6.3.11) it is better to develop a special scheme.

We are going to solve problem (6.3.11) by a smoothing technique discussed in Sect. 6.1. This means that we replace the function $\lambda_{\max}(X)$ by its smooth approximation $f_\mu(X) = \mu E(\frac{1}{\mu} X)$, defined by (6.3.9) with tolerance parameter $\mu > 0$. Note that

$$\begin{aligned} f_\mu(X) &= \mu \ln \left[ \sum_{i=1}^{n} e^{\lambda^{(i)}(X)/\mu} \right] \geq \lambda_{\max}(X), \\ f_\mu(X) &\leq \lambda_{\max}(X) + \mu \ln n. \end{aligned} \tag{6.3.12}$$

At the same time,

$$\nabla f_\mu(X) = \left[ \sum_{i=1}^{n} e^{\lambda^{(i)}(X)/\mu} \right]^{-1} \cdot \sum_{i=1}^{n} e^{\lambda^{(i)}(X)/\mu} u_i(X) u_i(X)^T, \tag{6.3.13}$$

where $u_i(X)$, $i = 1, \ldots, n$, are corresponding unit eigenvectors of the symmetric matrix $X$. Thus, at each test point $X$, the gradient $\nabla f_\mu(X)$ takes into account all eigenvalues of the matrix $X$. However, since the factors $e^{\lambda^{(i)}(X)/\mu}$ decrease very rapidly, it actually depends only on few largest eigenvalues. Their selection is made automatically by expression (6.3.13). The ranking of importance of the eigenvalues is done in a logarithmic scale controlled by the tolerance parameter $\mu$.

Let us analyze now the efficiency of the smoothing technique as applied to problem (6.3.11). Our goal is to find an $\epsilon$-solution $\bar{x} \in Q$ to problem (6.3.11):

$$\phi(\bar{y}) - \phi^* \leq \epsilon. \tag{6.3.14}$$

For that, we will try to find a $\frac{1}{2}\epsilon$-solution to the smooth problem

$$\text{Find } \phi_\mu^* \;=\; \min_{y\in Q}\{\phi_\mu(y) \;\overset{\text{def}}{=}\; f_\mu(C+A(y))\}, \tag{6.3.15}$$

with

$$\mu = \mu(\epsilon) \;=\; \tfrac{\epsilon}{2\ln n}. \tag{6.3.16}$$

Clearly, if $\phi_\mu(\bar{y}) - \phi_\mu^* \le \frac{1}{2}\epsilon$, then in view of (6.3.12) we have

$$\phi(\bar{y}) - \phi^* \le \phi_\mu(\bar{y}) - \phi_\mu^* + \mu \ln n \le \epsilon.$$

Let us analyze now the complexity of finding a $\frac{1}{2}\epsilon$-solution to problem (6.3.15) by the optimal method (6.1.19).

Let us fix some norm $\|h\|$ for $h \in \mathbb{R}^m$. Consider a prox-function $d(\cdot)$ of the set $Q$ with prox-center $x_0 \in Q$. We assume this function to be strongly convex on $Q$ with convexity parameter one. Define

$$\|A\| = \max_{h\in\mathbb{R}^m}\{\|A(h)\|_{(\infty)} : \; \|h\| = 1\}.$$

Note that this norm is quite small. Indeed,

$$\|A(h)\|_{(\infty)} \;=\; \lambda^{(1)}(|A(h)|) \;\le\; \langle A(h), A(h)\rangle_F^{1/2}, \quad h \in \mathbb{R}^m.$$

Therefore, for example, $\|A\| \le \|A\|_G \overset{\text{def}}{=} \max\limits_{\|h\|=1} \langle A(h), A(h)\rangle_F^{1/2}$.

Let us estimate the second derivative of the function $\phi_\mu(\cdot)$. For any $y$ and $h$ from $\mathbb{R}^m$, in view of inequality (6.3.10) we have

$$\langle \nabla\phi_\mu(y), h\rangle = \langle \nabla f_\mu(C + A(y)), h\rangle \;=\; \langle \nabla E(\tfrac{1}{\mu}(C + A(y))), A(h)\rangle_F,$$

$$\langle \nabla^2\phi_\mu(y)h, h\rangle = \tfrac{1}{\mu}\langle \nabla^2 E(C + A(y))A(h), A(h)\rangle_F$$

$$\le \tfrac{1}{\mu}\|A(h)\|^2_{(\infty)} \;\le\; \tfrac{1}{\mu}\|A\|^2 \cdot \|h\|^2.$$

Thus, by Theorem 6.1.1 the function $\phi_\mu$ has Lipschitz continuous gradient with the constant

$$L = \tfrac{1}{\mu}\|A\|^2 = \tfrac{2\ln n}{\epsilon}\|A\|^2.$$

Now taking into account the estimate (6.1.21), we conclude that the method (6.1.19), as applied to problem (6.3.15), has the following rate of convergence:

$$\phi_\mu(y_k) - \phi_\mu^* \le \tfrac{8 \ln n \|A\|^2 d(y_\mu^*)}{\epsilon \cdot (k+1)(k+2)},$$

where $y_\mu^* \in Q$ is the solution to (6.3.15). Hence, it is able to generate a $\frac{1}{2}\epsilon$-solution to this problem (which is an $\epsilon$-solution to problem (6.3.11)) at most after

$$\tfrac{4\|A\|}{\epsilon} \sqrt{d(y_\mu^*) \ln n} \tag{6.3.17}$$

iterations.

## 6.4 Minimizing the Local Model of an Objective Function

(A linear optimization oracle; The method of conditional gradients; Conditional gradients with contraction; Computation of primal-dual solution; Strong convexity of the composite term; The second-order trust-region method with contraction.)

### *6.4.1 A Linear Optimization Oracle*

In this section we consider numerical methods for solving the following *composite* minimization problem:

$$\min_x \left\{ \bar{f}(x) \overset{\text{def}}{=} f(x) + \Psi(x) \right\}, \tag{6.4.1}$$

where $\Psi$ is a *simple* closed convex function with bounded domain $Q \subset \mathbb{E}$, and $f$ is a convex function, which is differentiable on $Q$. Denote by $x^*$ one of the optimal solutions of (6.4.1), and $D \overset{\text{def}}{=} \operatorname{diam}(Q)$. As usual, our assumption on the simplicity of the function $\Psi$ means that some auxiliary optimization problems related to $\Psi$ are easily solvable. The complexity of these problems will be always discussed for corresponding optimization schemes.

The most important examples of the function $\Psi$ are as follows.

- $\Psi$ is an indicator function of a closed convex set $Q$:

$$\Psi(x) = \operatorname{Ind}_Q(x) \overset{\text{def}}{=} \begin{cases} 0, & x \in Q, \\ +\infty, & \text{otherwise}. \end{cases} \tag{6.4.2}$$

- $\Psi$ is a self-concordant barrier for a closed convex set $Q$ (see Sect. 5.3).
- $\Psi$ is a nonsmooth convex function with simple structure. In this case, we need to include in $\Psi$ an indicator function for a bounded domain. For example, it

could be

$$\Psi(x) = \begin{cases} \|x\|_{(1)}, \text{ if } \|x\|_{(1)} \leq R, \\ +\infty, \text{ otherwise.} \end{cases}$$

We assume that the function $f$ is represented by a Black-Box oracle. If it is a *first-order oracle*, we assume its gradients satisfy the following *Hölder condition*:

$$\|\nabla f(x) - \nabla f(y)\|_* \leq G_\nu \|x - y\|^\nu, \quad x, y \in Q. \tag{6.4.3}$$

The constant $G_\nu$ is formally defined for any $\nu \in (0, 1]$. For some values of $\nu$ it can be $+\infty$. Note that for any $x$ and $y$ in $Q$ we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{G_\nu}{1+\nu} \|y - x\|^{1+\nu}. \tag{6.4.4}$$

If this is a *second-order oracle*, we assume that its Hessians satisfy the Hölder condition

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq H_\nu \|x - y\|^\nu, \quad x, y \in Q. \tag{6.4.5}$$

In this case, for any $x$ and $y$ in $Q$ we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \tfrac{H_\nu \|y-x\|^{2+\nu}}{(1+\nu)(2+\nu)}. \tag{6.4.6}$$

Our assumption on the simplicity of the function $\Psi$ means exactly the following.

**Assumption 6.4.1** *For any $s \in \mathbb{E}^*$, the auxiliary problem*

$$\min_{x \in Q} \{\langle s, x \rangle + \Psi(x)\} \tag{6.4.7}$$

*is easily solvable. Denote by $v_\Psi(s) \in Q$ one of its optimal solutions.*

Thus, for our methods we assume that we can use a *linear optimization oracle*, related to the set $Q$. Indeed, in the case (6.4.2), this assumption implies that we are able to solve the problem

$$\min_x \{\langle s, x \rangle : x \in Q\}.$$

For some sets (e.g. convex hulls of finite number of points), this oracle has lower complexity than the standard auxiliary problem consisting in minimizing a prox-function plus a linear term (see, for example, Sect. 6.1.3).

In view of Theorem 3.1.23 the point $v_\Psi(s)$ is characterized by the following variational principle:

$$\langle s, x - v_\Psi(s) \rangle + \Psi(x) \geq \Psi(v_\Psi(s)), \quad x \in Q. \tag{6.4.8}$$

By Definition 3.1.5, this means that $-s \in \partial \Psi(v_\Psi(s))$.

In the sequel, we often need to estimate the partial sums of different series. For that, it is convenient to use the following lemma, the proof of which we leave as an exercise for the reader.

**Lemma 6.4.1** *Let the function $\xi(\tau)$, $\tau \in \mathbb{R}$, be decreasing and convex. Then, for any two integers $a$ and $b$, such that $[a - \frac{1}{2}, b+1] \subset \operatorname{dom} \xi$, we have*

$$\int\limits_{a}^{b+1} \xi(\tau) d\tau \;\le\; \sum_{k=a}^{b} \xi(k) \le \int\limits_{a-1/2}^{b+1/2} \xi(\tau)\, d\tau. \tag{6.4.9}$$

For example, for any $t \ge 0$ and $p \ge -t$, we have

$$\begin{aligned} \sum_{k=t}^{2t+p} \tfrac{1}{k+p+1} \;&\overset{(5.4.38)}{\ge}\; \int\limits_{t}^{2t+p+1} \tfrac{1}{\tau+p+1} d\tau \;=\; \ln(\tau + p + 1)\Big|_{t}^{2t+p+1} \\ &=\; \ln \tfrac{2t+2p+2}{t+p+1} \;=\; \ln 2. \end{aligned} \tag{6.4.10}$$

On the other hand, if $t \ge 1$, then

$$\begin{aligned} \sum_{k=t}^{2t+1} \tfrac{1}{(k+2)^2} \;&\overset{(5.4.38)}{\le}\; \int\limits_{t-1/2}^{2t+3/2} \tfrac{1}{(\tau+2)^2} d\tau \;=\; -\tfrac{1}{\tau+2}\Big|_{t-1/2}^{2t+3/2} \;=\; \tfrac{1}{t+3/2} - \tfrac{1}{2t+7/2} \\ &=\; \tfrac{4t+8}{(2t+3)(4t+7)} \;\le\; \tfrac{12}{11(2t+3)}. \end{aligned} \tag{6.4.11}$$

### 6.4.2 *The Method of Conditional Gradients with Composite Objective*

In order to solve problem (6.4.1), we apply the following method.

**Conditional Gradients with Composite Objective**

**1.** Choose an arbitrary point $x_0 \in Q$.

**2. For $t \ge 0$ iterate:** (a) Compute $v_t = v_\Psi(\nabla f(x_t))$.

(b) Choose $\tau_t \in (0, 1]$ and set $x_{t+1} = (1 - \tau_t)x_t + \tau_t v_t$.

(6.4.12)

It is clear that this method can solve only problems where the function $f$ has continuous gradient.

*Example 6.4.1* Let $\Psi(x) = \mathrm{Ind}_Q(x)$ with $Q = \{x \in \mathbb{R}^2 : (x^{(1)})^2 + (x^{(2)})^2 \leq 1\}$. Define

$$f(x) = \max\{x^{(1)}, x^{(2)}\}.$$

Then clearly $x_* = \left(\frac{-1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right)^T$. Let us choose in (6.4.12) $x_0 \neq x_*$.

For the function $f$, we can apply an oracle which returns at any $x \in Q$ a subgradient $\nabla f(x) \in \{(1,0)^T, (0,1)^T\}$. Then, for any feasible $x$, the point $v_\Psi(\nabla f(x))$ is equal either to $y_1 = (-1,0)^T$, or to $y_2 = (0,-1)^T$. Therefore, all points of the sequence $\{x_t\}_{t \geq 0}$, generated by method (6.4.12), belong to the triangle $\mathrm{Conv}\{x_0, y_1, y_2\}$, which does not contain the optimal point $x_*$. □

In order to justify the rate of convergence of method (6.4.12) for functions with Hölder continuous gradients, we apply a variant of the estimating sequences technique (see Sects. 2.2.1 and 6.1.3). For that, it is convenient to introduce in (6.4.12) new control variables. Consider a sequence of nonnegative weights $\{a_t\}_{t \geq 0}$. Define

$$A_t = \sum_{k=0}^{t} a_k, \quad \tau_t = \frac{a_{t+1}}{A_{t+1}}, \quad t \geq 0. \tag{6.4.13}$$

From now on, we assume that the parameter $\tau_t$ in method (6.4.12) is chosen in accordance with the rule (6.4.13). Define

$$\begin{aligned} V_0 &= \max_x \{\langle \nabla f(x_0), x_0 - x\rangle + \Psi(x_0) - \Psi(x)\}, \\ B_{\nu,t} &= a_0 V_0 + \left(\sum_{k=1}^{t} \frac{a_k^{1+\nu}}{A_k^\nu}\right) G_\nu D^{1+\nu}, \quad t \geq 0. \end{aligned} \tag{6.4.14}$$

It is clear that

$$\begin{aligned} V_0 &\overset{(6.4.6)}{\leq} \max_x \left\{ f(x_0) - f(x) + \tfrac{G_\nu}{1+\nu}\|x - x_0\|^{1+\nu} + \Psi(x_0) - \Psi(x) \right\} \\ &\leq \bar{f}(x_0) - \bar{f}(x_*) + \tfrac{G_\nu D^{1+\nu}}{1+\nu} \overset{\text{def}}{=} \Delta(x_0) + \tfrac{G_\nu D^{1+\nu}}{1+\nu}. \end{aligned} \tag{6.4.15}$$

**Theorem 6.4.1** *Let the sequence $\{x_t\}_{t \geq 0}$ be generated by method (6.4.12). Then, for any $\nu \in (0,1]$ with $G_\nu < +\infty$, any step $t \geq 0$, and any $x \in Q$ we have*

$$A_t(f(x_t) + \Psi(x_t)) \leq \sum_{k=0}^{t} a_k[f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \Psi(x)] + B_{\nu,t}. \tag{6.4.16}$$

*Proof* Indeed, in view of definition (6.4.14), for $t = 0$ inequality (6.4.16) is satisfied. Assume that it is valid for some $t \geq 0$. Then

$$\begin{aligned}
&\sum_{k=0}^{t+1} a_k[f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \Psi(x)] + B_{\nu,t} \\
&\overset{(6.4.16)}{\geq} A_t(f(x_t) + \Psi(x_t)) + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle + \Psi(x)] \\
&\geq A_{t+1} f(x_{t+1}) + A_t \Psi(x_t) + \langle \nabla f(x_{t+1}), a_{t+1}(x - x_{t+1}) + A_t(x_t - x_{t+1})\rangle \\
&\quad + a_{t+1}\Psi(x) \\
&\overset{(6.4.12)_b}{=} A_{t+1} f(x_{t+1}) + A_t \Psi(x_t) + a_{t+1}\left[\Psi(x) + \langle \nabla f(x_{t+1}), x - v_t\rangle\right] \\
&\overset{(6.4.12)_b}{\geq} A_{t+1}\left(f(x_{t+1}) + \Psi(x_{t+1})\right) + a_{t+1}\left[\Psi(x) - \Psi(v_t) + \langle \nabla f(x_{t+1}), x - v_t\rangle\right].
\end{aligned}$$

It remains to note that

$$\begin{aligned}
\Psi(x) - \Psi(v_t) + \langle \nabla f(x_{t+1}), x - v_t\rangle &\overset{(6.4.8)}{\geq} \langle \nabla f(x_{t+1}) - \nabla f(x_t), x - v_t\rangle \\
&\overset{(6.4.3)}{\geq} -\tau_t^{\nu} G_\nu D^{1+\nu}.
\end{aligned}$$

Thus, to ensure that (6.4.16) is valid for the next iteration, it is enough to choose

$$B_{\nu,t+1} = B_{\nu,t} + \frac{a_{t+1}^{1+\nu}}{A_{t+1}^{\nu}} G_\nu D^{1+\nu}. \qquad \square$$

**Corollary 6.4.1** *For any $t \geq 0$ with $A_t > 0$, and any $\nu \in (0, 1]$, we have*

$$\bar{f}(x_t) - \bar{f}(x_*) \leq \frac{1}{A_t} B_{\nu,t}. \tag{6.4.17}$$

Let us discuss now the possible variants for choosing the weights $\{a_t\}_{t\geq 0}$.

1. *Constant weights*. Let us choose $a_t \equiv 1$, $t \geq 0$. Then $A_t = t + 1$, and for $\nu \in (0, 1)$ we have

$$\begin{aligned}
B_{\nu,t} &= V_0 + \left(\sum_{k=1}^{t} \frac{1}{(1+k)^\nu}\right) G_\nu D^{1+\nu} \\
&\overset{(6.4.9)}{\leq} V_0 + G_\nu D^{1+\nu} \frac{1}{1-\nu}(1+\tau)^{1-\nu}\Big|_{1/2}^{t+1/2} \\
&\overset{(6.4.15)}{\leq} \Delta(x_0) + G_\nu D^{1+\nu}\left[\frac{1}{1+\nu} + \left(\frac{3}{2}\right)^{1-\nu} \frac{1}{1-\nu}\left(\left(1 + \frac{2}{3}t\right)^{1-\nu} - 1\right)\right].
\end{aligned}$$

Thus, for $\nu \in (0,1)$, we have $\frac{1}{A_t} B_{\nu,t} \le O(t^{-\nu})$. For the most important case $\nu = 1$, we have $\lim\limits_{\nu \to 1} \frac{1}{1-\nu} \left( \left(1 + \frac{2}{3}t\right)^{1-\nu} - 1 \right) = \ln(1 + \frac{2}{3}t)$. Therefore,

$$\bar{f}(x_t) - \bar{f}(x_*) \le \frac{1}{t+1} \left( \varDelta(x_0) + G_1 D^2 \left[ \frac{1}{2} + \ln(1 + \frac{2}{3}t) \right] \right). \tag{6.4.18}$$

In this situation, in method (6.4.12) we take $\tau_t \overset{(6.4.13)}{=} \frac{1}{t+1}$.

2. *Linear weights*. Let us choose $a_t \equiv t$, $t \ge 0$. Then $A_t = \frac{t(t+1)}{2}$, and for $\nu \in (0,1)$ with $t \ge 1$ we have

$$\begin{aligned} B_{\nu,t} &= \left( \sum_{k=1}^{t} \frac{2^\nu k^{1+\nu}}{k^\nu (1+k)^\nu} \right) G_\nu D^{1+\nu} \le \left( \sum_{k=1}^{t} 2^\nu k^{1-\nu} \right) G_\nu D^{1+\nu} \\ &\overset{(6.4.9)}{\le} G_\nu D^{1+\nu} \frac{2^\nu}{2-\nu} \tau^{2-\nu} \Big|_{1/2}^{t+1/2} = \frac{2^\nu}{2-\nu} \left[ \left(t + \frac{1}{2}\right)^{2-\nu} - \left(\frac{1}{2}\right)^{2-\nu} \right] G_\nu D^{1+\nu}. \end{aligned}$$

Thus, for $\nu \in (0,1)$, we again have $\frac{1}{A_t} B_{\nu,t} \le O(t^{-\nu})$. For the case $\nu = 1$, we get the following bound:

$$\bar{f}(x_t) - \bar{f}(x_*) \le \frac{4}{t+1} G_1 D^2, \quad t \ge 1. \tag{6.4.19}$$

As we can see, this rate of convergence is better than (6.4.18). In this case, in method (6.4.12) we take $\tau_t \overset{(6.4.13)}{=} \frac{2}{t+2}$, which is a standard recommendation for this scheme.

3. *Aggressive weights*. Let us choose, for example, $a_t \equiv t^2$, $t \ge 0$. Then $A_t = \frac{t(t+1)(2t+1)}{6}$. Note that for $k \ge 0$ we have $\frac{k^{2+\nu}}{(k+1)^\nu (2k+1)^\nu} \le \frac{k^{2-\nu}}{2^\nu}$. Therefore, for $\nu \in (0,1)$ with $t \ge 1$ we obtain

$$\begin{aligned} B_{\nu,t} &= \left( \sum_{k=1}^{t} \frac{6^\nu k^{2(1+\nu)}}{k^\nu (1+k)^\nu (2k+1)^\nu} \right) G_\nu D^{1+\nu} \le \left( \sum_{k=1}^{t} 3^\nu k^{2-\nu} \right) G_\nu D^{1+\nu} \\ &\overset{(6.4.9)}{\le} G_\nu D^{1+\nu} \frac{3^\nu}{3-\nu} \tau^{3-\nu} \Big|_{1/2}^{t+1/2} = \frac{3^\nu}{3-\nu} \left[ \left(t + \frac{1}{2}\right)^{3-\nu} - \left(\frac{1}{2}\right)^{3-\nu} \right] G_\nu D^{1+\nu}. \end{aligned}$$

For $\nu \in (0,1)$, we get again $\frac{1}{A_t} B_{\nu,t} \le O(t^{-\nu})$. For $\nu = 1$, we obtain

$$\bar{f}(x_t) - \bar{f}(x_*) \le \frac{9}{2t+1} G_1 D^2, \quad t \ge 1, \tag{6.4.20}$$

which is slightly worse than (6.4.19). The rule for choosing the coefficients $\tau_t$ in this situation is $\tau_t \overset{(6.4.13)}{=} \frac{6(t+1)}{(t+2)(2t+3)}$. It can be easily checked that a further increase of the rate of growth of coefficients $a_t$ makes the rate of convergence of method (6.4.12) even worse.

Note that the above rules for choosing the coefficients $\{\tau_t\}_{t\geq 0}$ in method (6.4.12) do not depend on the smoothness parameter $\nu \in (0,1]$. In this sense, method (6.4.12) is a *universal method* for solving the problem (6.4.1). Moreover, this method is *affine invariant*. Its behavior does not depend on the choice of norm in $\mathbb{E}$. Hence, its rate of convergence can be established with respect to the best norm describing the geometry of the feasible set.

### 6.4.3 Conditional Gradients with Contraction

In this section, we will use some special dual functions. Let $Q \subset E$ be a bounded closed convex set. For a closed convex function $F(\cdot)$ with $\operatorname{dom} F \supseteq \operatorname{int} Q$, we define its *restricted dual function*, (with respect to a central point $\bar{x} \in Q$), as follows:

$$F^*_{\bar{x},Q}(s) = \max_{x\in Q}\{\langle s, \bar{x} - x\rangle + F(\bar{x}) - F(x)\}, \quad s \in \mathbb{E}^*. \tag{6.4.21}$$

Clearly, this function is well defined for all $s \in \mathbb{E}^*$. Moreover, it is convex and nonnegative on $\mathbb{E}^*$.

We need to introduce in construction (6.4.21) an additional scaling parameter $\tau \in [0,1]$, which controls the size of the feasible set. For $s \in \mathbb{E}^*$, we call the function

$$F^*_{\tau,\bar{x},Q}(s) = \max_{x\in Q}\{\langle s, \bar{x} - y\rangle + F(\bar{x}) - F(y) : \; y = (1-\tau)\bar{x} + \tau x\} \tag{6.4.22}$$

the *scaled restricted dual* of the function $F$.

**Lemma 6.4.2** *For any $s \in \mathbb{E}^*$ and $\tau \in [0,1]$, we have*

$$F^*_{\bar{x},Q}(s) \geq F^*_{\tau,\bar{x},Q}(s) \;\geq\; \tau\, F^*_{\bar{x},Q}(s). \tag{6.4.23}$$

*Proof* Since for any $x \in Q$, the point $y = (1-\tau)\bar{x} + \tau x$ belongs to $Q$, the first inequality is trivial. On the other hand,

$$\begin{aligned} F^*_{\tau,\bar{x},Q}(s) &= \max_{x\in Q}\{\langle s, \tau(\bar{x} - x)\rangle + F(\bar{x}) - F(y) : \; y = (1-\tau)\bar{x} + \tau x \;\} \\ &\geq \max_{x\in Q}\{\langle s, \tau(\bar{x} - x)\rangle + F(\bar{x}) - (1-\tau)F(\bar{x}) - \tau F(x) \;\} \\ &= \tau F^*_{\bar{x},Q}(s). \qquad \square \end{aligned}$$

Let us consider a variant of method (6.4.12), which takes into account the composite form of the objective function in problem (6.4.1). For $\Psi(x) \equiv \operatorname{Ind}_Q(x)$, these two methods coincide. Otherwise, they generate different minimization sequences.

**Conditional Gradient Method with Contraction**

**1.** Choose an arbitrary point $x_0 \in Q$.

**2. For** $t \geq 0$ **iterate:** Choose a coefficient $\tau_t \in (0, 1]$ and compute

$$x_{t+1} = \arg\min_{x \in Q} \{ \langle \nabla f(x_t), y \rangle + \Psi(y) : \; y = (1 - \tau_t) x_t + \tau_t x \}.$$

(6.4.24)

This method can be seen as a *Trust-Region Scheme* with a linear model of the objective function. The trust region in method (6.4.24) is formed by a contraction of the initial feasible set. In Sect. 6.4.6, we will consider a more traditional trust-region method with quadratic model of the objective.

In view of Theorem 3.1.23 the point $x_{t+1}$ in method (6.4.24) is characterized by the following variational principle:

$$\begin{aligned} x_{t+1} &= (1 - \tau_t) x_t + \tau_t v_t, \; v_t \in Q, \\ &\Psi((1 - \tau_t) x_t + \tau_t x) + \tau_t \langle \nabla f(x_t), x - x_t \rangle \\ &\geq \Psi(x_{t+1}) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle, \quad x \in Q. \end{aligned} \tag{6.4.25}$$

Let us choose somehow the sequence of nonnegative weights $\{a_t\}_{t \geq 0}$, and define in (6.4.24) the coefficients $\tau_t$ in accordance to (6.4.13). Define now the estimating functional sequence $\{\phi_t(x)\}_{t \geq 0}$ as follows:

$$\begin{aligned} \phi_0(x) &= a_0 \bar{f}(x), \\ \phi_{t+1}(x) &= \phi_t(x) + a_{t+1} [ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \Psi(x) ], \quad t \geq 0. \end{aligned} \tag{6.4.26}$$

Clearly, for all $t \geq 0$ we have

$$\phi_t(x) \leq A_t \bar{f}(x), \quad x \in Q. \tag{6.4.27}$$

Define

$$C_{\nu,t} = a_0 \Delta(x_0) + \tfrac{1}{1+\nu}\left(\sum_{k=1}^{t} \frac{a_k^{1+\nu}}{A_k^{\nu}}\right) G_\nu D^{1+\nu}, \quad t \ge 0. \tag{6.4.28}$$

Let us introduce

$$\delta(x) \overset{\text{def}}{=} \max_{y\in Q}\{\langle \nabla f(x), x-y\rangle + \Psi(x) - \Psi(y)\} \overset{(6.4.21)}{\equiv} \Psi^*_{x,Q}(\nabla f(x)). \tag{6.4.29}$$

For problem (6.4.1), this value measures the level of satisfaction of the first-order optimality conditions at a point $x \in Q$. For any $x \in Q$, we have

$$\delta(x) \ge \bar f(x) - \bar f(x_*) \ge 0. \tag{6.4.30}$$

We call $\delta(x)$ the *total variation* of the linear model of the composite objective function in problem (6.4.1) over the feasible set. It justifies the first-order optimality conditions in our problem. Note that this value can be computed by a procedure for solving the auxiliary problem (6.4.7).

**Theorem 6.4.2** *Let the sequence $\{x_t\}_{t\ge 0}$ be generated by method (6.4.24). Then, for any $\nu \in (0,1]$ and any step $t \ge 0$, we have*

$$A_t \bar f(x_t) \le \phi_t(x) + C_{\nu,t}, \quad x \in Q. \tag{6.4.31}$$

*Moreover, for any $t \ge 0$ we have*

$$\bar f(x_t) - \bar f(x_{t+1}) \ge \tau_t \delta(x_t) - \tfrac{G_\nu D^{1+\nu}}{1+\nu} \tau_t^{1+\nu}. \tag{6.4.32}$$

*Proof* Let us prove inequality (6.4.31). For $t = 0$, we have $C_{\nu,0} = a_0[\bar f(x_0) - \bar f(x_*)]$. Thus, in this case (6.4.31) follows from (6.4.27).

Assume now that (6.4.31) is valid for some $t \ge 0$. In view of definition (6.4.13), optimality condition (6.4.25) can written in the following form:

$$\begin{aligned} a_{t+1}\langle \nabla f(x_t), x - x_t\rangle \ge A_{t+1}\,[&\Psi(x_{t+1}) - \Psi((1-\tau_t)x_t + \tau_t x) \\ &+ \langle \nabla f(x_t), x_{t+1} - x_t\rangle] \end{aligned}$$

for all $x \in Q$. Therefore,

$$\begin{aligned}
\phi_{t+1}(x) + C_{\nu,t} \quad &= \quad \phi_t(x) + C_{\nu,t} \\
&\quad +a_{t+1}[f(x_t) + \langle \nabla f(x_t), x - x_t\rangle + \Psi(x)] \\
&\overset{(6.4.25),(6.4.31)}{\geq} \quad A_t[f(x_t) + \Psi(x_t)] + a_{t+1}[f(x_t) + \Psi(x)] \\
&\quad +A_{t+1}[\Psi(x_{t+1}) - \Psi((1-\tau_t)x_t + \tau_t x) \\
&\quad +\langle \nabla f(x_t), x_{t+1} - x_t\rangle] \\
&\geq \quad A_{t+1}\,[f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \Psi(x_{t+1})] \\
&\overset{(6.4.4)}{\geq} \quad A_{t+1}\left[\bar{f}(x_{t+1}) - \tfrac{1}{1+\nu} G_\nu \|x_{t+1} - x_t\|^{1+\nu}\right].
\end{aligned}$$

It remains to note that $\|x_{t+1} - x_t\| = \tau_t \|x_t - v_t\| \overset{(6.4.13)}{\leq} \frac{a_{t+1}}{A_{t+1}} D$. Thus, we can take

$$C_{\nu,t+1} = C_{\nu,t} + \tfrac{1}{1+\nu} \tfrac{a_{t+1}^{1+\nu}}{A_{t+1}^{\nu}} G_\nu D^{1+\nu}.$$

In order to prove inequality (6.4.32), let us introduce the values

$$\begin{aligned}
\delta_\tau(x) \quad &\overset{\text{def}}{=} \quad \max_{u\in Q}\{\langle \nabla f(x), x - y\rangle + \Psi(x) - \Psi(y) : \ y = (1-\tau)x + \tau u\} \\
&\overset{(6.4.22)}{=} \Psi^*_{\tau,x,Q}(\nabla f(x)), \quad \tau \in [0,1].
\end{aligned}$$

Clearly,

$$\begin{aligned}
-\delta_{\tau_t}(x_t) \quad &= \quad \min_{x\in Q}\{\langle \nabla f(x_t), y - x_t\rangle + \Psi(y) - \Psi(x_t) : y = (1-\tau_t)x_t + \tau_t x\} \\
&= \quad \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \Psi(x_{t+1}) - \Psi(x_t) \\
&\overset{(6.4.4)}{\geq} \bar{f}(x_{t+1}) - \bar{f}(x_t) - \tfrac{G_\nu}{1+\nu}\|x_{t+1} - x_t\|^{1+\nu}.
\end{aligned}$$

Since $\|x_{t+1} - x_t\| \leq \tau_t D$, we conclude that

$$\bar{f}(x_t) - \bar{f}(x_{t+1}) \geq \delta_{\tau_t}(x_t) - \tfrac{G_\nu D^{1+\nu}}{1+\nu}\tau_t^{1+\nu} \overset{(6.4.23)}{\geq} \tau_t \delta(x_t) - \tfrac{G_\nu D^{1+\nu}}{1+\nu}\tau_t^{1+\nu}. \qquad \square$$

In view of (6.4.27), inequality (6.4.31) results in the following rate of convergence:

$$\bar{f}(x_t) - \bar{f}(x_*) \le \tfrac{1}{A_t} C_{\nu,t}, \quad t \ge 0. \tag{6.4.33}$$

For the linearly growing weights $a_t = t$, $A_t = \frac{t(t+1)}{2}$, $t \ge 0$, we have already seen that

$$C_{\nu,t} = \tfrac{1}{1+\nu} B_{\nu,t} \;\le\; \tfrac{2^\nu}{(1+\nu)(2-\nu)} \left[ \left(t + \tfrac{1}{2}\right)^{2-\nu} - \left(\tfrac{1}{2}\right)^{2-\nu} \right] G_\nu D^{1+\nu}.$$

In the case $\nu = 1$, this results in the following rate of convergence:

$$\bar{f}(x_t) - \bar{f}(x_*) \le \tfrac{2}{t+1} G_1 D^2, \quad t \ge 1. \tag{6.4.34}$$

Let us justify for this case the rate of convergence of the sequence $\{\delta(x_t)\}_{t \ge 1}$. We have $\tau_t \overset{(6.4.13)}{=} \frac{a_{t+1}}{A_{t+1}} = \frac{2}{t+2}$. On the other hand, for any $T \ge t$,

$$\begin{aligned} \tfrac{2G_1 D^2}{t+1} \;&\overset{(6.4.34)}{\ge}\; \bar{f}(x_t) - \bar{f}(x_*) \\ &\overset{(6.4.32)}{\ge} \sum_{k=t}^{T} \left[ \tau_k \delta(x_k) - \tfrac{1}{2} G_1 D^2 \tau_k^2 \right] + \bar{f}(x_{T+1}) - \bar{f}(x_*). \end{aligned} \tag{6.4.35}$$

Let $\delta_T^* = \min\limits_{0 \le t \le T} \delta(x_t)$. Then, choosing $T = 2t + 1$, we get

$$\begin{aligned} 2\ln 2 \cdot \delta_T^* \;&\overset{(6.4.10)}{\le} \left( \sum_{k=t}^{T} \tfrac{2}{k+2} \right) \delta_T^* \;\overset{(6.4.35)}{\le}\; 2G_1 D^2 \left[ \tfrac{1}{t+1} + \sum_{k=t}^{T} \tfrac{1}{(k+2)^2} \right] \\ &\overset{(6.4.11)}{\le} 2G_1 D^2 \left[ \tfrac{1}{t+1} + \tfrac{12}{11(2t+3)} \right] \;=\; 2G_1 D^2 \left[ \tfrac{2}{T+1} + \tfrac{12}{11(T+2)} \right] \\ &\le\; \tfrac{68}{11} \cdot \tfrac{G_1 D^2}{T+1}. \end{aligned}$$

Thus, in the case $\nu = 1$, for odd $T$, we get the following bound:

$$\delta_T^* \le \tfrac{34}{11 \ln 2} \cdot \tfrac{G_1 D^2}{T+1}. \tag{6.4.36}$$

### 6.4.4 Computing the Primal-Dual Solution

Note that both methods (6.4.12) and (6.4.24) admit computable accuracy certificates. For the first method, define

$$\ell_t = \frac{1}{A_t} \min_x \left\{ \sum_{k=0}^{t} a_k [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \Psi(x)] : \ x \in Q \right\}.$$

This value can be computed by the standard operation (6.4.7). Clearly,

$$\bar{f}(x_t) - \bar{f}(x_*) \leq \bar{f}(x_t) - \ell_t \overset{(6.4.16)}{\leq} \frac{1}{A_t} B_{\nu,t}. \tag{6.4.37}$$

For the second method, let us choose $a_0 = 0$. Then the estimating functions are linear:

$$\phi_t(x) = \sum_{k=1}^{t} a_k [f(x_{k-1}) + \langle \nabla f(x_{k-1}), x - x_{k-1} \rangle + \Psi(x)].$$

Therefore, defining $\hat{\ell}_t = \frac{1}{A_t} \min\limits_x \{\phi_t(x) : \ x \in Q\}$, we also have

$$\bar{f}(x_t) - \bar{f}(x_*) \leq \bar{f}(x_t) - \hat{\ell}_t \overset{(6.4.16)}{\leq} \frac{1}{A_t} C_{\nu,t}, \quad t \geq 1. \tag{6.4.38}$$

Accuracy certificates (6.4.37) and (6.4.38) justify that both methods (6.4.12) and (6.4.24) are able to recover some information on the optimal dual solution. However, in order to implement this ability, we need to open the Black Box and introduce an *explicit model* of the function $f(\cdot)$.

Let us assume that the function $f$ is representable in the following form:

$$f(x) = \max_u \{\langle Ax, u \rangle - g(u) : \ u \in Q_d\}, \tag{6.4.39}$$

where $A : \mathbb{E} \to \mathbb{E}_1^*$, $Q_d$ is a closed convex set in a finite-dimensional linear space $\mathbb{E}_2$, and the function $g(\cdot)$ is *p-uniformly convex* on $Q_d$:

$$\langle \nabla g(u_1) - \nabla g(u_2), u_1 - u_2 \rangle \geq \sigma_g \|u_1 - u_2\|^p, \quad u_1, u_2 \in Q_d, \tag{6.4.40}$$

where the *convexity degree* $p \geq 2$. Denote by $u(x) \in Q_d$ the unique optimal solution to optimization problem in (6.4.39).

**Lemma 6.4.3** *The function $f$ has Hölder continuous gradient $\nabla f(x) = A^* u(x)$ with parameter $\nu = \frac{1}{p-1}$ and constant $G_\nu = \left(\frac{1}{\sigma_g}\right)^\nu \|A\|^{1+\nu}$.*

*Proof* Let $u_1 = u(x_1)$, $u_2 = u(x_2)$, $g_1' = \nabla g(u_1)$, and $g_2' = \nabla g(u_2)$. Then, in view of the optimality condition (2.2.39), we have

$$\langle Ax_1 - g_1', u_2 - u_1 \rangle \leq 0, \quad \langle Ax_2 - g_2', u_1 - u_2 \rangle \;\leq\; 0.$$

Adding these two inequalities, we get

$$\langle A(x_1 - x_2), u_1 - u_2 \rangle \geq \langle g_1' - g_2', u_1 - u_2 \rangle \overset{(6.4.40)}{\geq} \sigma_g \| u_1 - u_2 \|^p.$$

Thus,

$$\begin{aligned} \| \nabla f(x_1) - \nabla f(x_2) \|^* = \| A^*(u_1 - u_2) \|^* \;&\leq\; \| A \| \cdot \| u_1 - u_2 \| \\ &\leq \| A \| \cdot \left( \tfrac{1}{\sigma_g} \| A(x_1 - x_2) \| \right)^{\frac{1}{p-1}} \\ &\leq \| A \|^{\frac{p}{p-1}} \left( \tfrac{1}{\sigma_g} \| x_1 - x_2 \| \right)^{\frac{1}{p-1}}. \qquad \Box \end{aligned}$$

Let us write down an *adjoint problem* to (6.4.1).

$$\begin{aligned} \min_x \{ \bar{f}(x) : \; x \in Q \} \overset{(6.4.39)}{=} & \min_x \left\{ \Psi(x) + \max_u \{ \langle Ax, u \rangle - g(u) : \; u \in Q_d \} \right\} \\ \geq & \max_{u \in Q_d} \left\{ -g(u) + \min_x \{ \langle A^* u, x \rangle + \Psi(x) \} \right\}. \end{aligned}$$

Thus, defining $\Phi(u) = \min\limits_x \{ \langle A^* u, x \rangle + \Psi(x) \}$, we get the following adjoint problem:

$$\max_{u \in Q_d} \left\{ \bar{g}(u) \overset{\text{def}}{=} -g(u) + \Phi(u) \right\}. \tag{6.4.41}$$

In this problem, the objective function is nonsmooth and uniformly strongly concave of degree $p$. Clearly, we have

$$\bar{f}(x) - \bar{g}(u) \geq 0, \quad x \in Q, \; u \in Q_d. \tag{6.4.42}$$

Let us show that both methods (6.4.12) and (6.4.24) are able to approximate the optimal solution to the problem (6.4.41).

Note that for any $\bar{x} \in Q$ we have

$$\begin{aligned} f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle \overset{(6.4.39)}{=} & \langle A\bar{x}, u(\bar{x}) \rangle - g(u(\bar{x})) + \langle A^* u(\bar{x}), x - \bar{x} \rangle \\ = \;\; & \langle Ax, u(\bar{x}) \rangle - g(u(\bar{x})). \end{aligned}$$

Therefore, defining for the first method (6.4.12) $u_t = \frac{1}{A_t}\sum\limits_{k=0}^{t} a_k u(x_k)$, we obtain

$$\begin{aligned}\ell_t &= \min_{x\in Q}\left\{\Psi(x) + \frac{1}{A_t}\sum_{k=0}^{t} a_k[\langle Ax, u(x_k)\rangle - g(u(x_k))]\right\}\\ &= \Phi(u_t) - \frac{1}{A_t}\sum_{k=0}^{t} a_k g(u(x_k)) \;\le\; \bar{g}(u_t).\end{aligned}$$

Thus, we get

$$0 \overset{(6.4.42)}{\le} \bar{f}(x_t) - \bar{g}(u_t) \;\le\; \bar{f}(x_t) - \ell_t \overset{(6.4.37)}{\le} \frac{1}{A_t} B_{\nu,t}, \quad t \ge 0. \tag{6.4.43}$$

For the second method (6.4.24), we choose $a_0 = 0$ and take $u_t = \frac{1}{A_t}\sum\limits_{k=1}^{t} a_k u(x_{k-1})$. In this case, by a similar reasoning, we get

$$0 \overset{(6.4.42)}{\le} \bar{f}(x_t) - \bar{g}(u_t) \;\le\; \bar{f}(x_t) - \hat{\ell}_t \overset{(6.4.38)}{\le} \frac{1}{A_t} C_{\nu,t}, \quad t \ge 1. \tag{6.4.44}$$

## 6.4.5 Strong Convexity of the Composite Term

In this section, we assume that the function $\Psi$ in problem (6.4.1) is *strongly convex* (see Sect. 3.2.6). In view of (3.2.37), this means that there exists a positive constant $\sigma_\Psi$ such that

$$\Psi(\tau x + (1-\tau)y) \le \tau\Psi(x) + (1-\tau)\Psi(y) - \frac{1}{2}\sigma_\Psi \tau(1-\tau)\|x-y\|^2 \tag{6.4.45}$$

for all $x, y \in Q$ and $\tau \in [0,1]$. Let us show that in this case CG-methods converge much faster. We demonstrate this for method (6.4.12).

In view of the strong convexity of $\Psi$, the variational principle (6.4.8) characterizing the point $v_t$ in method (6.4.12) can be strengthened:

$$\Psi(x) + \langle \nabla f(x_t), x - v_t\rangle \ge \Psi(v_t) + \frac{1}{2}\sigma_\Psi\|x - v_t\|^2, \quad x \in Q. \tag{6.4.46}$$

Let $V_0$ be defined as in (6.4.14). Define

$$\hat{B}_{\nu,t} = a_0 V_0 + \left(\sum_{k=1}^{t} \frac{a_k^{1+2\nu}}{A_k^{2\nu}}\right)\frac{G_\nu^2 D^{2\nu}}{2\sigma_\Psi}, \quad t \ge 0. \tag{6.4.47}$$

**Theorem 6.4.3** *Let the sequence $\{x_t\}_{t\geq 0}$ be generated by method (6.4.12), and assume the function $\Psi$ is strongly convex. Then, for any $\nu \in (0,1]$, any step $t \geq 0$, and any $x \in Q$ we have*

$$A_t(f(x_t) + \Psi(x_t)) \leq \sum_{k=0}^{t} a_k[f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \Psi(x)] + \hat{B}_{\nu,t}. \tag{6.4.48}$$

*Proof* The beginning of the proof of this statement is very similar to that of Theorem 6.4.1. Assuming that (6.4.48) is valid for some $t \geq 0$, we get the following inequality:

$$\begin{aligned}&\sum_{k=0}^{t+1} a_k[f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \Psi(x)] + B_{\nu,t}\\ &\geq A_{t+1}\left(f(x_{t+1}) + \Psi(x_{t+1})\right) + a_{t+1}\left[\Psi(x) - \Psi(v_t) + \langle \nabla f(x_{t+1}), x - v_t\rangle\right].\end{aligned}$$

Further,

$$\begin{aligned}&\Psi(x) - \Psi(v_t) + \langle \nabla f(x_{t+1}), x - v_t\rangle\\ &\overset{(6.4.46)}{\geq} \langle \nabla f(x_{t+1}) - \nabla f(x_t), x - v_t\rangle + \tfrac{1}{2}\sigma_\Psi \|x - v_t\|^2\\ &\overset{(4.2.3)}{\geq} -\tfrac{1}{2\sigma_\Psi}\|\nabla f(x_{t+1}) - \nabla f(x_t)\|_*^2\\ &\overset{(6.4.3)}{\geq} -\tfrac{1}{2\sigma_\Psi}\left(\tfrac{a_{t+1}^\nu}{A_{t+1}^\nu} G_\nu D^\nu\right)^2.\end{aligned}$$

Thus, to ensure that (6.4.48) is valid for the next iteration, it is enough to choose

$$\hat{B}_{\nu,t+1} = \hat{B}_{\nu,t} + \tfrac{1}{2\sigma_\Psi}\tfrac{a_{t+1}^{1+2\nu}}{A_{t+1}^{2\nu}} G_\nu^2 D^{2\nu}. \qquad \square$$

It can be easily checked that in our situation, the linear weights strategy $a_t \equiv t$ is not the best one. Let us choose $a_t = t^2$, $t \geq 0$. Then $A_t = \frac{t(t+1)(2t+1)}{6}$, and we get

$$\begin{aligned}\hat{B}_{\nu,t} &= \left(\sum_{k=1}^{t} \tfrac{6^{2\nu}k^{2(1+2\nu)}}{k^{2\nu}(k+1)^{2\nu}(2k+1)^{2\nu}}\right)\tfrac{G_\nu^2 D^{2\nu}}{2\sigma_\Psi} \leq \left(3^{2\nu}\sum_{k=1}^{t} k^{2(1-\nu)}\right)\tfrac{G_\nu^2 D^{2\nu}}{2\sigma_\Psi}\\ &\overset{(6.4.9)}{\leq} \tfrac{G_\nu^2 D^{2\nu}}{2\sigma_\Psi}\cdot\tfrac{3^{2\nu}}{3-2\nu}\tau^{3-2\nu}\Big|_{1/2}^{t+1/2} = \tfrac{3^{2\nu}}{3-2\nu}\left[(t+\tfrac{1}{2})^{3-2\nu} - \left(\tfrac{1}{2}\right)^{3-2\nu}\right]\tfrac{G_\nu^2 D^{2\nu}}{2\sigma_\Psi}.\end{aligned}$$

Thus, for $\nu \in (0, 1)$, we get $\frac{1}{A_t}\hat{B}_{\nu,t} \leq O(t^{-2\nu})$. For $\nu = 1$, we obtain

$$\bar{f}(x_t) - \bar{f}(x_*) \leq \tfrac{54}{(t+1)(2t+1)} \cdot \tfrac{G_1^2 D^2}{2\sigma_\Psi}, \tag{6.4.49}$$

which is much better than (6.4.19). This gives us an example of acceleration of the Conditional Gradient Method by a strong convexity assumption.

### 6.4.6 *Minimizing the Second-Order Model*

Let us assume now that in problem (6.4.1) the function $f$ is twice continuously differentiable. Then we can apply to this problem the following method.

**Composite Trust-Region Method with Contraction**

**1.** Choose an arbitrary point $x_0 \in Q$.

**2. For** $t \geq 0$ **iterate:** Define the coefficient $\tau_t \in (0, 1]$ and choose

$$\begin{aligned} x_{t+1} \in \operatorname{Arg}\min_y \Big\{ & \langle \nabla f(x_t), y - x_t\rangle + \tfrac{1}{2}\langle \nabla^2 f(x_t)(y - x_t), y - x_t\rangle \\ & +\Psi(y) : \ y \in (1 - \tau_t)x_t + \tau_t x, \ x \in Q \Big\}. \end{aligned}$$

(6.4.50)

Note that this scheme is well defined even if the Hessian of the function $f$ is positive semidefinite. Of course, in general, the computational cost of each iteration of this scheme can be big. However, in one important case, when $\Psi(\cdot)$ is an indicator function of a Euclidean ball, the complexity of each iteration of this scheme is dominated by the complexity of matrix inversion. Thus, method (6.4.50) can be easily applied to problems of the form

$$\min_x \{ f(x) : \ \|x - x_0\| \leq r\}, \tag{6.4.51}$$

where the norm $\|\cdot\|$ is Euclidean.

Let $H_\nu < +\infty$ for some $\nu \in (0, 1]$. In this section we assume that

$$\langle \nabla^2 f(x)h, h\rangle \leq L\|h\|^2, \quad x \in Q, \ h \in \mathbb{E}. \tag{6.4.52}$$

Let us choose a sequence of nonnegative weights $\{a_t\}_{t\geq 0}$, and define in (6.4.50) the coefficients $\{\tau_t\}_{t\geq 0}$ in accordance with (6.4.13). Define the estimating functional sequence $\{\phi_t(x)\}_{t\geq 0}$ by recurrent relations (6.4.26), where the sequence $\{x_t\}_{t\geq 0}$ is generated by method (6.4.50). Finally, define

$$\hat{C}_{\nu,t} = a_0 \Delta(x_0) + \left(\sum_{k=1}^{t} \frac{a_k^{2+\nu}}{A_k^{1+\nu}}\right) \frac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)} + \left(\sum_{k=1}^{t} \frac{a_k^2}{2A_k}\right) L D^2. \tag{6.4.53}$$

In our convergence results, we also estimate the *second-order optimality measure* for problem (6.4.1) at the current test points. Let us introduce

$$\theta(x) \overset{\text{def}}{=} \max_{y\in Q}\{\langle \nabla f(x), x-y\rangle - \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle + \Psi(x) - \Psi(y)\}. \tag{6.4.54}$$

For any $x \in Q$ we have $\theta(x) \geq 0$. We call $\theta(x)$ the *total variation* of the quadratic model of the composite objective function in problem (6.4.1) over the feasible set. Defining

$$F_x(y) = \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle + \Psi(y),$$

we get $\theta(x) = \Big(F_x\Big)^*_{x,Q}(\nabla f(x))$ (see definition (6.4.21)).

**Theorem 6.4.4** *Let the sequence* $\{x_t\}_{t\geq 0}$ *be generated by method* (6.4.50)*. Then, for any* $\nu \in [0,1]$ *and any step* $t \geq 0$ *we have*

$$A_t \bar{f}(x_t) \leq \phi_t(x) + \hat{C}_{\nu,t}, \quad x \in Q. \tag{6.4.55}$$

*Moreover, for any* $t \geq 0$ *we have*

$$\bar{f}(x_t) - \bar{f}(x_{t+1}) \geq \tau_t \theta(x_t) - \frac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)} \tau_t^{2+\nu}. \tag{6.4.56}$$

*Proof* Let us prove inequality (6.4.55). For $t = 0$, $\hat{C}_{\nu,0} = a_0[\bar{f}(x_0) - \bar{f}(x_*)]$. Therefore, this inequality is valid.

In view of Theorem 3.1.23 the point $x_{t+1}$ is characterized by the following variational principle:

$$x_{t+1} = (1-\tau_t)x_t + \tau_t v_t, \quad v_t \in Q,$$

$$\Psi(y) + \langle \nabla f(x_t) + \nabla^2 f(x_t)(x_{t+1} - x_t), y - x_{t+1}\rangle \geq \Psi(x_{t+1}),$$

$$\forall\ y = (1-\tau_t)x_t + \tau_t x, \quad x \in Q.$$

Therefore, in view of definition (6.4.13), for any $x \in Q$ we have

$$\begin{aligned} a_{t+1}\langle \nabla f(x_t), x - x_t\rangle \quad &\geq \quad A_{t+1}\langle \nabla f(x_t) + \nabla^2 f(x_t)(x_{t+1} - x_t), x_{t+1} - x_t\rangle \\ &\quad + a_{t+1}\langle \nabla^2 f(x_t)(x_{t+1} - x_t), x_t - x\rangle \\ &\quad + A_{t+1}[\Psi(x_{t+1}) - \Psi((1-\tau_t)x_t + \tau_t x)] \\ &\overset{(6.4.52)}{\geq} A_{t+1}\langle \nabla f(x_t) + \tfrac{1}{2}\nabla^2 f(x_t)(x_{t+1} - x_t), x_{t+1} - x_t\rangle \\ &\quad + A_{t+1}[\Psi(x_{t+1}) - \Psi((1-\tau_t)x_t + \tau_t x)] - \tfrac{a_{t+1}^2}{2A_{t+1}} L D^2. \end{aligned}$$

Hence,

$$\begin{aligned} &A_t \bar{f}(x_t) + a_{t+1}[f(x_t) + \langle \nabla f(x_t), x - x_t\rangle + \Psi(x)] \\ \geq\ & A_t \Psi(x_t) + A_{t+1}[f(x_t) + \langle \nabla f(x_t) + \tfrac{1}{2}\nabla^2 f(x_t)(x_{t+1} - x_t), x_{t+1} - x_t\rangle] \\ & + a_{t+1}\Psi(x) + A_{t+1}[\Psi(x_{t+1}) - \Psi((1-\tau_t)x_t + \tau_t x)] - \tfrac{a_{t+1}^2}{2A_{t+1}} L D^2 \\ \overset{(6.4.6)}{\geq}\ & A_{t+1}[f(x_{t+1}) + \Psi(x_{t+1})] - A_{t+1}\tfrac{H_\nu \|x_{t+1} - x_t\|^{2+\nu}}{(1+\nu)(2+\nu)} - \tfrac{a_{t+1}^2}{2A_{t+1}} L D^2 \\ \geq\ & A_{t+1}\bar{f}(x_{t+1}) - \tfrac{a_{t+1}^{2+\nu}}{A_{t+1}^{1+\nu}} \cdot \tfrac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)} - \tfrac{a_{t+1}^2}{2A_{t+1}} L D^2. \end{aligned}$$

Thus, if (6.4.55) is valid for some $t \geq 0$, then

$$\begin{aligned} \phi_{t+1}(x) + \hat{C}_{\nu,t} &\geq A_t \bar{f}(x_t) + a_{t+1}[f(x_t) + \langle \nabla f(x_t), x - x_t\rangle + \Psi(x)] \\ &\geq A_{t+1}\bar{f}(x_{t+1}) - \tfrac{a_{t+1}^{2+\nu}}{A_{t+1}^{1+\nu}} \cdot \tfrac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)} - \tfrac{a_{t+1}^2}{2A_{t+1}} L D^2. \end{aligned}$$

Therefore, we can take $\hat{C}_{\nu,t+1} = \hat{C}_{\nu,t} + \frac{a_{t+1}^{2+\nu}}{A_{t+1}^{1+\nu}} \cdot \frac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)} + \frac{a_{t+1}^2}{2A_{t+1}} L D^2$.

In order to justify inequality (6.4.56), let us introduce the values

$$\begin{aligned} \theta_t(\tau) \quad &\overset{\text{def}}{=} \quad \max_{x \in Q}\{\langle \nabla f(x_t), x_t - y\rangle - \tfrac{1}{2}\langle \nabla^2 f(x_t)(y - x_t), y - x_t\rangle \\ &\qquad + \Psi(x_t) - \Psi(y) : \ y = (1-\tau)x_t + \tau x\} \\ &\overset{(6.4.22)}{=} \left(F_{x_t}\right)^*_{\tau, x_t, Q}(\nabla f(x_t)), \quad \tau \in [0, 1]. \end{aligned}$$

Clearly,

$$\begin{aligned}-\theta_t(\tau_t) &= \min_{x\in Q}\{\langle\nabla f(x_t), y-x_t\rangle - \tfrac{1}{2}\langle\nabla^2 f(x_t)(y-x_t), y-x_t\rangle \\ &\quad +\Psi(y)-\Psi(x_t) : y=(1-\tau_t)x_t+\tau_t x\} \\ &= \langle\nabla f(x_t), x_{t+1}-x_t\rangle - \tfrac{1}{2}\langle\nabla^2 f(x_t)(x_{t+1}-x_t), x_{t+1}-x_t\rangle \\ &\quad +\Psi(x_{t+1})-\Psi(x_t) \\ &\overset{(6.4.6)}{\geq} \bar{f}(x_{t+1})-\bar{f}(x_t)-\tfrac{H_\nu}{(1+\nu)(2+\nu)}\|x_{t+1}-x_t\|^{2+\nu}.\end{aligned}$$

Since $\|x_{t+1}-x_t\| \leq \tau_t D$, we conclude that

$$\bar{f}(x_t)-\bar{f}(x_{t+1}) \geq \theta_t(\tau_t) - \tfrac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)}\tau_t^{2+\nu} \overset{(6.4.23)}{\geq} \tau_t\theta(x_t) - \tfrac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)}\tau_t^{2+\nu}. \square$$

Thus, inequality (6.4.55) ensures the following rate of convergence of method (6.4.50)

$$\bar{f}(x_t)-\bar{f}(x_*) \leq \tfrac{1}{A_t}\hat{C}_{\nu,t}. \tag{6.4.57}$$

A particular expression of the right-hand side of this inequality for different values of $\nu \in [0,1]$ can be obtained in exactly the same way as it was done in Sect. 6.4.2. Here, we restrict ourselves only to the case when $\nu = 1$ and $a_t = t^2$, $t \geq 0$. Then $A_t = \frac{t(t+1)(2t+1)}{6}$, and

$$\begin{aligned}\sum_{k=1}^{t}\frac{a_k^3}{A_k^2} &= \sum_{k=1}^{t}\frac{36k^6}{k^2(k+1)^2(2k+1)^2} \;\leq\; 18t, \\ \sum_{k=1}^{t}\frac{a_k^2}{2A_k} &= \sum_{k=1}^{t}\frac{3k^4}{k(k+1)(2k+1)} \;\leq\; \tfrac{3}{2}\sum_{k=1}^{t}k = \tfrac{3}{4}t(t+1).\end{aligned}$$

Thus, we get

$$\bar{f}(x_t)-\bar{f}(x_*) \leq \tfrac{18H_1D^3}{(t+1)(2t+1)} + \tfrac{9LD^2}{2(2t+1)}. \tag{6.4.58}$$

Note that the rate of convergence (6.4.58) is worse than the convergence rate of cubic regularization of the Newton method (see Sect. 4.2.3). However, to the best of our knowledge, inequality (6.4.58) gives us the first global rate of convergence of an optimization scheme belonging to the family of trust-region methods. In view of inequality (6.4.55), the optimal solution of the dual problem (6.4.41) can be approximated by method (6.4.50) with $a_0 = 0$ in the same way as it was suggested in Sect. 6.4.4 for Conditional Gradient Methods.

Let us now estimate the rate of decrease of the values $\theta(x_t)$, $t \geq 0$, in the case when $\nu = 1$. Note that $\tau_t \overset{(6.4.13)}{=} \frac{a_{t+1}}{A_{t+1}} = \frac{6(t+1)}{(t+2)(2t+3)}$. It is easy to see that these coefficients satisfy the following inequalities:

$$\frac{3}{t+3} \leq \tau_t \leq \frac{6}{2t+5}, \quad t \geq 0. \tag{6.4.59}$$

Therefore, choosing the total number of steps $T = 2t + 2$, we have

$$\begin{aligned} \sum_{k=t}^{T} \tau_k &\overset{(6.4.59)}{\geq} 3 \sum_{k=t}^{2t+2} \frac{1}{k+3} \overset{(6.4.10)}{\geq} 3 \ln 2, \\ \sum_{k=t}^{T} \tau_k^3 &\overset{(6.4.59)}{\leq} \sum_{k=t}^{2t+2} \frac{27}{(k+5/2)^3} \overset{(6.4.11)}{\leq} -\frac{27}{2(k+5/2)^2}\Bigg|_{t-1/2}^{2t+5/2} \\ &= \frac{27}{2}\left[\frac{1}{(t+2)^2} - \frac{1}{(2t+5)^2}\right] = \frac{27}{2}\left[\frac{4}{(T+2)^2} - \frac{1}{(T+3)^2}\right] \\ &= \frac{27(3T+8)(T+4)}{2(T+2)^2(T+3)^2} \leq \frac{81}{2(T+1)(T+2)}. \end{aligned} \tag{6.4.60}$$

Now we can use the same trick as at the end of Sect. 6.4.2. Define

$$\theta_T^* = \min_{0 \leq t \leq T} \theta(x_t).$$

Then

$$\begin{aligned} \frac{36 H_1 D^3}{T(T-1)} + \frac{9LD^2}{2(T-1)} &\overset{(6.4.58)}{\geq} \bar{f}(x_t) - \bar{f}(x_*) \geq \sum_{k=t}^{T} (\bar{f}(x_k) - \bar{f}(x_{k+1})) \\ &\overset{(6.4.56)}{\geq} \theta_T^* \sum_{k=t}^{T} \tau_k - \frac{H_1 D^3}{6} \sum_{k=t}^{T} \tau_k^3 \\ &\overset{(6.4.60)}{\geq} 3\theta_T^* \ln 2 - \frac{27 H_1 D^3}{4(T+1)(T+2)}. \end{aligned}$$

Thus, for even $T$, we get the following bound:

$$\begin{aligned} \theta_T^* &\leq \frac{3}{\ln 2}\left[\frac{4H_1 D^3}{T(T-1)} + \frac{3H_1 D^3}{4(T+1)(T+2)} + \frac{LD^2}{2(T-1)}\right] \\ &\leq \frac{3}{\ln 2}\left[\frac{5H_1 D^3}{T(T-1)} + \frac{LD^2}{2(T-1)}\right]. \end{aligned} \tag{6.4.61}$$

# Chapter 7
# Optimization in Relative Scale

In many applications, it is difficult to relate the number of iterations in an optimization scheme with the desired accuracy of the solution since the corresponding inequality contains unknown parameters (Lipschitz constant, distance to the optimum). However, in many cases the required level of relative accuracy is quite understandable. To develop 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. The availability of the 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. We also present a barrier subgradient method, which computes an approximate maximum of a positive convex function with certain relative accuracy. We show how to apply this method to solve problems of fractional covering, maximal concurrent flow, semidefinite relaxation, online optimization, portfolio management, and others. Finally, we consider a class of strictly positive functions, for which a kind of quasi-Newton method is developed.

## 7.1 Homogeneous Models of an Objective Function

(The conic unconstrained minimization problem; The subgradient approximation scheme; Structural optimization; Application examples: Linear Programming, Minimization of the spectral radius; The truss topology design problem.)

Y. Nesterov, *Lectures on Convex Optimization*, Springer Optimization
and Its Applications 137, https://doi.org/10.1007/978-3-319-91578-4_7

### *7.1.1 The Conic Unconstrained Minimization Problem*

Quite often, in the theoretical justification of convex optimization methods it is assumed that problems have bounded feasible sets. Besides its technical convenience, this assumption allows us to introduce a reasonable scale for measuring the absolute accuracy of an approximate solution. In the cases when the initial problem does not possess this property, some algorithms require an artificial bounding of the domain (the "big $M$" approach). This approach is, perhaps, acceptable for polynomial-time methods, where the "big M" enters the complexity estimates only inside a logarithm term (see Chap. 5). However, it is clear that for gradient-type methods, such a strategy cannot work.

In fact, this is almost a philosophical question: Do the problems with unbounded feasible sets really arise in practice? And if so, how they should be treated? Actually, there is at least one, very important class of such problems, namely, the problems obtained by Lagrangian relaxations of inequality constraints (see Sects. 1.3.3 and 3.1.7). If there were some reasonable bounds on the dual variables for these constraints, then it would be natural to incorporate them into the primal problem. Then, instead of constraints in the primal problem, we could have an additional term in the objective function.

Another difficulty is related to the way of bounding unbounded feasible sets. It is not always possible to find a reasonable localization set a priori, without collecting additional information on the topology of the problem by some auxiliary computations.

In this chapter, we suggest an alternative way of treating the convex minimization problems. Namely, we are going to compute their approximate solutions in *relative* scale. We will see that this idea works at least for a special class of *conic* unconstrained minimization problems.[1] These are the problems of minimizing a positively homogeneous convex function over a convex set, which is separated from the origin. In order to compute an approximate solution to this problem with a certain *relative accuracy*, we need to know a John ellipsoid for the subdifferential of the objective function evaluated at the origin. We will see that in many cases all necessary information about the objective function can be easily obtained by analyzing its structure.

In what follows, we say that the value $f(\bar{x})$ approximates the optimal value $f^* > 0$ with *relative accuracy* $\delta$ if

$$f^* \leq f(\bar{x}) \leq (1+\delta) f^*.$$

In this chapter, it is convenient to use the following notation for the balls in $\mathbb{E}$ with respect to $\|\cdot\|$:

$$B_{\|\cdot\|}(r) = \{x \in \mathbb{E} : \|x\| \leq r\}.$$

[1] By this term we mean problems with no functional constraints.

The notation $\pi_{Q,\|\cdot\|}(x)$ is used for the projection of a point $x$ onto the set $Q$ with respect to the norm $\|\cdot\|$. For the sake of notation, if no ambiguity arises, the indication of the norm is omitted.

Finally, in the case $\mathbb{E} = \mathbb{R}^n$, $I_n$ denotes the unit matrix in $\mathbb{R}^n$, $e_i$ denotes the $i$th coordinate vector, and $\bar{e}_n$ stands for the vector of all ones. For an $n \times n$ matrix $X$ we denote by $\lambda_1(X), \dots, \lambda_n(X)$ its spectrum of eigenvalues numbered in decreasing order.

The most general form of the optimization problem considered in this section is as follows:

$$\text{Find} \quad f^* = \min_{x \in Q_1} f(x), \tag{7.1.1}$$

where $f$ is a convex positively homogeneous function of degree one (see the end of Sect. 3.1.6), and $Q_1 \subset \mathbb{E}$ is a closed convex set, which does not contain the origin. In many applications, the role of $Q_1$ is played by an affine subspace

$$\mathscr{L} = \{x \in \mathbb{E} : \; Cx = b\},$$

where $b \in \mathbb{E}_1$, $b \neq 0$, and $C : \mathbb{E} \to \mathbb{E}_1$. Without loss of generality, we can assume that $C$ is non-degenerate.

Our main assumption on problem (7.1.1) is that

$$\operatorname{dom} f \equiv \mathbb{E}, \quad 0 \in \operatorname{int} \partial f(0). \tag{7.1.2}$$

In other words, we assume that $f$ is a support function of a convex compact set containing the origin in its interior. Then $f^* > 0$, and the problem of finding an approximate solution to (7.1.1) with a certain *relative accuracy* becomes well posed. In what follows, we call the setting (7.1.1), (7.1.2) the *conic* unconstrained minimization problem.

Note that any unconstrained minimization problem

$$\min_{y \in \mathbb{E}} \phi(y),$$

with convex objective $\phi(\cdot)$, can be rewritten in the form (7.1.1) by simple homogenization:

$$x = (y, \tau) \in \mathbb{E} \times \mathbb{R}_+, \quad f(x) = \tau\phi(y/\tau), \quad Cx \equiv \tau, \quad b = 1$$

(see Example 3.1.2(6)). However, in general, we cannot guarantee that such a function satisfies assumption (7.1.2).

Let us look at the following examples.

*Example 7.1.1* Let our initial problem consist in finding approximately an unconstrained minimum of the function

$$\phi_\infty(y) = \max_{1 \le i \le m} |\langle a_i, y\rangle + c^{(i)}|, \quad y \in \mathbb{R}^{n-1}.$$

Let us introduce $x = \begin{pmatrix} y \\ \tau \end{pmatrix}$, and $\hat{a}_i = \begin{pmatrix} a_i \\ c^{(i)} \end{pmatrix}$, $i = 1, \dots, m$. Let

$$A^T = \left(\hat{a}_1 \dots, \hat{a}_m\right), \quad F_\infty(v) = \max_{1 \le i \le m} |v^{(i)}|,$$

$$p = 1, \quad C = (\underbrace{0, \dots, 0}_{(n-1) \text{ times}}, 1), \quad b = 1.$$

Then for positive $\tau$ we can define

$$f(x) = \tau\phi_\infty(y/\tau) \equiv F_\infty(Ax), \quad Q_1 \; = \; \mathscr{L}.$$

Thus, this description of $f(\cdot)$ can be extended onto the whole space.

In a similar way, for the function

$$\phi_1(y) = \sum_{i=1}^m |\langle a_i, y\rangle + c^{(i)}|, \quad y \in \mathbb{R}^{n-1},$$

we can get a representation (7.1.1), which satisfies (7.1.2). In this case, we use $f(x) = F_1(Ax)$ with

$$F_1(v) = \sum_{i=1}^m |v^{(i)}|.$$

However, for the function

$$\phi(y) = \max_{1 \le i \le m} \left\{\langle a_i, y\rangle + c^{(i)}\right\}, \quad y \in \mathbb{R}^{n-1},$$

the above lifting cannot guarantee (7.1.2). □

Let us fix some norm $\|\cdot\|$ in $\mathbb{E}$, and define the dual norm in the standard way:

$$\|g\|^* = \max_{\|x\| \le 1} \langle s, x\rangle, \quad g \in \mathbb{E}^*. \tag{7.1.3}$$

Then we can rewrite our main assumption (7.1.2) in a quantitative form. Let $\gamma_0 \le \gamma_1$ be some positive values satisfying the following *asphericity* condition:

$$B_{\|\cdot\|^*}(\gamma_0) \subseteq \partial f(0) \subseteq B_{\|\cdot\|^*}(\gamma_1). \tag{7.1.4}$$

Thus, by (7.1.2) we just assume that such values are well defined. Note that these values depend on the choice of the norm $\|\cdot\|$. In the sequel, this choice will always be evident from the context.

Denote by

$$\alpha = \tfrac{\gamma_0}{\gamma_1} < 1,$$

the *asphericity coefficient* of the function $f$. As we will see later, this parameter is crucial for complexity bounds of finding approximate solutions to problem (7.1.1) with a certain relative accuracy.

Note that in many situations it is reasonable to choose $\|\cdot\|$ as an ellipsoidal norm. In view of John's theorem, for a good variant of this norm we can guarantee that

$$\alpha \geq \tfrac{1}{n}, \tag{7.1.5}$$

where $n = \dim \mathbb{E}$. Moreover, if $\partial f(0)$ is symmetric:

$$f(x) = f(-x) \quad \forall x \in \mathbb{E},$$

then the lower bound for ellipsoidal norms is even better:

$$\alpha \geq \tfrac{1}{\sqrt{n}}. \tag{7.1.6}$$

(We will prove both variants of John's Theorem in Sect. 7.2.) Of course, it may be difficult to find a norm which is good for a particular objective function $f$. However, in this case we can try to employ our knowledge of its structure.

For example, it may happen that we know a self-concordant barrier $\psi(\cdot)$ for the convex set $\partial f(0)$ (see Sect. 5.3), and $\nabla\psi(0) = 0$. Then we can use

$$\|v\|^* = \langle v, \nabla^2\psi(0)v\rangle^{1/2}, \quad \|x\| = \langle [\nabla^2\psi(0)]^{-1}x, x\rangle^{1/2}.$$

In this case, it is possible to choose

$$\gamma_0 = 1, \quad \gamma_1 = \nu + 2\sqrt{\nu},$$

where $\nu$ is the parameter of the barrier $\psi(\cdot)$ (see Theorem 5.3.9).

For some important problems the subdifferential $\partial f(0)$ is a polyhedral set. Then the following result may be useful.

**Lemma 7.1.1** *Let* $f(x) = \max\limits_{1\leq i\leq m} \langle a_i, x\rangle$, $x \in \mathbb{R}^n$. *Assume that the matrix*

$$A = (a_1, \ldots, a_m)$$

*has full row rank and* $\sum\limits_{i=1}^{m} a_i = 0$ *(thus,* $m > n$*). Then the norm*

$$\|x\| = \left[ \sum_{i=1}^{m} \langle a_i, x\rangle^2 \right]^{1/2}$$

*is well defined . We can choose* $\gamma_1 = 1$ *and* $\gamma_0 = \frac{1}{\sqrt{m(m-1)}}$.

*Proof* Note that the matrix $G = \sum\limits_{i=1}^{m} a_i a_i^T$ is non-degenerate. Then

$$\|v\|^* = \langle v, G^{-1}v\rangle^{1/2}$$

(see Lemma 3.1.20), and therefore for any $i = 1, \dots, m$ we have

$$\begin{aligned}(\|a_i\|^*)^2 \;=\; \langle a_i, G^{-1}a_i\rangle &= \max_{x\in\mathbb{R}^n}\{2\langle a_i, x\rangle - \langle Gx, x\rangle\}\\ &= \max_{x\in\mathbb{R}^n}\left\{2\langle a_i, x\rangle - \sum_{k=1}^{m}\langle a_k, x\rangle^2\right\}\\ &\le \max_{x\in\mathbb{R}^n}\{2\langle a_i, x\rangle - \langle a_i, x\rangle^2\} \;=\; 1.\end{aligned}$$

Since $\partial f(0) = \mathrm{Conv}\ \{a_i,\ i = 1, \dots, m\}$, we can take $\gamma_1 = 1$.

On the other hand, for any $x \in \mathbb{R}^n$ we have $\sum\limits_{i=1}^{m}\langle a_i, x\rangle = 0$. Therefore,

$$\begin{aligned}\langle Gx, x\rangle &= \sum_{i=1}^{m}\langle a_i, x\rangle^2\\ &\le \max_{s\in\mathbb{R}^m}\left\{\sum_{i=1}^{m}(s^{(i)})^2 :\ \sum_{i=1}^{m} s^{(i)} = 0,\ s^{(i)} \le f(x),\ i = 1, \dots, m\right\}.\end{aligned}$$

In view of Corollary 3.1.2, the extremum in the above maximization problem is attained, for example, at

$$\hat{s} = f(x)\cdot(\bar{e}_m - me_1).$$

This means that $\langle Gx, x\rangle \le m(m-1)f^2(x)$. Hence, $f(x) \ge \frac{\|x\|}{\sqrt{m(m-1)}}$. In view of representation (3.1.41), this justifies the choice $\gamma_0 = \frac{1}{\sqrt{m(m-1)}}$. □

The possibility of employing another structural representation of problem (7.1.1) is discussed in Sect. 7.1.3.

Let us conclude this section with a statement which supports our ability to solve problem (7.1.1) with a certain relative accuracy.

Denote by $x_0$ the projection of the origin onto the set $Q_1$ with respect to the norm $\| \cdot \|$[2]:

$$\|x_0\| = \min_{x \in Q_1} \|x\|.$$

**Theorem 7.1.1**

1. *For any $x \in \mathbb{R}^n$, we have*

$$\gamma_0 \cdot \|x\| \le f(x) \;\le\; \gamma_1 \cdot \|x\|. \tag{7.1.7}$$

*Therefore the function $f$ is Lipschitz continuous on $\mathbb{E}$ in the norm $\| \cdot \|$ with Lipschitz constant $\gamma_1$. Moreover,*

$$\alpha f(x_0) \le \gamma_0 \cdot \|x_0\| \;\le\; f^* \;\le\; f(x_0) \;\le\; \gamma_1 \cdot \|x_0\|. \tag{7.1.8}$$

2. *For any optimal solution $x^*$ to (7.1.1), we have*

$$\|x_0 - x^*\| \le \tfrac{2}{\gamma_0} f^* \;\le\; \tfrac{2}{\gamma_0} f(x_0). \tag{7.1.9}$$

*If the norm $\| \cdot \|$ is Euclidean, then this inequality can be strengthened as follows:*

$$\|x_0 - x^*\| \le \tfrac{1}{\gamma_0} f^* \;\le\; \tfrac{1}{\gamma_0} f(x_0). \tag{7.1.10}$$

*Proof* For any $x \in \mathbb{E}$, we have

$$f(x) \overset{(3.1.41)}{=} \max_v \{\langle v, x\rangle : \; v \in \partial f(0)\} \;\ge\; \max_v \{\langle v, x\rangle : \; v \in B_{\|\cdot\|_*}(\gamma_0)\} \;=\; \gamma_0 \|x\|,$$

$$f(x) \overset{(3.1.41)}{=} \max_v \{\langle v, x\rangle : \; v \in \partial f(0)\} \;\le\; \max_v \{\langle v, x\rangle : \; v \in B_{\|\cdot\|_*}(\gamma_1)\} \;=\; \gamma_1 \|x\|.$$

Therefore, for any $x$ and $h \in \mathbb{E}$, we have

$$f(x+h) \le f(x) + f(h) \le f(x) + \gamma_1 \|h\|.$$

Moreover,

$$f^* = \min_{x \in Q_1} f(x) \;\ge\; \min_{x \in Q_1} \gamma_0 \|x\| \;=\; \gamma_0 \|x_0\|.$$

[2] Recall that this can be any general norm.

Hence, in view of (7.1.7) we have

$$f^* \geq \gamma_0 \|x_0\| \;\geq\; \alpha f(x_0),$$

$$f^* \leq f(x_0) \;\leq\; \gamma_1 \|x_0\|.$$

In order to prove the second statement, note that in view of the first item of the theorem we have

$$\|x_0 - x^*\| \leq \|x_0\| + \|x^*\| \;\leq\; \frac{2}{\gamma_0} \cdot f^*.$$

For the Euclidean norm $\|x\| = \langle Gx, x\rangle^{1/2}$ with $G \succ 0$, this bound can be strengthened. Indeed, in this case $\langle Gx_0, x^* - x_0\rangle \overset{(2.2.39)}{\geq} 0$. Therefore,

$$\begin{aligned} \|x_0 - x^*\|^2 = \|x_0\|^2 - 2\langle Gx_0, x^*\rangle + \|x^*\|^2 \;&\leq\; \|x^*\|^2 - \|x_0\|^2 \\ &< \|x^*\|^2. \qquad \square \end{aligned}$$

## 7.1.2 *The Subgradient Approximation Scheme*

Let us discuss now different possibilities for finding an approximate solution to problem (7.1.1). For the sake of simplicity, we assume that the norm $\| \cdot \|$ is Euclidean.

The first of our schemes is based on the standard Subgradient Method for minimizing non-smooth convex functions. Denote by $g(x)$ an arbitrary subgradient of the function $f$ at point $x$. Consider the simplest variant of the Subgradient Method as applied to problem (7.1.1).

```
**Subgradient Method** $G_N(R)$

**for** $k := 0$ **to** $N$ **do:**   Compute $f(x_k)$ and $g(x_k)$.
    $x_{k+1} := \pi_{Q_1}\left(x_k - \frac{R}{\sqrt{N+1}} \cdot \frac{g(x_k)}{\|g(x_k)\|^*}\right)$.          (7.1.11)

**Output:**   $\bar{x} = \arg\min_x \{f(x) : \; x = x_0, \ldots, x_N\}$.
```

In what follows, the output of this process $\bar{x} \in \mathbb{E}$ is denoted by $G_N(R)$. In view of Theorem 3.2.2, the rate of convergence of this method is as follows:

$$f(G_N(R)) - f^* \le \frac{\gamma_1}{\sqrt{N+1}} \cdot \frac{\|x_0 - x^*\|^2 + R^2}{2R}. \tag{7.1.12}$$

Thus, in order to be efficient, the Subgradient Method needs a good estimate for the distance between the starting point $x_0$ and the solution $x^*$:

$$R \;\approx\; \|x_0 - x^*\|.$$

In our case, this estimate could be obtained from the first inequality in (7.1.10). However, since the value $f^*$ is not known in advance, we will use the second part of this inequality:

$$\hat{\rho} \overset{\text{def}}{=} \frac{1}{\gamma_0} f(x_0) \;\ge\; \|x_0 - x^*\|. \tag{7.1.13}$$

The performance of the corresponding scheme is given by the following statement.

**Theorem 7.1.2** *For a fixed $\delta$ from $(0, 1)$, let us choose*

$$N = \left\lfloor \frac{1}{\alpha^4 \delta^2} \right\rfloor. \tag{7.1.14}$$

*Then $f(G_N(\hat{\rho})) \le (1+\delta) \cdot f^*$.*

*Proof* In view of inequality (7.1.12), the choice (7.1.14) and inequalities (7.1.10), (7.1.8), we have

$$f(G_N(\hat{\rho})) - f^* \le \alpha^2 \delta \gamma_1 \cdot \frac{\|x_0 - x^*\|^2 + \hat{\rho}^2}{2\hat{\rho}} \;\le\; \alpha^2 \delta \gamma_1 \hat{\rho} \;=\; \alpha \delta f(x_0)$$
$$\le \delta \cdot f^*. \qquad \square$$

Note that we pay a high price for the poor estimate of the initial distance. If we were be able to use the first part of inequality (7.1.10), then the corresponding complexity bound could be much better. Let us show that a better bound for the distance to the optimal solution can be derived from the trivial observation that $f^* \le f(x)$ for any point $x$ from $Q_1$.

Denote by $\delta \in (0, 1)$ the desired relative accuracy. Let

$$\hat{N} = \left\lfloor \frac{e}{\alpha^2} \cdot \left(1 + \frac{1}{\delta}\right)^2 \right\rfloor,$$

where $e$ is the base of the exponent. Consider the following *restarting strategy*. Set $\hat{x}_0 = x_0$, and for $t \ge 1$ iterate

$$\boxed{\begin{array}{l} \hat{x}_t := G_{\hat{N}}\left(\frac{1}{\gamma_0} f(\hat{x}_{t-1})\right); \\ \\ \textbf{if } f(\hat{x}_t) \ge \frac{1}{\sqrt{e}} f(\hat{x}_{t-1}) \textbf{ then } T := t\textbf{and} \text{ Stop.} \end{array}} \tag{7.1.15}$$

**Theorem 7.1.3** *The number of points generated by the process (7.1.15) is bounded:*

$$T \le 1 + 2\ln\frac{1}{\alpha}. \tag{7.1.16}$$

*The last generated point satisfies the inequality $f(\hat{x}_T) \le (1+\delta)f^*$. The total number of lower-level gradient steps in the process (7.1.15) does not exceed*

$$\frac{e}{\alpha^2} \cdot \left(1 + \frac{1}{\delta}\right)^2 \cdot \left(1 + 2\ln\frac{1}{\alpha}\right). \tag{7.1.17}$$

*Proof* By simple induction, it is easy to prove that at the beginning of stage $t$ in (7.1.15) the following inequality holds:

$$\left(\frac{1}{\sqrt{e}}\right)^{t-1} f(x_0) \ge f(\hat{x}_{t-1}), \quad t \ge 1.$$

Thus, in view of inequality (7.1.8), at the last stage $T$ of the process we have

$$\left(\frac{1}{\sqrt{e}}\right)^{T-1} f(x_0) \ge f(\hat{x}_{T-1}) \ \ge\ f^* \ \ge\ \alpha f(x_0).$$

This leads to inequality (7.1.16).

In view of (7.1.10), we have $\|x_0 - x^*\| \le \frac{1}{\gamma_0} f^* \le \frac{1}{\gamma_0} f(\hat{x}_{T-1})$. Therefore, at the last stage of the process, using (7.1.12) and the termination rule in (7.1.15), we get

$$\begin{aligned} f(\hat{x}_T) - f^* &\le \frac{\gamma_1}{\sqrt{\hat{N}+1}} \cdot \frac{1}{\gamma_0} \cdot f(\hat{x}_{T-1}) \ \le\ \frac{\sqrt{e}}{\alpha\sqrt{\hat{N}+1}} \cdot f(\hat{x}_T) \\ &\le \frac{\delta}{1+\delta} \cdot f(\hat{x}_T). \qquad \square \end{aligned}$$

## 7.1.3 Direct Use of the Problem Structure

In Sect. 7.1.2 we have shown that the outer and inner ellipsoidal approximations of the set $\partial f(0)$ are the key ingredients of minimization schemes for computing an approximate solution to problem (7.1.1) in relative scale. However, in order to find an ellipsoidal norm, which is good for our problem, we need to employ its structure somehow. In this section, we introduce a model of problem (7.1.1) which is suitable both for the explicit indication of such a norm and for applying the smoothing technique described in Sect. 6.1. We will see that the efficiency of the latter approach significantly dominates that of the Subgradient Method.

Since the objective function $f$ in problem (7.1.1) is positive homogeneous, the simplest possible structure of such an object could be as follows. Let us assume that the objective function $f$ is a composition of two objects, a linear operator $A(x)$ and a *simple* nonlinear convex homogeneous function $F$. In other words, assume that $f(x) = F(A(x))$. Let us introduce this object in a formal way. In this section we switch to the notation of Sect. 6.1, choosing $\mathbb{E}_1 = \mathbb{R}^n$ and $\mathbb{E}_2 = \mathbb{R}^m$.

Let $Q_2$ be a closed bounded convex set in $\mathbb{R}^m$ containing the origin in its interior. Define a convex homogeneous function $F$ as follows:

$$F(v) = \max_{u \in Q_2} \langle v, u \rangle_{\mathbb{R}^m}. \tag{7.1.18}$$

Further, let $A$ be an $m \times n$-matrix which has a full column rank (thus, $m \geq n$). Define the objective function

$$f(x) = F(Ax), \quad x \in \mathbb{R}^n. \tag{7.1.19}$$

Clearly, $f$ is a convex function with degree of homogeneity one. Our problem of interest is still (7.1.1), which we repeat for convenience here:

$$\text{Find } f^* = \min_{x \in Q_1} \ f(x). \tag{7.1.20}$$

Since $\partial F(0) \equiv Q_2$, we have $\partial f(0) = A^T Q_2$ (see Lemma 3.1.11). Thus, problem (7.1.20) satisfies the main assumption (7.1.2).

Let $\| \cdot \|_{\mathbb{R}^m}$ denote the standard Euclidean norm in $\mathbb{R}^m$:

$$\|u\|_{\mathbb{R}^m} = \left[ \sum_{i=1}^{m} \left(u^{(i)}\right)^2 \right]^{1/2}, \quad u \in \mathbb{R}^m.$$

Let us introduce the following characteristics of the function $F$:

$$\gamma_0(F) = \max_{r>0}\{r : \ B_{\|\cdot\|_{R^m}}(r) \subseteq \partial F(0)\},$$

$$\gamma_1(F) = \min_{r>0}\{r : \ B_{\|\cdot\|_{\mathbb{R}^m}}(r) \supseteq \partial F(0)\},$$

$$\alpha(F) = \tfrac{\gamma_1(F)}{\gamma_0(F)} \geq 1.$$

For the sets from Example 7.1.1, these values are as follows:

$$\begin{aligned} &\gamma_0(F_1) = \tfrac{1}{\sqrt{m}},\ \gamma_1(F_1) = 1, \qquad \alpha(F_1) = \sqrt{m}, \\ &\gamma_0(F_\infty) = 1, \quad \gamma_1(F_\infty) = \sqrt{m},\ \alpha(F_\infty) = \sqrt{m}. \end{aligned} \tag{7.1.21}$$

Let us define now the following Euclidean norm in the primal space:

$$\| x \|_{\mathbb{R}^n} = \| Ax \|^*_{\mathbb{R}^m}, \quad x \in \mathbb{R}^n. \tag{7.1.22}$$

Since $A$ is non-degenerate, this norm is well defined. Defining $G = A^T A \succ 0$, we get the following representations:

$$\begin{aligned} \| x \|_{\mathbb{R}^n} &= \langle Gx, x \rangle^{1/2} = \left[ \sum_{i=1}^m \langle a_i, x \rangle^2 \right]^{1/2}, \\ \| g \|^*_{\mathbb{R}^n} &= \langle g, G^{-1} g \rangle^{1/2}, \end{aligned} \tag{7.1.23}$$

where $a_i$, $i = 1, \ldots, m$, denote the columns of the matrix $A^T$.

**Lemma 7.1.2** *For norm $\| \cdot \|_{\mathbb{R}^n}$, condition (7.1.4) holds with*

$$\gamma_0 = \gamma_0(F), \quad \gamma_1 = \gamma_1(F).$$

*Thus, we can take $\alpha = \alpha(F) = \frac{\gamma_0(F)}{\gamma_1(F)}$.*

*Proof* Since $\partial f(0) = A^T Q_2$, we have the following representation for the support function of this set:

$$\xi(x) \overset{\text{def}}{=} \max_{s \in \partial f(0)} \langle s, x \rangle_{\mathbb{R}^n} = \max_{u \in Q_2} \langle A^T u, x \rangle_{\mathbb{R}^m} = \max_{u \in Q_2} \langle Ax, u \rangle_{\mathbb{R}^m}.$$

Thus,

$$\xi(x) \le \max_{\| u \|_2 \le \gamma_1(F)} \langle Ax, u \rangle_{\mathbb{R}^m} = \gamma_1(F) \| Ax \|^*_{\mathbb{R}^m} = \gamma_1(F) \| x \|_{\mathbb{R}^n},$$

$$\xi(x) \ge \max_{\| u \|_{\mathbb{R}^m} \le \gamma_0(F)} \langle Ax, u \rangle_{\mathbb{R}^m} = \gamma_0(F) \| Ax \|^*_{\mathbb{R}^m} = \gamma_0(F) \| x \|_{\mathbb{R}^n}.$$

Hence, in view of Corollary 3.1.5, $\partial f(0) \subseteq B_{\| \cdot \|^*_1}(\gamma_1(F))$, and $\partial f(0) \supseteq B_{\| \cdot \|^*_1}(\gamma_0(F))$. □

Note that for many simple sets $Q_2$, parameters $\gamma_1(F)$ and $\gamma_0(F)$ are easily available (see, for example, (7.1.21)). Therefore, metric (7.1.23) can be used to find an approximate solution to the corresponding problems by the Subgradient

Method (7.1.15). However, the main advantage of representation (7.1.19) is related to the possibility of employing the smoothing technique of Sect. 6.1. Let us show how this can be done.

Problem (7.1.20) differs from problem (6.1.10) only in one aspect: it can have an unbounded primal feasible set. Thus, a straightforward application of the efficient smoothing technique to (7.1.20) is impossible. However, we can introduce an artificial bound on the size of the optimal solution using the information provided by inequality (7.1.10). Define

$$Q_1(\rho) = \{x \in Q_1 : \ \|x - x_0\|_{\mathbb{R}^n} \leq \rho\}.$$

In view of (7.1.10), we have $x^* \in Q_1(\hat{\rho})$ for $\hat{\rho} = \frac{1}{\gamma_0(F)} f(x_0)$. Thus, problem (7.1.20) is equivalent to the following:

$$\begin{aligned} \text{Find } f^* &= \min_{x \in \mathbb{R}^n} \{f(x) : \ x \in Q_1(\hat{\rho})\} \\ &= \min_{x \in Q_1(\hat{\rho})} \max_{u \in Q_2} \langle Ax, u\rangle_{\mathbb{R}^m} \\ &\overset{(6.1.34)}{=} \max_{u \in \mathbb{R}^m} \{\phi_{\hat{\rho}}(u) : \ u \in Q_2\}, \end{aligned} \tag{7.1.24}$$

where $\phi_\rho(u) = \min\limits_{x \in Q_1(\rho)} \langle Ax, u\rangle_{\mathbb{R}^m}$. Thus, we have managed to represent our problem in the form required by Sect. 6.1.

Let us introduce the objects necessary for applying the smoothing technique. In the primal space, we choose the prox-function $d_1(x) = \frac{1}{2}\|x - x_0\|^2_{\mathbb{R}^n}$. This function has convexity parameter equal to one. Its maximum on the feasible set $Q_1(\hat{\rho})$ does not exceed $D_1 = \frac{1}{2}\hat{\rho}^2$.

Similarly, for the dual feasible set, we choose $d_2(u) = \frac{1}{2}\|u\|^2_{\mathbb{R}^m}$. Then its convexity parameter is one, and the maximum of this function on the dual feasible set $Q_2$ is smaller than $D_2 = \frac{1}{2}\gamma_1^2(F)$. It remains to note that

$$\begin{aligned} \|A\|_{1,2} &= \max_{x,u}\{\langle Ax, u\rangle_{\mathbb{R}^m} : \ \|x\|_{\mathbb{R}^n} \leq 1, \ \|u\|_{\mathbb{R}^m} \leq 1\} \\ &= \max_{x}\{\|Ax\|^*_{\mathbb{R}^m} : \ \|x\|_{\mathbb{R}^n} \leq 1\} \\ &\overset{(7.1.22)}{=} \max_{x}\{\|x\|_{\mathbb{R}^n} : \ \|x\|_{\mathbb{R}^n} \leq 1\} \ = \ 1. \end{aligned} \tag{7.1.25}$$

For the reader's convenience we present here the algorithm (6.1.19) adopted for our needs. This method is applied to a smooth approximation of the objective function $f$:

$$f_\mu(x) = \max_{u \in Q_2}\{\langle Ax, u\rangle_{\mathbb{R}^m} - \mu d_2(u)\}, \quad x \in \mathbb{R}^n. \tag{7.1.26}$$

In view of Theorem 6.1.1, this function has Lipschitz continuous gradient

$$\nabla f_\mu(x) = A^T u_\mu(x),$$

where $u_\mu(x)$ is a unique solution to the optimization problem in (7.1.26). In view of equality (7.1.25), the Lipschitz constant for the gradient is equal to $\frac{1}{\mu}$.

```
Method $S_N(R)$

Set $\mu = \frac{2R}{\gamma_1(F)\cdot\sqrt{N(N+1)}}$ and $v_0 = x_0$.

for $k := 0$ to $N - 1$ do
    $y_k = \frac{k}{k+2} x_k + \frac{2}{k+2} v_k$,
    $u_\mu(y_k) = \arg\max\limits_{u \in Q_2} \{\langle A y_k, u\rangle_{\mathbb{R}^m} - \frac{\mu}{2}\|u\|^2_{\mathbb{R}^m}\}$,
    $v_{k+1} = \arg\min\limits_{x \in Q_1(R)} \left\{ \frac{1}{2\mu}\|x - x_0\|^2_{\mathbb{R}^n} + \langle Ax, \sum\limits_{i=0}^{k} \frac{i+1}{2} u_\mu(y_i)\rangle_{\mathbb{R}^m} \right\}$,
    $x_{k+1} = \frac{k}{k+2} x_k + \frac{2}{k+2} v_{k+1}$.

Output: $\bar{x} := x_N$.
```

(7.1.27)

In what follows, we denote the output $\bar{x} \in \mathbb{R}^n$ of this process by $S_N(R)$. It is easy to check that all conditions of Theorem 6.1.3 are satisfied. Thus, if $\|x_0 - x^*\|_{\mathbb{R}^n} \leq R$, then the output of this process satisfies inequality

$$f(S_N(R)) - f^* \leq \frac{2\gamma_1(F)R}{\sqrt{N(N+1)}}. \tag{7.1.28}$$

This observation has an important corollary.

**Theorem 7.1.4** *For* $\delta \in (0, 1)$, *let*

$$N = \left\lfloor \frac{2}{\alpha^2(F)\,\delta} \right\rfloor. \tag{7.1.29}$$

*Then* $f\left(S_N\left(\frac{1}{\gamma_0(F)} f(x_0)\right)\right) \leq (1 + \delta) f^*$.

*Proof* Since $\|x_0 - x^*\|_{\mathbb{R}^n} \overset{(7.1.10)}{\leq} \frac{1}{\gamma_0(F)} f(x_0)$, and $N+1 \overset{(7.1.29)}{\geq} \frac{2}{\alpha^2(F)\,\delta}$, from (7.1.28) and (7.1.8) we have

$$f(S_N(R)) - f^* \leq \delta \cdot \alpha(F) f(x_0) \leq \delta \cdot f^*. \qquad \square$$

Note that the complexity bound (7.1.29) of the scheme (7.1.27) is lower even than the bound of the Subgradient Method (7.1.15) with a recursively updated estimate for the distance to the optimum. Let us show that a similar updating strategy can also accelerate scheme (7.1.27).

Let $\delta \in (0, 1)$ be the desired relative accuracy. Let

$$\tilde{N} = \left\lfloor \frac{2e}{\alpha(F)} \cdot \left(1 + \frac{1}{\delta}\right) \right\rfloor .$$

Consider the following restarting strategy. Set $\hat{x}_0 = x_0$. For $t \geq 1$ iterate

$$\boxed{\begin{array}{l} \hat{x}_t := S_{\tilde{N}}\left(\frac{1}{\gamma_0(F)} f(\hat{x}_{t-1})\right); \\[2ex] \textbf{if } f(\hat{x}_t) \geq \frac{1}{e} f(\hat{x}_{t-1}) \textbf{ then } T := t \textbf{ and } \text{Stop}. \end{array}} \tag{7.1.30}$$

**Theorem 7.1.5** *The number of points $T$ generated by scheme (7.1.30) is bounded as follows:*

$$T \leq 1 + \ln \frac{1}{\alpha(F)}. \tag{7.1.31}$$

*The last generated point satisfies inequality $f(\hat{x}_T) \leq (1+\delta) f^*$. The total number of lower-level steps in the process (7.1.30) does not exceed*

$$\frac{2e}{\alpha(F)} \cdot \left(1 + \frac{1}{\delta}\right) \cdot \left(1 + \ln \frac{1}{\alpha(F)}\right). \tag{7.1.32}$$

*Proof* By simple induction it is easy to prove that at the beginning of stage $t$ the following inequality holds:

$$\left(\frac{1}{e}\right)^{t-1} f(x_0) \geq f(\hat{x}_{t-1}), \quad t \geq 1.$$

Thus, in view of Item 1 of Theorem 7.1.1, at the last stage $T$ of the process we have

$$\left(\frac{1}{e}\right)^{T-1} f(x_0) \geq f(\hat{x}_{T-1}) \;\geq\; f^* \;\geq\; \alpha(F) f(x_0).$$

This leads to inequality (7.1.31).

Note that $\|x_0 - x^*\| \le \frac{1}{\gamma_0(F)} f^* \le \frac{1}{\gamma_0(F)} f(\hat{x}_{T-1})$. Therefore, at the last stage of the process in view of inequality (7.1.28) and the termination rule in (7.1.30) we have

$$\begin{aligned} f(\hat{x}_T) - f^* &\le \tfrac{2\gamma_1(F)}{\tilde{N}+1} \cdot \tfrac{1}{\gamma_0(F)} \cdot f(\hat{x}_{T-1}) \;\le\; \tfrac{2e}{\alpha(F)\cdot(\tilde{N}+1)} \cdot f(\hat{x}_T) \\ &\le \tfrac{\delta}{1+\delta} \cdot f(\hat{x}_T). \end{aligned} \qquad \square$$

### 7.1.4 Application Examples

In this section, we discuss the complexity of implementation of the schemes presented in Sect. 7.1.3 as applied to different structural classes of optimization problems.

#### 7.1.4.1 Linear Programming

Let $\hat{A}$ be an $m \times (n-1)$-matrix, $m \ge n$, which has a full column rank. For a given vector $c \in \mathbb{R}^m$, consider the following optimization problem:

$$\text{Find } f^* = \max_{u \in \mathbb{R}^m} \left\{ \langle c, u \rangle : \; \hat{A}^T u = 0, \; |u^{(i)}| \le 1, \; i = 1, \dots, m \right\}. \tag{7.1.33}$$

This problem is non-trivial only if the column rank of matrix $A = (\hat{A}, c)$ is equal to $n$, which we assume to be true.

Problem (7.1.33) can be rewritten in the adjoint form. Define

$$\phi_1(y) = \max_{u \in \mathbb{R}^m} \left\{ \langle c, u \rangle + \langle y, \hat{A}^T u \rangle : \; |u^{(i)}| \le 1, \; i = 1, \dots, m \right\} = \sum_{i=1}^{m} |\langle a_i, y \rangle + c_i|,$$

where $a_i$ are the columns of the matrix $\hat{A}^T$. Then

$$f^* = \min_{y \in \mathbb{R}^{n-1}} \phi_1(y).$$

In Example 7.1.1 we have already seen that the latter minimization problem can be represented in the form (7.1.19)–(7.1.20) with $x = (y^T, \tau)^T$, and $F_1(v) = \sum\limits_{i=1}^{m} |v^{(i)}|$. Thus,

$$Q_2 = \{u \in \mathbb{R}^m : \; |u^{(i)}| \le 1, \; i = 1, \dots, m\}.$$

Choosing $\|u\|_{(2)} = \left[\sum\limits_{i=1}^{m} (u^{(i)})^2\right]^{1/2}$, we get

$$\gamma_0(F_\infty) = 1, \quad \gamma_1(F_\infty) = \sqrt{m}, \quad \alpha(F_\infty) = \tfrac{1}{\sqrt{m}}.$$

Therefore, in view of Theorem 7.1.5, in order to estimate $f^*$ with relative accuracy $\delta \in (0, 1)$ we need at most

$$2e \cdot m^{1/2} \cdot \left(1 + \tfrac{1}{2}\ln m\right) \cdot \left(1 + \tfrac{1}{\delta}\right)$$

iterations of the scheme $S_N(R)$.

For this method, we need to compute and invert the matrix $G = A^T A$. If $A$ is dense, this takes $O(n^2 m)$ operations. Further, each iteration of the scheme $S_N(R)$ requires $O(nm)$ operations:

- Multiplication of matrix $A$ by $y_k$ takes $O(mn)$ operations.
- Since the set $Q_2$ and the norm $\|u\|_{\mathbb{R}^m}$ have separable structure, computation of $u_\mu(x_k)$ needs $O(m)$ operations.
- Computation of $v_{k+1}$ needs one multiplication of $A^T$ by a vector, and finding the projection onto a set with representation

$$Q_1(R) = \{x \in \mathbb{R}^n : \ Cx = 1, \ \|x\|_{\mathbb{R}^n} \leq R\}$$

  in the Euclidean metric $\|\cdot\|_{\mathbb{R}^n}$. Since $C \in \mathbb{R}^{1\times n}$, such a projection can be found by a closed-form expression.

Thus, the total amount of computations in the scheme is of the order of

$$O\left(n^2 m + \tfrac{1}{\delta} \cdot n m^{1.5} \ln m\right) \tag{7.1.34}$$

operations. The first ingredient of this estimate is dominant when $\delta > \frac{\sqrt{m}}{n} \ln m$.

Note that for problem (7.1.33) we can apply a standard short-step path-following scheme (5.3.25). Each iteration of this scheme needs $O(n^2 m)$ operations. Therefore its worst-case efficiency estimate is as follows:

$$O\left(n^2 m^{1.5} \ln \tfrac{m}{\delta}\right). \tag{7.1.35}$$

Another possibility is to solve this problem by the ellipsoid method (3.2.53). In this case, the total complexity of its solution is

$$O\left(n^3 m \ln \tfrac{m}{\delta}\right). \tag{7.1.36}$$

Comparing the bounds (7.1.34), (7.1.35), and (7.1.36), we conclude that the scheme (7.1.30) is the best when $\delta$ is not too small, say

$$\delta > O\left(\frac{1}{n}\max\left\{1, \frac{\sqrt{m}}{n}\right\}\right).$$

#### 7.1.4.2 Minimization of the Spectral Radius

Denote by $\mathbb{S}^n$ the space of symmetric $n \times n$-matrices. For $X \in \mathbb{S}^n$, we can define its spectral radius:

$$\rho(X) = \max_{1 \le i \le n} |\lambda_i(X)|.$$

Note that this function is convex on $\mathbb{S}^n$. For a vector of decision variables $x \in \mathbb{R}^p$, let us introduce a linear operator $A(x)$:

$$A(x) = \sum_{i=1}^{p} x^{(i)} A_i \in \mathbb{S}^n.$$

Now we can define the following objective function in problem (7.1.20):

$$f(x) = \rho(A(x)). \tag{7.1.37}$$

Assume also that the constraints in problem (7.1.20), (7.1.37) are linear and very simple. For example, it could be $x^{(1)} = 1$.

In order to treat the problem (7.1.20), (7.1.37) we need to represent the upper-level function $\rho(X)$ in a special form (7.1.18). Let

$$Q_2 = \left\{X \in \mathbb{S}^n : \sum_{i=1}^{n} |\lambda_i(X)| \le 1\right\}.$$

Let us endow the space $\mathbb{S}^n$ with the standard Frobenius norm:

$$\|X\|_F = \langle X, X\rangle_F^{1/2}, \quad \langle X, Y\rangle_F \overset{\text{def}}{=} \sum_{i,j=1}^{n} X^{(i,j)} Y^{(i,j)}, \quad X, Y \in \mathbb{S}^n.$$

**Lemma 7.1.3** *Let $Q_2$ be a closed convex set such that*

$$B_{\|\cdot\|_F}\left(\frac{1}{\sqrt{n}}\right) \subset Q_2 \subset B_{\|\cdot\|_F}(1). \tag{7.1.38}$$

*Moreover,* $\rho(X) = \max\limits_{U \in Q_2} \langle X, U\rangle$.

*Proof* For any $X \in S^n$, we have:

$$\begin{aligned}\rho(X) &= \min_{\tau\in\mathbb{R}}\{\tau :\ \tau I_n \succeq X,\ \tau I_n \succeq -X\}\\ &= \min_{\tau\in\mathbb{R}}\max_{Y_1,Y_2\succeq 0}[\tau + \langle X-\tau I_n, Y_1\rangle_F - \langle X+\tau I_n, Y_2\rangle_F]\\ &= \max_{Y_1,Y_2\succeq 0}\{\langle X, Y_1-Y_2\rangle_F :\ \langle I_n, Y_1+Y_2\rangle_F = 1\}.\end{aligned}$$

Let $U = Y_1 - Y_2$ and $V = Y_1 + Y_2$. Then

$$\rho(X) = \max_{U\in S^n}\{\langle X, U\rangle_F,\ U \in \hat{Q}\},$$

where $\hat{Q} = \{U :\ \exists V \succeq \pm U,\ \langle I_n, V\rangle_F = 1\}$. It is clear that the set $\hat{Q}$ is closed, convex and bounded. Let us prove that $\hat{Q} = Q_2$.

Indeed, we can always represent $U$ by its orthogonal basis of eigenvectors:

$$U = B\Lambda B^T, \quad BB^T = I_n,$$

where $\Lambda$ is a diagonal matrix. Assume that $U \in Q_2$. Define a diagonal matrix $\hat{\Lambda}$ with the following diagonal entries:

$$\hat{\Lambda}^{(i,i)} = |\Lambda^{(i,i)}|/[\sum_{j=1}^n |\Lambda^{(j,j)}|], \quad i = 1,\dots,n.$$

Then $V = B\hat{\Lambda}B^T \succeq \pm U$ and $\langle I_n, V\rangle_F = 1$. Thus $Q_2 \subseteq \hat{Q}$.

Conversely, if $U \in \hat{Q}$, then there exists a $V \in S^n$ such that $B^T V B \succeq \pm\Lambda$. Therefore

$$\langle Vb_i, b_i\rangle_F \geq |\Lambda^{(i,i)}|, \quad i = 1,\dots,n,$$

where $b_i$ are the columns of the matrix $B$. Hence,

$$1 = \langle I_n, V\rangle_F = \langle BB^T, V\rangle_F = \langle I_n, B^T V B\rangle_F = \sum_{i=1}^n \langle Vb_i, b_i\rangle_F \geq \sum_{i=1}^n |\lambda_i(U)|.$$

Thus, $\hat{Q} \subseteq Q_2$ and we conclude that $\hat{Q} = Q_2$.

It remains to prove inclusion (7.1.38). Indeed, if $\|U\|_F^2 \leq \frac{1}{n}$, that is $\sum_{i=1}^n \lambda_i^2(U) \leq \frac{1}{n}$, then

$$\sum_{i=1}^n |\lambda_i(U)| \leq \sqrt{n}\cdot\left[\sum_{i=1}^n \lambda_i^2(U)\right]^{1/2} \leq 1.$$

Conversely, if $\sum\limits_{i=1}^{n} |\lambda_i(U)| \leq 1$, then $\sum\limits_{i=1}^{n} \lambda_i^2(U) \leq \left[ \sum\limits_{i=1}^{n} |\lambda_i(U)| \right]^2 \leq 1$. □

Thus, in view of inclusion (7.1.38) we have

$$\gamma_0(\rho) = \tfrac{1}{\sqrt{n}}, \quad \gamma_1(\rho) = 1, \quad \alpha(\rho) = \tfrac{1}{\sqrt{n}}.$$

Hence, in view of Theorem 7.1.5, the total number of iterations of the method $S_N(R)$ does not exceed

$$2e\sqrt{n}\left(1 + \tfrac{1}{2}\ln n\right) \cdot \left(1 + \tfrac{1}{\delta}\right).$$

In order to apply this approach, we need to compute and invert the matrix $G$. In our situation, $G$ is the matrix of the following quadratic form:

$$\langle Gx, x\rangle = \langle A(x), A(x)\rangle_F.$$

Thus, $G^{(i,j)} = \langle A_i, A_j\rangle_F$, $i, j = 1, \ldots, p$. If the matrices $A_i$ are dense, the computation of this matrix takes $O(p^2n^2)$ arithmetic operations and the inversion takes $O(p^3)$ operations. Since we assume $p < \frac{n(n+1)}{2}$, the total cost of the preliminary computation is of the order of $O(p^2n^2)$ operations.

Further, the most expensive operations at each step of the method $S_N(R)$ are as follows.

- Computation of the value of the bilinear form $\langle A(x), U\rangle_F$ and its gradients takes $O(pn^2)$ operations.
- Finding a projection of point $X$ onto the set $Q_2$ with respect to the standard Frobenius norm. The most expensive part of this operation consists in solving an eigenvalue problem for the matrix $X$. This can be done in $O(n^3)$ operations.
- The total amount of operations in the space $\mathbb{R}^p$ does not exceed $O(p^2)$.

Thus, the complexity of each iteration of $S_N(R)$ is of the order of $O(n^2(n+p))$ operations. Hence, in total, the method (7.1.30) requires

$$O\left(n^2p^2 + \tfrac{1}{\delta} \cdot n^{2.5}(p+n)\ln n\right) \tag{7.1.39}$$

arithmetic operations.

Let us compare this estimate with the worst-case complexity of a short-step path-following scheme as applied to the problem (7.1.20)–(7.1.37). For this method, the most expensive computations at each iteration are the computations of the elements of the Hessian of the barrier function. In accordance with Lemma 5.4.6, these are the values

$$\langle X^{-1}A_iX^{-1}, A_j\rangle_F, \quad i, j = 1, \ldots, p.$$

Such a computation needs $O(pn^2(p+n))$ operations. Thus, the total complexity of the interior-point method is of the order of

$$O\left(pn^{2.5}(p+n)\ln\frac{n}{\delta}\right)$$

operations. Comparing this estimate with (7.1.39) we see that the gradient method is better if the required relative accuracy is not too small:

$$\delta \geq O\left(\tfrac{1}{p}\right).$$

#### 7.1.4.3 The Truss Topology Design Problem

In this problem, we have a set of points

$$x_i \in \mathbb{R}^2, \quad i = 1, \dots, n+p,$$

connected by a set of arcs $(i_k, j_k)$, $k = 1, \dots, m$. We always assume that $j_k > i_k$. Each arc has a nonnegative weight $t^{(k)}$, and the sum of all weights is equal to one. The nodes $x_{n+1}, \dots, x_{n+p}$ are fixed. To all other nodes we can apply external forces

$$f_i \in \mathbb{R}^2, \quad i = 1, \dots, n, \quad f \overset{\text{def}}{=} (f_1, \dots, f_n)^T \in \mathbb{R}^{2n}.$$

The goal is to find an optimal design vector

$$t \overset{\text{def}}{=} (t^{(1)}, \dots, t^{(m)})^T \in \Delta_m \equiv \left\{ t \in \mathbb{R}^m_+ : \sum_{i=1}^m t^{(i)} = 1 \right\}$$

which minimizes the total *stiffness* $\psi(t)$ of the system.

To define the stiffness, we can always assume that $i_k < n$, $k = 1, \dots, m$, allowing no arcs between fixed nodes. For each arc $k$, define vectors

$$d_k = \tfrac{x_{i_k} - x_{j_k}}{\|x_{i_k} - x_{j_k}\|^2} \in \mathbb{R}^2, \quad k = 1, \dots, m,$$

where $\| \cdot \|$ is the standard Euclidean norm in $\mathbb{R}^2$. Now we can define the constraint vector $a_k = (a_{k,1}, \dots, a_{k,n})^T \in \mathbb{R}^{2n}$, which is composed by the following two-dimensional vectors:

$$a_{k,q} = \begin{cases} d_k, \text{ if } q = i_k, \\ -d_k, \text{ if } q = j_k \text{ and } j_k \leq n, \\ 0, \text{ otherwise.} \end{cases} \quad q = 1, \dots, n.$$

Let $B(t) = \sum\limits_{k=1}^{m} t^{(k)} a_k a_k^T$. Then the *truss topology design* problem can be written as follows

$$\text{Find } \psi^* = \inf_t \{ \langle [B(t)]^{-1} f, f \rangle : \ t \in \text{rint}\, \Delta_m \}. \tag{7.1.40}$$

This problem is well defined if and only if the matrix $G \overset{\text{def}}{=} B(\bar{e}_m)$ is positive definite.

Let us show how this problem can be rewritten in the form (7.1.19)–(7.1.20).

$$\begin{aligned}
\psi^* &= \inf_{t \in \text{rint}\, \Delta_m} \langle [B(t)]^{-1} f, f \rangle \\
&= \inf_{t \in \text{rint}\, \Delta_m} \max_{x \in \mathbb{R}^{2n}} [2\langle f, x \rangle - \langle B(t)x, x \rangle] \\
&= \max_{x \in \mathbb{R}^{2n}} \inf_{t \in \text{rint}\, \Delta_m} \left[ 2\langle f, x \rangle - \sum_{k=1}^{m} t^{(k)} \langle a_k, x \rangle^2 \right] \\
&= \max_{x \in \mathbb{R}^{2n}} \left[ 2\langle f, x \rangle - \max_{1 \le k \le m} \langle a_k, x \rangle^2 \right] \\
&= \max_{x \in \mathbb{R}^{2n}} \frac{\langle f, x \rangle^2}{\max\limits_{1 \le k \le m} \langle a_k, x \rangle^2}
\end{aligned}$$

(in the last step we perform a maximization of the objective function along direction $x$ by multiplying it by a positive factor).

Thus, we can consider the problem

$$\text{Find } f^* = \min_{x \in \mathbb{R}^{2n}} \{ f(x) \overset{\text{def}}{=} \max_{1 \le k \le m} |\langle a_k, x \rangle| : \ \langle f, x \rangle = 1 \}, \tag{7.1.41}$$

which is exactly in the desired form (7.1.19)–(7.1.20). Let $A$ be an $m \times (2n)$-matrix with the rows $a_k^T$. Then, using the notation of Example 7.1.1 the objective function of this problem can be written as

$$f(x) = F_\infty(Ax).$$

In view of (7.1.21) we have $\alpha(F_\infty) = \frac{1}{\sqrt{m}}$. Therefore, in order to find an approximate solution to (7.1.41) with relative accuracy $\delta$, the method (7.1.30) needs at most

$$2e\sqrt{m} \left( 1 + \tfrac{1}{2} \ln m \right) \cdot \left( 1 + \tfrac{1}{\delta} \right) \tag{7.1.42}$$

iterations of the scheme $S_N(R)$. The most expensive operations of each iteration of the latter scheme are as follows.

- Computation of the value and the gradients of the bilinear form $\langle Ax, u\rangle$ needs $O(m)$ operations (recall that $A$ is sparse).
- Euclidean projection on $Q_2 \subset \mathbb{R}^m$ needs $O(m \ln m)$ operations.
- All steps in the primal space need $O(n^2)$ operations.

Note that the preliminary computation of the matrix $G$ needs $O(m + n^2)$ operations, but its inversion costs $O(n^3)$. Since $m \leq \frac{n(n+1)}{2}$, we come to the following upper bound for the total computational effort of the method (7.1.30):

$$O\left(n^3 + \tfrac{1}{\delta} \cdot (n^2 + m \ln m) \cdot \sqrt{m} \ln m\right) \tag{7.1.43}$$

arithmetic operations. For a dense truss with $m = O(n^2)$ this estimates becomes

$$O\left(\tfrac{n^3}{\delta} \ln^2 n\right)$$

arithmetic operations.

## 7.2 Rounding of Convex Sets

(Computing rounding ellipsoids; John's Theorem; Rounding by diagonal ellipsoids; Minimizing the maximal absolute value of linear functions; Bilinear matrix games with non-negative coefficients; Minimizing the spectral radius of symmetric matrices.)

### *7.2.1 Computing Rounding Ellipsoids*

Among modern methods for solving problems of Linear Programming (LP-problems, for short), the Interior-Point Methods (IPM) are considered to be the most efficient. However, these methods are based on an expensive machinery. For an LP-problem with $n$ variables and $m$ inequality constraints, $(m > n)$, in order to get an approximate solution with *absolute* accuracy $\epsilon$, these methods need to perform

$$O(\sqrt{m} \ln \tfrac{m}{\epsilon})$$

iterations of Newton's Method (see Sect. 5.4). Recall that for problems with dense data, each iteration can take up to $O(n^2 m)$ operations.

Clearly these bounds leave considerable room for competition with gradient-type methods, for which each iteration is much cheaper. However, the main drawback of the latter schemes is their relatively slow convergence. In general, the gradient schemes need $O\left(\frac{C_0}{\epsilon^2}\right)$ iterations in order to find an $\epsilon$-solution to the problem (see Sect. 3.2). In this estimate, a strong dependence on $\epsilon$ is coupled with the presence of a constant $C_0$, which depends on the norm of the matrix of constraints, the size of the solution, etc, and which can be uncontrollably large. Consequently, the classical gradient-type schemes can compete with IPM only on very large problems.

However, in Chap. 6 we have shown that it is possible to use the special structure of LP-problems in order to get gradient-type schemes which converge in $O\left(\frac{C_1}{\epsilon}\right)$ iterations. Moreover, it was shown that, for some LP-problems, the constant $C_1$ can be found explicitly and that it is reasonably small. In Sect. 7.1 this result was extended to cover minimization schemes for finding an approximate solution with a certain *relative* accuracy. Namely, it was shown that for some classes of LP-problems it is possible to compute an approximate solution of relative accuracy $\delta$ with $O(\frac{\sqrt{m}}{\delta})$ iterations of a gradient-type scheme. Recall that for many applications the concept of relative accuracy is very attractive since it adapts automatically to *any* size of the solution. So, there is no necessity to fight against big and unknown constants. For many problems in Economics and Engineering, the level of relative accuracy of the order 1.5–0.05% is completely acceptable.

The approach of Sect. 7.1 is applicable to special *conic* unconstrained minimization problems. They consist in minimization of a non-negative positively homogeneous convex function $f$, dom $f = \mathbb{R}^n$, on a closed convex set separated from zero. In order to compute a solution to this problem with some relative accuracy, we need to know a rounding ellipsoid for the subdifferential of $f$ at the origin. It was shown that for some LP-problems it is possible to use the structure of the objective function in order to compute such an ellipsoid with radius $O\left(\sqrt{m}\right)$.

It is well known that, for any centrally symmetric set in $\mathbb{R}^n$, there exists a $\sqrt{n}$-rounding ellipsoid. Moreover, a good approximation to such an ellipsoid can be easily computed. It appears that this ellipsoid provides us with a good norm, allowing us to solve the corresponding minimization problem up to a certain relative accuracy. In this section, we analyze two non-trivial classes of LP-problems and show that for both classes the approximate solutions with relative accuracy $\delta$ can be computed in $O\left(\frac{\sqrt{n \ln m}}{\delta} \ln n\right)$ iterations of a gradient-type method.

At the same time, the preliminary computation of the rounding ellipsoids in both situations is reasonably cheap: it takes $O(n^2 m \ln m)$ operations at most. Up to a logarithmic factor, this estimate coincides with the complexity of finding a projection onto a linear subspace in $\mathbb{R}^m$ defined by $n$ linear equations. However, we will see that the consequent optimization process is even cheaper.

Let us recall some notation. In this section, it is convenient to identify $\mathbb{E}$ and $\mathbb{E}^*$ with $\mathbb{R}^n$. A symmetric $n \times n$-matrix $G \succ 0$ defines a norm on $\mathbb{R}^n$:

$$\|x\|_G = \langle Gx, x\rangle^{1/2}, \quad x \in \mathbb{R}^n.$$

The dual norm is defined in the usual way:

$$\|s\|_G^* = \sup_x \{\langle s, x\rangle : \|x\|_G \leq 1\} = \langle s, G^{-1}s\rangle^{1/2}, \quad s \in \mathbb{R}^n.$$

For a closed convex bounded set $C \subset \mathbb{R}^n$, $\xi_C(x)$ denotes its support function:

$$\xi_C(x) = \max_{s \in C} \langle s, x\rangle, \quad x \in \mathbb{R}^n.$$

Thus $\partial \xi_C(0) = C$.

Finally, $D(a)$ denotes a diagonal $n \times n$-matrix with vector $a \in \mathbb{R}^n$ at the diagonal. In this setting, $e_k \in \mathbb{R}^n$ denotes the $k$th coordinate vector, and $\bar{e}_n \in \mathbb{R}^n$ denotes the vector of all ones. Thus, $I_n \equiv D(\bar{e}_n)$. As before, the notation $\mathbb{R}^n_+$ is used for the positive orthant and $\Delta_n \equiv \{x \in \mathbb{R}^n_+ : \langle \bar{e}_n, x\rangle = 1\}$ denotes the standard simplex in $\mathbb{R}^n$.

In this section, we analyze efficient algorithms for constructing rounding ellipsoids for different types of convex sets. An ellipsoid $W_r(v, G) \subset \mathbb{R}^n$ is usually represented in the following form:

$$W_r(v, G) = \{s \in \mathbb{R}^n : \|s - v\|_G^* \equiv \langle s - v, G^{-1}(s - v)\rangle^{1/2} \leq r\},$$

where $G \succ 0$ is a symmetric $n \times n$-matrix. If $v = 0$, we often use the notation $W_r(G)$. An ellipsoid $W_1(v, G)$ is called a *$\beta$-rounding* for a convex set $C \subset \mathbb{R}^n$, $\beta \geq 1$, if

$$W_1(v, G) \subseteq C \subseteq W_\beta(v, G).$$

We call $\beta$ the *radius* of ellipsoidal rounding.

#### 7.2.1.1 Convex Sets with Central Symmetry

Let $G \succ 0$. For an arbitrary $g \in \mathbb{R}^n$, consider the set $C_{\pm g}(G) = \mathrm{Conv}\,\{W_1(G), \pm g\}$. For $\alpha \in [0, 1]$ define

$$G(\alpha) = (1 - \alpha)G + \alpha g g^T.$$

**Lemma 7.2.1** *For any $\alpha \in [0, 1)$, the following inclusion holds:*

$$W_1(G(\alpha)) \subset C_{\pm g}(G). \tag{7.2.1}$$

*If the value $\sigma \overset{\mathrm{def}}{=} \frac{1}{n}(\|g\|_G^*)^2 - 1$ is positive, then the function*

$$V(\alpha) \overset{\mathrm{def}}{=} \ln \frac{\det G(\alpha)}{\det G(0)} = \ln(1 + \alpha(n(1 + \sigma) - 1)) + (n - 1)\ln(1 - \alpha),$$

*attains its maximum at* $\alpha^* = \frac{\sigma}{n(1+\sigma)-1}$. *Moreover,*

$$\begin{aligned} V(\alpha^*) &= \ln(1+\sigma) + (n-1)\ln\tfrac{(n-1)(1+\sigma)}{n(1+\sigma)-1} \\ &\geq \ln(1+\sigma) - \tfrac{\sigma}{1+\sigma} \;\geq\; \tfrac{\sigma^2}{(1+\sigma)(2+\sigma)}. \end{aligned} \tag{7.2.2}$$

*Proof* For any $x \in \mathbb{R}^n$, we have

$$\begin{aligned} \xi_{W_1(G(\alpha))}(x) = \langle G(\alpha)x, x\rangle^{1/2} &= [(1-\alpha)\langle Gx, x\rangle + \alpha\langle g, x\rangle^2]^{1/2} \\ &\leq \max\{\langle Gx, x\rangle^{1/2}, |\langle g, x\rangle|\} \\ &= \max\{\xi_{W_1(G)}(x), \xi_{\text{Conv }\{\pm g\}}(x)\} \;=\; \xi_{C_{\pm g}(G)}(x). \end{aligned}$$

Hence, in view of Corollary 3.1.5, inclusion (7.2.1) is proved.

Furthermore,

$$\begin{aligned} V(\alpha) &= \ln\det(G^{-1/2}G(\alpha)G^{-1/2}) \\ &= \ln\det\big((1-\alpha)I_n + \alpha G^{-1/2}gg^T G^{-1/2}\big) \\ &= \ln\big(1-\alpha+\alpha(\|g\|_G^*)^2\big) + (n-1)\ln(1-\alpha) \\ &= \ln\left(1+\alpha(n(1+\sigma)-1)\right) + (n-1)\ln(1-\alpha). \end{aligned}$$

Hence, in view of Theorem 2.1.1, the global optimality condition for the function $V(\cdot)$ is as follows:

$$\tfrac{n-1}{1-\alpha} = \tfrac{n(1+\sigma)-1}{1+\alpha(n(1+\sigma)-1)}.$$

The only solution of this equation is $\alpha^* = \frac{\sigma}{n(1+\sigma)-1}$. Note that

$$\begin{aligned} V(\alpha^*) &= \ln(1+\sigma) + (n-1)\ln\tfrac{(n-1)(1+\sigma)}{n(1+\sigma)-1} \\ &= \ln(1+\sigma) - (n-1)\ln\left(1+\tfrac{\sigma}{(n-1)(1+\sigma)}\right) \\ &\geq \ln(1+\sigma) - \tfrac{\sigma}{1+\sigma} \;=\; \tfrac{\sigma^2}{1+\sigma} - \omega(\sigma) \\ &\overset{(5.1.23)}{\geq} \tfrac{\sigma^2}{(1+\sigma)(2+\sigma)}. \end{aligned}$$

□

In this section, we are interested in solving the following problem. Let $C$ be a convex centrally symmetric body, i.e. int $C \neq \emptyset$, and $x \in C \Leftrightarrow -x \in C$. For a given $\gamma > 1$, we need to find an ellipsoidal rounding for $C$ of radius $\gamma\sqrt{n}$. An initial approximation to the solution of our problem is given by a matrix $G_0 \succ 0$ such that $W_1(G_0) \subseteq C$, and $C \subseteq W_R(G_0)$ for a certain $R \geq 1$.

Let us look at a particular variant of such a problem.

*Example 7.2.1* Consider a set of vectors $a_i \in \mathbb{R}^n$, $i = 1, \ldots, m$, which span the whole space $\mathbb{R}^n$. Let the set $C$ be defined as follows:

$$C = \text{Conv }\{\pm a_i,\ i = 1, \ldots, m\}. \tag{7.2.3}$$

We choose $G_0 = \frac{1}{m}\sum\limits_{i=1}^{m} a_i a_i^T$. Note that for any $x \in \mathbb{R}^n$, we have $\xi_C(x) = \max\limits_{1 \leq i \leq m} |\langle a_i, x\rangle|$. Therefore,

$$\xi_{W_1(G_0)}(x) = \left[\frac{1}{m}\sum_{i=1}^{m}\langle a_i, x\rangle^2\right]^{1/2} \leq \xi_C(x),$$

$$\xi_{W_{\sqrt{m}}(G_0)}(x) = m^{1/2}\left[\frac{1}{m}\sum_{i=1}^{m}\langle a_i, x\rangle^2\right]^{1/2} \geq \xi_C(x).$$

Thus, in view of Corollary 3.1.5, $W_1(G_0) \subseteq C \subseteq W_{\sqrt{m}}(G_0)$. □

Let us analyze the following algorithmic scheme.

```
For $k \geq 0$ iterate:

1. Compute $g_k \in C :\ \|g_k\|^*_{G_k} = r_k \overset{\text{def}}{=} \max\limits_{g}\{\|g\|^*_{G_k} :\ g \in C\}$.

2. If $r_k \leq \gamma n^{1/2}$ then Stop else                                  (7.2.4)

       $\alpha_k = \frac{r_k^2 - n}{n(r_k^2 - 1)}, \quad G_{k+1} = (1 - \alpha_k)G_k + \alpha_k g_k g_k^T.$

end.
```

The complexity bound for this scheme is given by the following statement.

**Theorem 7.2.1** *Let $R \geq 1$ and $W_1(G_0) \subseteq C \subseteq W_R(G_0)$. Then scheme (7.2.4) terminates after*

$$2n \frac{\gamma^2}{(\gamma-1)^2} \ln R \tag{7.2.5}$$

*iterations at most.*

*Proof* Note that the coefficient $\alpha_k$ in Step 2 of (7.2.4) is chosen in accordance with Lemma 7.2.1. Since the method runs as long as $\sigma_k \overset{\mathrm{def}}{=} \frac{1}{n} r_k^2 - 1 \geq \gamma^2 - 1$, in view of inequality (7.2.2), at each step $k \geq 0$ we have

$$\ln \det G_{k+1} \geq \ln \det G_k + 2 \ln \gamma - \frac{\gamma^2-1}{\gamma^2}. \tag{7.2.6}$$

Note that

$$2 \ln \gamma - \frac{\gamma^2-1}{\gamma^2} = \frac{(\gamma^2-1)^2}{\gamma^2} - \omega(\gamma^2 - 1) \overset{(5.1.23)}{\geq} \frac{(\gamma^2-1)^2}{\gamma^2} - \frac{(\gamma^2-1)^2}{1+\gamma^2}$$

$$= \frac{(\gamma^2-1)^2}{\gamma^2(1+\gamma^2)} \geq \frac{1}{\gamma^2}(\gamma - 1)^2.$$

At the same time, for any $k \geq 0$ we get

$$\det(G_k)^{1/2} \cdot \mathrm{vol}_n\,(W_1(I_n)) = \mathrm{vol}_n\,(W_1(G_k)) \leq \mathrm{vol}_n\,(C) \leq \mathrm{vol}_n\,(W_R(G_0))$$

$$= R^n \cdot \det(G_0)^{1/2} \cdot \mathrm{vol}_n\,(W_1(I_n)).$$

Hence, $\ln \det G_k - \ln \det G_0 \leq 2n \ln R$, and we get bound (7.2.5) by summing up inequalities (7.2.6). □

Let us estimate the total arithmetical complexity of the scheme (7.2.4) as applied to a particular symmetric convex set (7.2.3). In this situation, it is reasonable to recursively update the inverse matrices $H_k \overset{\mathrm{def}}{=} G_k^{-1}$, and the set of values

$$\nu_k^{(i)} = \langle a_i, H_k a_i \rangle, \quad i = 1, \ldots, m,$$

which we treat as a vector $\nu_k \in \mathbb{R}^m$. A modified variant of the scheme (7.2.4) is as follows.

```
A. Compute $H_0 = \left[\frac{1}{m}\sum\limits_{i=1}^{m} a_i a_i^T\right]^{-1}$ and the vector $\nu_0 \in \mathbb{R}^m$.

B. For $k \geq 0$ iterate:

    1. Find $i_k$ :  $\nu_k^{(i_k)} = \max\limits_{1 \leq i \leq m} \nu_k^{(i)}$. Set $r_k = [\nu^{(i_k)}]^{1/2}$.

    2. If $r_k \leq \gamma n^{1/2}$ then Stop else

        2.1. Set $\sigma_k = \frac{1}{n} r_k^2 - 1$, $\alpha_k = \frac{\sigma_k}{r_k^2 - 1}$, $x_k = H_k a_{i_k}$.

        2.2. Update $H_{k+1} := \frac{1}{1-\alpha_k}\left[H_k - \frac{\alpha_k}{1+\sigma_k} \cdot x_k x_k^T\right]$.

        2.3. Update $\nu_{k+1}^{(i)} := \frac{1}{1-\alpha_k}\left[\nu_k^{(i)} - \frac{\alpha_k}{1+\sigma_k} \cdot \langle a_i, x_k \rangle^2\right]$,
                 $i = 1, \dots, m$.

  end.
```

(7.2.7)

Let us estimate the arithmetical complexity of this scheme. For simplicity, we assume that the matrix $A = (a_1, \dots, a_m)$ is dense. We write down only the leading polynomial terms in the complexity of the corresponding computations, where we count only multiplications.

- **Phase A** takes $\frac{mn^2}{2}$ operations to compute the matrix $G_0$, plus $\frac{n^3}{6}$ operations to compute its inverse, and $\frac{mn^2}{2}$ operations to compute the vector $\nu_0$.
- **Step 2.1** takes $n^2$ operations.
- **Step 2.2** takes $\frac{n^2}{2}$ operations.
- **Step 2.3** takes $mn$ operations.

Using now the estimate (7.2.5) with $R = \sqrt{m}$ (see Example 7.2.1), we conclude that for $\gamma > 1$ and the centrally symmetric set (7.2.3), the scheme (7.2.7) can find a $\gamma\sqrt{n}$-rounding in

$$\frac{n^2}{6}(n + 6m) + \frac{\gamma^2}{(\gamma-1)^2} n^2 (2m + 3n) \ln m$$

arithmetic operations. Note that for a sparse matrix $A$ the complexity of **Phase A** and **Step 2.3** will be much lower.

*Remark 7.2.1* Note that the process (7.2.4) with eliminated stopping criterion can be used to prove a symmetric version of John's theorem.

Indeed, all matrices generated by this process have the following form:

$$G_k = \sum_{i=1}^{m} \lambda_k^{(i)} a_i a_i^T, \quad \lambda_k \in \mathbb{R}^m_+, \quad \sum_{i=1}^{m} \lambda_k^{(i)} = 1.$$

Therefore, $I_n = \sum\limits_{i=1}^{m} \lambda_k^{(i)} G_k^{-1/2} a_i a_i^T G_k^{-1/2}$. Taking the trace of both sides of this equality, we get

$$n = \sum_{i=1}^{m} \lambda_k^{(i)} (\|a_i\|^*_{G_k})^2 \;\le\; r_k^2.$$

On the other hand, we have seen that

$$\ln\det G_{k+1} \overset{(7.2.6)}{\ge} \ln\det G_k + \ln(1+\sigma_k) - \tfrac{\sigma_k}{1+\sigma_k} \overset{(5.1.23)}{\ge} \tfrac{1}{r_k^2}(r_k - \sqrt{n})^2.$$

Therefore, by the same reasoning as in the proof of Theorem 7.2.1, after $N$ iterations of the scheme we get

$$\sum_{k=0}^{N} \left(1 - \tfrac{\sqrt{n}}{r_k}\right)^2 \le 2n \ln R.$$

Defining $r_N^* = \min\limits_{0 \le k \le N} r_k$, we have $\frac{\sqrt{n}}{r_N^*} \ge 1 - \left(\frac{2n}{N+1} \ln R\right)^{1/2}$. Thus, $r_N^* \to \sqrt{n}$ as $N \to \infty$. Since the sequence of matrices $\{G_k\}$ is compact, we conclude that there exists a limiting matrix $G_*$ with rounding coefficient $\beta = \sqrt{n}$.

Thus, we have proved a symmetric version of John's Theorem for the set $\mathscr{C}$ defined by (7.2.3). Since the quality of our rounding does not depend on the number of points $m$, we can use the fact that any general symmetric convex set can be approximated by a convex combination of finite number of points with arbitrary accuracy. Thus, our statement is also valid for general sets.

Note that the process (7.2.4) always constructs a matrix with rounding coefficient $\beta = \sqrt{n}$. Of course, there exist symmetric sets with much better rounding. It will be interesting to develop an efficient procedure which can adjust to the exact rounding coefficient for a particular convex set. □

#### 7.2.1.2 General Convex Sets

For an arbitrary $g$ from $\mathbb{R}^n$, consider the set $C_g(G) = \text{Conv}\ \{W_1(G), g\}$. In view of Lemma 3.1.3 support function of this set is as follows:

$$\xi_{C_g(G)}(x) = \max\{\|x\|_G, \langle g, x\rangle\}, \quad x \in \mathbb{R}^n.$$

Define $r = \|g\|^*_G$, and

$$G(\alpha) = (1-\alpha)G + \left(\tfrac{\alpha}{r} + \left(\tfrac{r-1}{2}\right)^2 \cdot \left(\tfrac{\alpha}{r}\right)^2\right) \cdot gg^T, \quad \alpha \in [0, 1].$$

**Lemma 7.2.2** *For all* $\alpha \in [0, 1)$*, the ellipsoid*

$$E_\alpha = \{s \in \mathbb{R}^n : \|s - \tfrac{r-1}{2r} \cdot \alpha g\|^*_{G(\alpha)} \le 1\}$$

*belongs to the set* $C_g(G)$*. If* $r \ge n$*, then the function*

$$V(\alpha) \overset{\text{def}}{=} \ln \tfrac{\det G(\alpha)}{\det G(0)} \ = \ 2\ln\left(1 + \alpha \cdot \tfrac{r-1}{2}\right) + (n-1)\ln(1-\alpha)$$

*attains its maximum at* $\alpha^* = \frac{2}{n+1} \cdot \frac{r-n}{r-1}$*. Moreover,*

$$\begin{aligned} V(\alpha^*) &= 2\ln\tfrac{r-1}{n+1} + (n-1)\ln\tfrac{(n-1)(r+1)}{(n+1)(r-1)} \\ &\ge 2\left[\ln(1+\sigma) - \tfrac{\sigma}{1+\sigma}\right] \overset{(5.1.23)}{\ge} \ \tfrac{2\sigma^2}{(1+\sigma)(2+\sigma)}, \end{aligned} \tag{7.2.8}$$

*where* $\sigma = \frac{r-n}{n+1}$.

*Proof* In view of Corollary 3.1.5, we need to prove that for all $x \in \mathbb{E}$

$$\begin{aligned} \xi_{E_\alpha}(x) &\equiv \alpha \cdot \tfrac{r-1}{2r} \cdot \langle g, x\rangle + \left[(1-\alpha)\|x\|^2_G + \left(\tfrac{\alpha}{r} + \left(\tfrac{r-1}{2}\right)^2 \cdot \left(\tfrac{\alpha}{r}\right)^2\right)\langle g, x\rangle^2\right]^{1/2} \\ &\le \xi_{C_g(G)}(x) \ = \ \max\{\|x\|_G, \langle g, x\rangle\}. \end{aligned}$$

If $\|x\|_G \le \langle g, x\rangle$, then

$$\xi_{E_\alpha}(x) \le \alpha \cdot \tfrac{r-1}{2r} \cdot \langle g, x\rangle + |1 - \alpha \cdot \tfrac{r-1}{2r}| \cdot \langle g, x\rangle \ = \ \langle g, x\rangle.$$

Otherwise, we have $-r\|x\|_G \le \langle g, x\rangle \le \|x\|_G$. Note that the value $\xi_{E_\alpha}(x)$ depends on $\langle g, x\rangle$ in a convex way. Therefore, in view of Corollary 3.1.2, its maximum is achieved at the end points of the feasible interval for $\langle g, x\rangle$. For the end point

$\langle g, x\rangle = \|x\|_G$, we have already proved that $\xi_{E_\alpha}(x) = \|x\|_G$. Consider now the case $\langle g, x\rangle = -r\|x\|_G$. Then,

$$\begin{aligned}\xi_{E_\alpha}(x) &= -\alpha \cdot \tfrac{r-1}{2} \cdot \|x\|_G + \left[(1-\alpha)\|x\|_G^2 + \left(\tfrac{\alpha}{r} + \left(\tfrac{r-1}{2}\right)^2 \cdot \left(\tfrac{\alpha}{r}\right)^2\right) r^2 \|x\|_G^2\right]^{1/2}\\ &= \|x\|_G.\end{aligned}$$

Thus, we have proved that $E_\alpha \subseteq C_g(G)$ for any $\alpha \in [0, 1)$. Further,

$$\begin{aligned}V(\alpha) &= \ln\det(G^{-1/2}G(\alpha)G^{-1/2})\\ &= \ln\det\left((1-\alpha)I_n + \left(\tfrac{\alpha}{r} + \left(\tfrac{r-1}{2}\right)^2 \cdot \left(\tfrac{\alpha}{r}\right)^2\right) G^{-1/2}gg^*G^{-1/2}\right)\\ &= \ln\left(1 - \alpha + \left(\tfrac{\alpha}{r} + \left(\tfrac{r-1}{2}\right)^2 \cdot \left(\tfrac{\alpha}{r}\right)^2\right) \cdot r^2\right) + (n-1)\ln(1-\alpha)\\ &= 2\ln\left(1 + \alpha \cdot \tfrac{r-1}{2}\right) + (n-1)\ln(1-\alpha).\end{aligned}$$

Hence, in view of Theorem 2.1.1, the optimality condition for the concave function $V(\cdot)$ is as follows:

$$\tfrac{n-1}{1-\alpha} = \tfrac{r-1}{1+\alpha\cdot\frac{r-1}{2}}.$$

Thus, the maximum is attained at $\alpha^* = \frac{2}{n+1} \cdot \frac{r-n}{r-1}$. Defining $\sigma = \frac{r-n}{n+1}$, we get

$$\begin{aligned}V(\alpha^*) &= 2\ln\left(1 + \alpha^* \cdot \tfrac{r-1}{2}\right) + (n-1)\ln(1-\alpha^*)\\ &= 2\ln(1+\sigma) - (n-1)\ln\left(1 + \tfrac{2(r-n)}{(n-1)(r+1)}\right)\\ &\geq 2\ln(1+\sigma) - \tfrac{2(r-n)}{r+1} = 2\left[\ln(1+\sigma) - \tfrac{\sigma}{1+\sigma}\right]. \qquad \square\end{aligned}$$

In this section, we are interested in solving the following problem. Let $C \subset \mathbb{R}^n$ be a convex set with nonempty interior. For a given $\gamma > 1$, we need to find a $\gamma n$-rounding for $C$. An initial approximation to the solution of this problem is given by a point $v_0$ and a matrix $G_0 \succ 0$ such that $W_1(v_0, G_0) \subseteq Q \subseteq W_R(v_0, G_0)$ for certain $R \geq 1$. We assume that $n \geq 2$.

Let us analyze the following algorithmic scheme.

```
For $k \geq 0$ iterate:

1. Compute $g_k \in C: \ \|g_k - v_k\|^*_{G_k} = r_k \overset{\mathrm{def}}{=} \max\limits_{g \in C} \ \|g - v_k\|^*_{G_k}$.

2. If $r_k \leq \gamma n$ then Stop else

      $\alpha_k = \frac{2}{n+1} \cdot \frac{r_k - n}{r_k - 1}, \quad \beta_k = \frac{\alpha_k}{r_k} + \left(\frac{r_k - 1}{2}\right)^2 \cdot \left(\frac{\alpha_k}{r_k}\right)^2,$        (7.2.9)

      $v_{k+1} = v_k + \alpha_k \frac{r_k - 1}{2r_k}(g_k - v_k),$

      $G_{k+1} = (1 - \alpha_k)G_k + \beta_k \cdot (g_k - v_k)(g_k - v_k)^T.$

end.
```

The complexity bound for this scheme is given by the following statement.

**Theorem 7.2.2** *Let $W_1(v_0, G_0) \subseteq C \subseteq W_R(v_0, G_0)$ for some $R \geq 1$. Then scheme (7.2.9) terminates after*

$$\frac{(1+2\gamma)(2+\gamma)}{2(\gamma-1)^2} \cdot n \ln R \tag{7.2.10}$$

*iterations at most.*

*Proof* Note that the coefficient $\alpha_k$, vector $v_{k+1}$ and matrix $G_{k+1}$ in Step 2 of (7.2.9) are chosen in accordance with Lemma 7.2.2. Since the method runs as long as

$$\sigma_k \overset{\mathrm{def}}{=} \frac{r_k - n}{n+1} \ \geq \ \frac{n}{n+1}(\gamma - 1) \ \geq \ \frac{2}{3}(\gamma - 1),$$

in view of inequality (7.2.8), at each step $k \geq 0$ we have

$$\ln \det G_{k+1} \geq \ln \det G_k + \frac{2\sigma_k^2}{(1+\sigma_k)(2+\sigma_k)} \ \geq \ \ln \det G_k + \frac{4(\gamma-1)^2}{(1+2\gamma)(2+\gamma)}. \tag{7.2.11}$$

Note that for any $k \geq 0$, we have

$$\begin{aligned} \det(G_k)^{1/2} \cdot \mathrm{vol}_n\,(W_1(I_n)) = \mathrm{vol}_n\,(W_1(G_k)) \ &\leq \ \mathrm{vol}_n\,(C) \ \leq \ \mathrm{vol}_n\,(W_R(G_0)) \\ &= R^n \cdot \det(G_0)^{1/2} \cdot \mathrm{vol}_n\,(W_1(I_n)). \end{aligned}$$

Hence, $\ln\det G_k - \ln\det G_0 \le 2n \ln R$, and we get bound (7.2.10) by summing up the inequalities (7.2.11). □

Note that in the case $C = \mathrm{Conv}\,\{a_i,\ i = 1,\dots,m\}$, scheme (7.2.9) can be implemented efficiently in the same style as (7.2.7). We leave the derivation of this modification and its complexity analysis as an exercise for the reader. The starting rounding ellipsoid for such a set $C$ can be chosen as follows.

**Lemma 7.2.3** *Assume that the set $C = Conv\,\{a_i,\ i = 1,\dots,m\}$ has nonempty interior. Define*

$$\hat{a} = \frac{1}{m}\sum_{i=1}^m a_i, \quad G = \frac{1}{R^2}\sum_{i=1}^m (a_i - \hat{a})(a_i - \hat{a})^T,$$

*where $R = \sqrt{m(m-1)}$. Then $W_1(\hat{a}, G) \subset C \subset W_R(\hat{a}, G)$.*

*Proof* For any $x \in \mathbb{R}^n$ and $r > 0$, we have

$$\xi_{W_r(\hat{a},G)}(x) = \langle \hat{a}, x\rangle + r\|x\|_G = \langle \hat{a}, x\rangle + \frac{r}{R}\left[\sum_{i=1}^m \langle a_i - \hat{a}, x\rangle^2\right]^{1/2}.$$

Thus, we have $\xi_{W_R(\hat{a},G)}(x) \ge \max\limits_{1\le i\le m} \langle a_i, x\rangle = \xi_C(x)$. Hence, $W_R(\hat{a}, G) \supset C$. Further, let

$$\tau_i = \langle a_i - \hat{a}, x\rangle, \quad i = 1,\dots,m, \quad \text{and}$$

$$\hat{\tau} = \max_{1\le i\le m} \langle a_i, x\rangle - \langle \hat{a}, x\rangle \ge 0.$$

Note that $\sum\limits_{i=1}^m \tau_i = 0$ and $\tau_i \le \hat{\tau}$ for all $i$. Therefore,

$$\begin{aligned}\xi_{W_1(\hat{a},G)}(x) - \langle \hat{a}, x\rangle &\le \frac{1}{R}\max_{\tau_i}\left\{\left[\sum_{i=1}^m \tau_i^2\right]^{1/2} : \sum_{i=1}^m \tau_i = 0,\ \tau_i \le \hat{\tau},\ i = 1,\dots,m\right\}\\ &= \frac{\hat{\tau}}{R}\sqrt{m(m-1)} = \max_{1\le i\le m}\langle a_i, x\rangle - \langle \hat{a}, x\rangle\\ &= \xi_C(x) - \langle \hat{a}, x\rangle.\end{aligned}$$

Thus, in view of Corollary 3.1.5, $W_1(\hat{a}, G) \subset C$. □

*Remark 7.2.2* In the same way as it was done in Remark 7.2.1, we can use algorithm (7.2.9) to prove John's Theorem for general convex sets. We leave this reasoning as an exercise for the reader. □

#### 7.2.1.3 Sign-Invariant Convex Sets

We call a set $C \subset \mathbb{R}^n$ *sign-invariant* if, for any point $g$ from $C$, an arbitrary change of signs of its entries leaves the point inside $C$. In other words, for any $g \in C \bigcap \mathbb{R}^n_+$, we have

$$B(g) \equiv \{s \in \mathbb{R}^n : -g \leq s \leq g\} \subseteq C.$$

Examples of such sets are given by unit balls of $\ell_p$-norms or by Euclidean norms generated by diagonal matrices.

Clearly, any sign-invariant set is centrally symmetric. Thus, in view of Lemma 7.2.1, for such a set there exists a $\sqrt{n}$-ellipsoidal rounding (this is John's Theorem). We will see that an important additional feature of sign-invariant sets is that the matrix of the corresponding quadratic form can be *diagonal*.

Let $D \succ 0$ be a diagonal matrix. Let us choose an arbitrary vector $g \in \mathbb{R}^n_+$. Define

$$C = \text{Conv } \{W_1(D), B(g)\},$$

$$G(\alpha) = (1-\alpha)D + \alpha D^2(g).$$

Clearly $C$ is a sign-invariant set. Consider the function

$$V(\alpha) = \ln \tfrac{\det G(0)}{\det G(\alpha)} \;=\; -\sum_{i=1}^{n} \ln\left(1 + \alpha(\tau_i - 1)\right), \quad \alpha \in [0, 1),$$

where $\tau_i = \frac{(g^{(i)})^2}{D^{(i)}}$, $i = 1, \ldots, n$. Note that $V(\cdot)$ is a standard self-concordant function (see Sect. 5.1). For our analysis it is important that

$$\begin{aligned} V'(0) &= n - \sum_{i=1}^{n} \tau_i = n - \left(\|g\|^*_D\right)^2, \quad \text{and} \\ V''(0) &= \sum_{i=1}^{n} (\tau_i - 1)^2. \end{aligned} \tag{7.2.12}$$

**Lemma 7.2.4** *For any $\alpha \in [0, 1]$, $W_1(G(\alpha)) \subseteq C$. Assuming that $(\|g\|^*_D)^2 > n$, define the step*

$$\alpha^* \overset{\text{def}}{=} \tfrac{(\|g\|^*_D)^2 - n}{(2(\|g\|^*_D)^2 - n)\cdot(\|g\|^*_D)^2}.$$

*Then, $\alpha^* \in (0, \frac{1}{n}]$, and for any $\gamma \in \left(1, \frac{1}{\sqrt{n}}\|g\|^*_D\right]$ we have*

$$V(\alpha^*) \leq \ln\left(1 + \tfrac{\gamma^2 - 1}{\gamma^2}\right) - \tfrac{\gamma^2 - 1}{\gamma^2} \;<\; 0. \tag{7.2.13}$$

*Proof* For any $\alpha \in [0, 1]$ and $x \in \mathbb{R}^n$, we get

$$\begin{aligned}
[\xi_{W_1(G(\alpha))}(x)]^2 &= (1-\alpha)\langle Dx, x\rangle + \alpha \sum_{i=1}^{n} (g^{(i)} x^{(i)})^2 \\
&\leq (1-\alpha)\langle Dx, x\rangle + \alpha \left( \sum_{i=1}^{n} g^{(i)} \cdot |x^{(i)}| \right)^2 \\
&\leq \left[\max\{\xi_{W_1(D)}(x), \xi_{B(g)}(x)\}\right]^2 \; = \; [\xi_C(x)]^2.
\end{aligned}$$

Further, let $S = \sum\limits_{i=1}^{n} \tau_i = (\|g\|_D^*)^2$. By assumption, $S > n$. Therefore,

$$\begin{aligned}
V''(0) &\leq \max_{\tau} \left\{ \sum_{i=1}^{n} (\tau_i - 1)^2 : \; \sum_{i=1}^{n} \tau_i = S, \; \tau_i \geq 0, \; i = 1 \dots n \right\} \\
&= (S-1)^2 + n - 1 \; < \; S^2.
\end{aligned}$$

Since $V(\cdot)$ is a standard self-concordant function, by inequality (5.1.16) we have:

$$\begin{aligned}
V(\alpha) &\leq V(0) + \alpha \cdot V'(0) + \omega_*(\alpha \cdot (V''(0))^{1/2}) \\
&\leq -\alpha \cdot (S-n) + \omega_*(\alpha \cdot S),
\end{aligned} \tag{7.2.14}$$

where $\omega_*(\tau) = -\tau - \ln(1-\tau)$. By Theorem 2.1.1, the minimum of the right-hand side of this inequality is attained at the solution of the equation

$$S - n = \tfrac{\alpha_* S^2}{1 - \alpha_* S}.$$

Thus, $\alpha_* = \frac{S-n}{S\cdot(2S-n)} < \frac{1}{n}$. By Lemma 5.1.4, the decrease of the right-hand side in (7.2.14) is equal to

$$\omega\left(1 - \tfrac{n}{S}\right) \geq \omega(1 - \gamma^{-2}),$$

where $\omega(t) = t - \ln(1+t)$. □

**Corollary 7.2.1** *For any sign-symmetric set $C \subset \mathbb{R}^n$ with nonempty interior, there exists a diagonal matrix $D \succ 0$ such that*

$$W_1(D) \subseteq C \subseteq W_{\sqrt{n}}(D).$$

*Proof* For $R$ big enough, the set $\{D \succeq 0 : \; W_1(D) \subseteq C \subseteq W_R(D)\}$ is nonempty, closed, and bounded. Therefore, the existence of $\sqrt{n}$-rounding follows from inequality (7.2.13). □

For us, Corollary 7.2.1 is important because of the following consequence.

**Lemma 7.2.5** *Let all vectors $a_i \in \mathbb{R}^n$, $i = 1, \dots, m$, have nonnegative coefficients. Assume that there exists a diagonal matrix $D \succ 0$ such that*

$$W_1(D) \subseteq Conv\ \{B(a_i),\ i = 1, \dots, m\} \subseteq W_{\gamma\sqrt{n}}(D)$$

*for certain $\gamma \geq 1$. Then the function $f(x) = \max\limits_{1 \leq i \leq m} \langle a_i, x\rangle$ satisfies the inequalities*

$$\|x\|_D \leq f(x) \leq \gamma\sqrt{n} \cdot \|x\|_D \quad \forall x \in \mathbb{R}^n_+. \tag{7.2.15}$$

*Proof* Consider the function: $\hat{f}(x) = \max\limits_{1 \leq i \leq m} \sum\limits_{j=1}^{n} a_i^{(j)} |x^{(j)}|$. In view of Lemma 3.1.13, its subdifferential can be expressed as follows:

$$\partial \hat{f}(0) = \text{Conv}\ \{B(a_i),\ i = 1, \dots, m\}.$$

Thus, for any $x \in \mathbb{R}^n$ we have

$$\|x\|_D = \max_s \{\langle s, x\rangle : \ s \in W_1(D)\} \ \leq \ \max_s \{\langle s, x\rangle : \ s \in \partial \hat{f}(0)\} \equiv \hat{f}(x)$$

$$\leq \max_s \{\langle s, x\rangle : \ s \in W_{\gamma\sqrt{n}}(D)\} \ = \ \gamma\sqrt{n} \cdot \|x\|_D.$$

It remains to note that $\hat{f}(x) \equiv f(x)$ for all $x \in \mathbb{R}^n_+$. □

**Corollary 7.2.2** *Let $a_i \in \mathbb{R}^n_+$, $i = 1, \dots, m$. Consider the set*

$$\mathscr{F} = \{x \in \mathbb{R}^n_+ : \ \langle a_i, x\rangle \leq b_i,\ i = 1, \dots, m\}$$

*with $b_i > 0$, $i = 1, \dots, m$. Then there exists a diagonal matrix $D \succ 0$ such that*

$$W_1(D) \bigcap \mathbb{R}^n_+ \subset \mathscr{F} \subset W_{\sqrt{n}}(D) \bigcap \mathbb{R}^n_+. \tag{7.2.16}$$

*Proof* Consider $f(x) = \max\limits_{1 \leq i \leq m} \frac{1}{b_i}\langle a_i, x\rangle$. In view of Corollary 7.2.1 the assumptions of Lemma 7.2.5 are satisfied with $\gamma = 1$. Since $\mathscr{F} = \{x \in \mathbb{R}^n_+ : \ f(x) \leq 1\}$, the inclusions (7.2.16) follow from inequalities (7.2.15). □

In this section, we are interested in finding a diagonal ellipsoidal rounding for the following sign-symmetric set:

$$C = \text{Conv}\ \{B(a_i),\ i = 1, \dots, m\}, \tag{7.2.17}$$

where $a_i \in \mathbb{R}^n_+ \setminus \{0\}$, $i = 1, \dots, m$. Our main assumption on the data is as follows:

$$\hat{a} \overset{\text{def}}{=} \tfrac{1}{m} \sum_{i=1}^{m} a_i \; > \; 0.$$

Let $\hat{D} = D^2(\hat{a})$.

**Lemma 7.2.6** $W_1(\hat{D}) \subset C \subset W_{m\sqrt{n}}(\hat{D})$.

*Proof* Since $\hat{a} \in C$, we have $W_1(\hat{D}) \subset B(\hat{a}) \subseteq C$. On the other hand,

$$C \subseteq B(m\,\hat{a}) \subset \left\{ x \in \mathbb{R}^n : \sum_{i=1}^{n} \left( \tfrac{x^{(i)}}{m\,\hat{a}^{(i)}} \right)^2 \leq n \right\} \; = \; W_{m\sqrt{n}}(\hat{D}). \qquad \square$$

For the sign-symmetric set $C \subset \mathbb{R}^n$ defined by (7.2.17), consider the following algorithmic scheme which finds a diagonal rounding of radius $\gamma\sqrt{n}$ with

$$\gamma > \left[ 1 + \tfrac{1}{\sqrt{n}} \right]^{1/2}.$$

```
Set $D_0 = \hat{D}$.

For $k \geq 0$ iterate:

1. Compute $i_k : \; \| a_{i_k} \|^*_{D_k} = r_k \overset{\text{def}}{=} \max_{1 \leq i \leq m} \| a_i \|^*_{D_k}$.

2. If $r_k \leq \gamma\sqrt{n}$ then Stop else                                   (7.2.18)

    $\beta_k := \sum_{j=1}^{n} \left( \frac{(a_{i_k}^{(j)})^2}{D_k^{(j)}} - 1 \right)^2, \quad \alpha_k := \frac{r_k^2 - n}{\beta_k + (r_k^2 - n)\beta_k^{1/2}}$,

    $D_{k+1} := (1 - \alpha_k) D_k + \alpha_k D^2(a_{i_k})$.

end.
```

Note that this scheme applies the rules described in Lemma 7.2.4 using the notation $\beta_k$ for $V''(0)$. Therefore, exactly as in Theorems 7.2.1 and 7.2.2, we can prove the following statement.

**Theorem 7.2.3** *For $\gamma \geq \left[1 + \frac{1}{\sqrt{n}}\right]^{1/2}$, the scheme (7.2.18) terminates at most after*

$$\left[\tfrac{\gamma^2-1}{\gamma^2} - \ln\left(1 + \tfrac{\gamma^2-1}{\gamma^2}\right)\right]^{-1} \cdot n(\ln n + 2\ln m)$$

*iterations.*

Note that the number of operations during each iteration of the scheme (7.2.18) is proportional to the number of nonzero elements in the matrix $A = (a_1, \ldots, a_m)$.

### 7.2.2 Minimizing the Maximal Absolute Value of Linear Functions

Consider the following problem of Linear Programming:

$$\min_{y \in \mathbb{R}^{n-1}} \max_{1 \leq i \leq m} |\langle \bar{a}_i, y \rangle - c_i|. \tag{7.2.19}$$

Defining $a_i = (\bar{a}_i^T, -c_i)^T$, $i = 1, \ldots, m$, $x = \begin{pmatrix} y \\ \tau \end{pmatrix} \in \mathbb{R}^n$ and $d = e_n$, we can rewrite this problem in a conic form (see Sect. 7.1):

$$\text{Find } f^* = \min_x \left\{ f(x) \overset{\text{def}}{=} \max_{1 \leq i \leq m} |\langle a_i, x \rangle| : \ \langle d, x \rangle = 1 \right\}. \tag{7.2.20}$$

In Sect. 7.1, in order to construct an ellipsoidal rounding for $\partial f(0)$, we used the composite structure of the function $f(\cdot)$. However, the radius of this rounding was quite large, of the order $O(\sqrt{m})$. Now, by method (7.2.4) we can efficiently precompute a rounding ellipsoid for this set which radius is proportional to $O(\sqrt{n})$. Let us show that this leads to a much more efficient minimization scheme.

Let us fix some $\gamma > 1$. Assume that using the process (7.2.4) we managed to construct an ellipsoidal rounding for the centrally symmetric set $\partial f(0)$ of radius $\gamma\sqrt{n}$:

$$W_1(G) \ \subseteq \ \partial f(0) \ \equiv \ \text{Conv}\,\{\pm a_i, \ i = 1, \ldots, m\} \ \subseteq \ W_{\gamma\sqrt{n}}(G).$$

The immediate consequences are as follows:

$$\|x\|_G \ \leq \ f(x) \equiv \sup_s \{\langle s, x \rangle : \ s \in \partial f(0)\} \leq \gamma\sqrt{n} \cdot \|x\|_G, \tag{7.2.21}$$

$$\|a_i\|_G^* \ \leq \ \gamma\sqrt{n}, \quad i = 1, \ldots, m. \tag{7.2.22}$$

Let us now fix a smoothing parameter $\mu > 0$. Consider the following approximation of the function $f(\cdot)$:

$$f_\mu(x) = \mu \ln\left(\sum_{i=1}^{m} \left[e^{\langle a_i,x\rangle/\mu} + e^{-\langle a_i,x\rangle/\mu}\right]\right).$$

Clearly $f_\mu(\cdot)$ is convex and continuously differentiable infinitely many times on $\mathbb{R}^n$. Moreover,

$$f(x) \le f_\mu(x) \le f(x) + \mu\ln(2m), \quad \forall x \in \mathbb{R}^n. \tag{7.2.23}$$

Finally, note that for any point $x$ and any direction $h$ from $\mathbb{R}^n$ we have

$$\langle \nabla f_\mu(x), h\rangle = \sum_{i=1}^{m} \lambda_\mu^{(i)}(x)\cdot\langle a_i, h\rangle,$$

$$\lambda_\mu^{(i)}(x) = \tfrac{1}{\omega_\mu(x)}\cdot\left(e^{\langle a_i,x\rangle/\mu} - e^{-\langle a_i,x\rangle/\mu}\right), \quad i = 1,\dots,m,$$

$$\omega_\mu(x) = \sum_{i=1}^{m}\left(e^{\langle a_i,x\rangle/\mu} + e^{-\langle a_i,x\rangle/\mu}\right).$$

Therefore, the expression for the Hessian is as follows:

$$\langle \nabla^2 f_\mu(x)h, h\rangle = \tfrac{1}{\mu}\sum_{i=1}^{m}\tfrac{\langle a_i,h\rangle^2}{\omega_\mu(x)}\left(e^{\langle a_i,x\rangle/\mu} + e^{-\langle a_i,x\rangle/\mu}\right) - \tfrac{1}{\mu}\left(\sum_{i=1}^{m}\lambda_\mu^{(i)}(x)\cdot\langle a_i,h\rangle\right)^2.$$

In view of (7.2.22), we have

$$\langle \nabla^2 f_\mu(x)h, h\rangle \le \tfrac{1}{\mu}\left(\max_{1\le i\le m}\|a_i\|_G^*\right)^2\cdot\|h\|_G^2 \le \tfrac{\gamma^2 n}{\mu}\cdot\|h\|_G^2.$$

In view of Theorem 2.1.6, this implies that the gradient of the function $f_\mu(\cdot)$ is Lipschitz continuous in the metric $\|\cdot\|_G$ with Lipschitz constant $L_\mu = \frac{\gamma^2 n}{\mu}$:

$$\|\nabla f_\mu(x) - \nabla f_\mu(y)\|_G^* \le L_\mu\|x-y\|_G \quad \forall x, y \in \mathbb{E}.$$

Our approach is very similar to that of Sect. 7.1. Consider the problem

$$\min_x\{\phi(x);\ x \in Q\}, \tag{7.2.24}$$

where $Q$ is a closed convex set and the differentiable convex function $\phi(\cdot)$ has a gradient which is Lipschitz continuous in the Euclidean norm $\| \cdot \|_G$ with constant $L$. Let us write down here the optimal method (2.2.63) for solving the problem (7.2.24).

```
Method $S(\phi, L, Q, G, x_0, N)$

Set $v_0 = x_0$. For $k = 0, \dots, N-1$ do
    1. Set $y_k = \frac{k}{k+2} x_k + \frac{2}{k+2} v_k$.
    2. Compute $\nabla \phi(y_k)$.
    3. $v_{k+1} = \arg\min\limits_{v \in Q} \left[ \langle \sum\limits_{i=0}^{k} \frac{i+1}{2} \nabla \phi(y_i), v - x_0 \rangle + \frac{L}{2} \| v - x_0 \|_G^2 \right]$.
    4. $x_{k+1} := \frac{k}{k+2} x_k + \frac{2}{k+2} v_{k+1}$.

Return: $S(\phi, L, Q, G, x_0, N) \equiv x_N$.
```

(7.2.25)

In accordance with Theorem 6.1.2, the output of this scheme $x_N$ satisfies the following inequality

$$\phi(x_N) - \phi(x_\phi^*) \le \frac{2L \| x_0 - x_\phi^* \|_G^2}{N(N+1)}, \tag{7.2.26}$$

where $x_\phi^*$ is an optimal solution to problem (7.2.24).

As in Sect. 7.1, we are going to use the scheme (7.2.25) in order to compute an approximate solution to (7.2.20) with a certain relative accuracy $\delta > 0$. Define

$$\begin{aligned} Q(r) &= \{ x \in \mathbb{R}^n : \ \langle d, x \rangle = 1, \ \| x \|_G \le r \}, \\ x_0 &= \frac{G^{-1} d}{\langle d, G^{-1} d \rangle}, \\ \tilde{N} &= \left\lfloor 2 e \gamma \sqrt{2n \ln(2m)} \left( 1 + \frac{1}{\delta} \right) \right\rfloor . \end{aligned}$$

Consider the following method.

```
Set $\hat{x}_0 = x_0$.

For $t \geq 1$ iterate:

    $\mu_t := \frac{\delta f(\hat{x}_{t-1})}{2e(1+\delta)\ln(2m)}$;    $L_{\mu_t} := \frac{\gamma^2 n}{\mu_t}$;

    $\hat{x}_t := S\left(f_{\mu_t}, L_{\mu_t}, Q(f(\hat{x}_{t-1})), G, x_0, \tilde{N}\right)$;

    If $f(\hat{x}_t) \geq \frac{1}{e} f(\hat{x}_{t-1})$ then $T := t$ and Stop.
```
(7.2.27)

**Theorem 7.2.4** *The number of points generated by method (7.2.27) is bounded as follows:*

$$T \leq 1 + \ln(\gamma\sqrt{n}). \tag{7.2.28}$$

*The last point of the process satisfies inequality $f(\hat{x}_T) \leq (1+\delta)f^*$. The total number of lower-level steps in the process (7.2.27) does not exceed*

$$2\gamma e(1 + \ln(\gamma\sqrt{n}))\sqrt{2n\ln(2m)}\left(1 + \tfrac{1}{\delta}\right). \tag{7.2.29}$$

*Proof* Let $x^*$ be an optimal solution to the problem (7.2.20). Note that all points $\hat{x}_t$ generated by (7.2.27) are feasible for (7.2.20). Therefore, in view of (7.2.21)

$$f(\hat{x}_t) \geq f^* \geq \|x^*\|_G.$$

Thus, $x^* \in Q(f(\hat{x}_t))$ for any $t \geq 0$. Let

$$f_t^* = f_{\mu_t}(x_t^*) = \min_x \{f_{\mu_t}(x) : \; x \in Q(f(\hat{x}_{t-1}))\}.$$

Since $x^* \in Q(f(\hat{x}_t))$, in view of (7.2.23) we have

$$f_t^* \leq f_{\mu_t}(x^*) \leq f^* + \mu_t \ln(2m).$$

By the first part of (7.2.23), $f(\hat{x}_t) \leq f_{\mu_t}(\hat{x}_t)$. Note that

$$\|x_0 - x_t^*\|_G \;\leq\; \|x_t^*\|_G \;\leq\; f(\hat{x}_{t-1}), \quad t \geq 1.$$

In view of (7.2.26), we have at the last iteration $T$:

$$\begin{aligned} f(\hat{x}_T) - f^* &\le f_{\mu_T}(\hat{x}_T) - f^*_T + \mu_T \ln(2m) \\ &\le \frac{2L_{\mu_T} f^2(\hat{x}_{T-1})}{\left(\tilde{N}+1\right)^2} + \mu_T \ln(2m) \;=\; \frac{2\gamma^2 n f^2(\hat{x}_{T-1})}{\mu_T\left(\tilde{N}+1\right)^2} + \mu_T \ln(2m) \\ &\le \frac{f^2(\hat{x}_{T-1})\,\delta^2}{4\mu_T e^2 \ln(2m)(1+\delta)^2} + \mu_T \ln(2m) \;=\; 2\mu_T \ln(2m). \end{aligned}$$

Further, in view of the choice of $\mu_t$ and the stopping criterion in (7.2.27), we have

$$2\mu_T \ln(2m) = \frac{\delta f(\hat{x}_{T-1})}{e(1+\delta)} \le \frac{\delta f(\hat{x}_T)}{1+\delta}.$$

Thus $f(\hat{x}_T) \le (1+\delta) f^*$.

It remains to prove the estimate (7.2.28) for the number of steps in the upper-level of the process. Indeed, by a simple induction it is easy to prove that at the beginning of stage $t$ the following inequality holds:

$$\left(\frac{1}{e}\right)^{t-1} f(x_0) \ge f(\hat{x}_{t-1}), \quad t \ge 1.$$

Note that $x_0$ is the projection of the origin on the hyperplane $\langle d, x\rangle = 1$. Therefore, in view of inequalities (7.2.21), we have

$$f^* \ge \|x^*\|_G \;\ge\; \|x_0\|_G \;\ge\; \frac{1}{\gamma\sqrt{n}} f(x_0).$$

Thus, at the final step of the scheme we have

$$\left(\frac{1}{e}\right)^{T-1} f(x_0) \ge f(\hat{x}_{T-1}) \ge f^* \ge \frac{1}{\gamma\sqrt{n}} f(x_0).$$

This leads to the bound (7.2.28). □

Recall that the preliminary stage of method (7.2.27), that is, the computation of $\gamma\sqrt{n}$-rounding for $\partial f(0)$ with relative accuracy $\gamma > 1$, can be performed by procedure (7.2.4) in

$$\frac{n^2}{6}(n+6m) + \frac{\gamma^2}{(\gamma-1)^2} n^2(2m+3n)\ln m \;=\; O(n^2(n+m)\ln m)$$

arithmetic operations. Since each step of method (7.2.25) takes $O(mn)$ operations, the complexity of the preliminary stage is dominant if $\delta$ is not too small, say $\delta > \frac{1}{\sqrt{n}}$.

### *7.2.3 Bilinear Matrix Games with Non-negative Coefficients*

Let $A = (a_1, \dots, a_m)$ be an $n \times m$-matrix with nonnegative coefficients. Consider the problem

$$\text{Find } f^* = \min_{x \in \Delta_n} \left\{ f(x) \overset{\text{def}}{=} \max_{1 \le i \le m} \langle a_i, x \rangle \right\}. \tag{7.2.30}$$

Note that this format can be used for different standard problem settings. Consider, for example, the *linear packing problem*

$$\text{Find } \psi^* = \max_{y \in \mathbb{R}^n_+} \left\{ \langle c, y \rangle : \ \langle a_i, y \rangle \le b^{(i)}, \ i = 1, \dots, m \right\},$$

where all entries of vectors $a_i$ are non-negative, $b > 0 \in \mathbb{R}^m$, and $c > 0 \in \mathbb{R}^n$. Then

$$\psi^* = \max_{y \in \mathbb{R}^n_+} \left\{ \langle c, y \rangle : \ \max_{1 \le i \le m} \tfrac{1}{b^{(i)}} \langle a_i, y \rangle \le 1 \right\} = \max_{y \in \mathbb{R}^n_+} \frac{\langle c, y \rangle}{\max\limits_{1 \le i \le m} \frac{1}{b^{(i)}} \langle a_i, y \rangle}$$

$$= \left[ \min_{y \in \mathbb{R}^n_+} \left\{ \max_{1 \le i \le m} \tfrac{1}{b^{(i)}} \langle a_i, y \rangle : \ \langle c, y \rangle = 1 \right\} \right]^{-1}$$

$$= \left[ \min_{x \in \Delta_n} \ \max_{1 \le i \le m} \tfrac{1}{b^{(i)}} \langle D^{-1}(c) a_i, x \rangle \right]^{-1}.$$

As usual, we can approximate the objective function $f(\cdot)$ in (7.2.30) by the following smooth function:

$$f_\mu(x) = \mu \ln \left( \sum_{i=1}^m e^{\langle a_i, x \rangle / \mu} \right).$$

In this case, the following relations hold:

$$f(x) \le f_\mu(x) \le f(x) + \mu \cdot \ln m, \quad \forall x \in \mathbb{R}^n. \tag{7.2.31}$$

Define

$$\hat{f}(x) = \max_{1 \le i \le m} \sum_{j=1}^n a_i^{(j)} |x^{(j)}|.$$

Note that the subdifferential of the homogeneous function $\hat{f}(\cdot)$ at the origin is as follows:

$$\partial f(0) = \text{Conv } \{B(a_i), \ i = 1, \dots, m\}.$$

In Sect. 7.2.1.3, we have seen that it is possible to compute a diagonal matrix $D \succ 0$ such that

$$W_1(D) \subseteq \partial \hat{f}(0) \;\subseteq\; W_{2\sqrt{n}}(D),$$

(this corresponds to the choice $\gamma = 2$ in scheme (7.2.18)). In view of Lemma 7.2.5, using this matrix we can define a Euclidean norm $\| \cdot \|_D$ such that

$$\|x\|_D \leq f(x) \;\leq\; 2\sqrt{n} \cdot \|x\|_D, \quad \forall x \in \mathbb{R}^n_+. \tag{7.2.32}$$

Moreover, in this norm the sizes of all $a_i$ are bounded by $2\sqrt{n}$.

Now, using the same reasoning as in Sect. 7.2.2, we can show that for any $x$ and $h$ from $\mathbb{R}^n$

$$\langle \nabla^2 f_\mu(x)h, h\rangle \leq \tfrac{4n}{\mu} \cdot \|h\|_D^2.$$

Hence, the gradient of this function is Lipschitz continuous with respect to the norm $\| \cdot \|_D$ with constant $\frac{4n}{\mu}$. This implies that the function $f_\mu(\cdot)$ can be minimized by the efficient method (6.1.19).

Let us fix some relative accuracy $\delta > 0$. Define

$$Q(r) = \{x \in \Delta_n : \; \|x\|_D \leq r\},$$

$$x_0 = \tfrac{D^{-1}\bar{e}_n}{\langle \bar{e}_n, D^{-1}\bar{e}_n\rangle},$$

$$\tilde{N} = \left\lfloor 4e\sqrt{2n\ln m}\left(1 + \tfrac{1}{\delta}\right)\right\rfloor.$$

Consider the following method.

```
Set $\hat{x}_0 = x_0$.

For $t \geq 1$ iterate:

    $\mu_t := \frac{\delta f(\hat{x}_{t-1})}{2e(1+\delta)\ln m}$;    $L_{\mu_t} := \frac{4n}{\mu_t}$;                (7.2.33)

    $\hat{x}_t := S\left(f_{\mu_t}, L_{\mu_t}, Q(f(\hat{x}_{t-1})), D, x_0, \tilde{N}\right)$;

    If $f(\hat{x}_t) \geq \frac{1}{e} f(\hat{x}_{t-1})$ then $T := t$ and Stop.
```

Justification of this scheme is very similar to that of (7.2.27).

**Theorem 7.2.5** *The number of points generated by method (7.2.27) is bounded as follows:*

$$T \le 1 + \ln(2\sqrt{n}). \tag{7.2.34}$$

*The last point of the process satisfies the inequality* $f(\hat{x}_T) \le (1+\delta) f^*$. *The total number of lower-level steps in the process (7.2.27) does not exceed*

$$4e(1 + \ln(2\sqrt{n}))\sqrt{2n \ln m}\left(1 + \tfrac{1}{\delta}\right). \tag{7.2.35}$$

*Proof* Let $x^*$ be an optimal solution to the problem (7.2.30). Note that all points $\hat{x}_t$ generated by (7.2.33) are feasible. Therefore, in view of (7.2.32),

$$f(\hat{x}_t) \ge f^* \ge \|x^*\|_D.$$

Thus, $x^* \in Q(f(\hat{x}_t))$ for any $t \ge 0$. Define

$$f_t^* = f_{\mu_t}(x_t^*) \; = \; \min_x \{ f_{\mu_t}(x) : \; x \in Q(f(\hat{x}_{t-1}))\}.$$

Since $x^* \in Q(f(\hat{x}_t))$, in view of (7.2.31), we have

$$f_t^* \le f_{\mu_t}(x^*) \; \le \; f^* + \mu_t \ln m.$$

By the first part of (7.2.31) $f(\hat{x}_t) \le f_{\mu_t}(\hat{x}_t)$. Note that

$$\|x_0 - x_t^*\|_D \le \|x_t^*\|_D \; \le \; f(\hat{x}_{t-1})$$

for all $t \ge 1$. Thus, in view of (7.2.26), at the last iteration $T$, we have:

$$\begin{aligned} f(\hat{x}_T) - f^* &\le f_{\mu_T}(\hat{x}_T) - f_T^* + \mu_T \ln m \; \le \; \frac{2L_{\mu_T} f^2(\hat{x}_{T-1})}{\left(\tilde{N}+1\right)^2} + \mu_T \ln m \\ &= \frac{8nf^2(\hat{x}_{T-1})}{\mu_T\left(\tilde{N}+1\right)^2} + \mu_T \ln m \; \le \; \frac{f^2(\hat{x}_{T-1})\,\delta^2}{4\mu_T e^2 \ln m (1+\delta)^2} + \mu_T \ln m \\ &= 2\mu_T \ln m. \end{aligned}$$

Further, in view of the choice of $\mu_T$ and the stopping criterion, we have

$$2\mu_T \ln m = \frac{\delta f(\hat{x}_{T-1})}{e(1+\delta)} \; \le \; \frac{\delta f(\hat{x}_T)}{1+\delta}.$$

Thus, $f(\hat{x}_T) \le (1+\delta) f^*$.

It remains to prove the estimate (7.2.34) for the number of steps of the upper-level process. Indeed, by simple induction it is easy to prove that at the beginning of stage $t$ the following inequality holds:

$$\left(\frac{1}{e}\right)^{t-1} f(x_0) \geq f(\hat{x}_{t-1}), \quad t \geq 1.$$

Note that $x_0$ is the projection of the origin at the hyperplane $\langle \bar{e}_n, x\rangle = 1$. Therefore, in view of inequalities (7.2.32), we have

$$f^* \geq \|x^*\|_D \ \geq \ \|x_0\|_D \geq \frac{1}{2\sqrt{n}} f(x_0).$$

Thus, at the last step of the scheme we have

$$\left(\frac{1}{e}\right)^{T-1} f(x_0) \geq f(\hat{x}_{T-1}) \ \geq \ f^* \ \geq \ \frac{1}{2\sqrt{n}} f(x_0).$$

This leads to the bound (7.2.34). □

Thus, we have seen that the scheme (7.2.33) needs $O\left(\frac{\sqrt{n \ln m}}{\delta} \ln n\right)$ iterations of the gradient scheme (7.2.25). Since the matrix $D$ is diagonal, each iteration of this scheme is very cheap. Its complexity is proportional to the number of nonzero elements in the matrix $A$. Note also that in Step 3 of scheme (7.2.25) it is necessary to compute projections onto the set $Q(r)$, which is an intersection of the simplex and a diagonal ellipsoid. However, since $D$ is a diagonal matrix, this can be done in $O(n \ln n)$ operations by relaxing the only equality constraint and arranging a one-dimensional search in the corresponding Lagrange multiplier.

### 7.2.4 *Minimizing the Spectral Radius of Symmetric Matrices*

For a matrix $X \in \mathbb{S}_n$, define its spectral radius:

$$\begin{aligned}\rho(X) &= \max_{1\leq i\leq n} |\lambda^{(i)}(X)| \ = \ \max\{\lambda^{(1)}(X), -\lambda^{(n)}(X)\}\\ &= \min_{\tau}\{\tau : \ \tau I_n \succeq \pm X\}.\end{aligned}$$

In view of Theorem 3.1.7, $\rho(X)$ is a convex function on $\mathbb{S}_n$. In this section, we consider the following optimization problem:

$$\text{Find } \phi_* \ = \ \min_{y\in Q}\{\phi(y) \overset{\text{def}}{=} \rho(A(y))\}, \tag{7.2.36}$$

where $Q \subset \mathbb{R}^m$ is a closed convex set separated from the origin, and $A(\cdot)$ is a linear operator from $\mathbb{R}^m$ to $\mathbb{S}_n$:

$$A(y) = \sum_{i=1}^{m} y^{(i)} A_i \in \mathbb{S}_n, \quad y \in \mathbb{R}^m.$$

We assume that matrices $\{A_i\}_{i=1}^m$ are linearly independent. Hence, the matrix $G \in \mathbb{S}_m$ with elements

$$G^{(i,j)} = \langle A_i, A_j \rangle_M, \quad i, j = 1, \dots, m,$$

is positive definite. Denote by $r$ the maximal rank of $A(y)$:

$$r = \max_{y \in \mathbb{R}^m} \operatorname{rank} A(y) \le \min \left\{ n, \sum_{i=1}^{m} \operatorname{rank} A_i \right\}.$$

We are going to solve (7.2.36) using a variant of the smoothing technique, which is applicable for solving structural convex optimization problems in *relative scale*. Note that in view of our assumptions $\phi^*$ is strictly positive.

First of all, we approximate a non-smooth objective function in (7.2.36) by a smooth one. For that, we use $F_p(X)$ defined by (6.3.6). Note that

$$\begin{aligned} F_p(X) &= \tfrac{1}{2} \langle X^{2p}, I_n \rangle_M^{1/p} \ge \tfrac{1}{2}\rho^2(X), \\ F_p(X) &\le \tfrac{1}{2}\rho^2(X) \cdot (\operatorname{rank} X)^{1/p}. \end{aligned} \tag{7.2.37}$$

Consider the problem

$$\text{Find } f_p^* = \min_{y \in \mathbb{R}^m} \{ f_p(y) \overset{\text{def}}{=} F_p(A(y)) : \ y \in Q \}. \tag{7.2.38}$$

From (7.2.37), we can see that

$$\tfrac{1}{2}\phi_*^2 \le f_p^* \le \tfrac{1}{2}\phi_*^2 \cdot r^{1/p}. \tag{7.2.39}$$

Our goal is to find a point $\bar{y} \in Q$ which solves (7.2.36) with relative accuracy $\delta > 0$:

$$\phi(\bar{y}) \le (1 + \delta)\phi_*.$$

Let us choose an integer $p$ which satisfies the following inequality

$$p(\delta) \overset{\text{def}}{=} \tfrac{1+\delta}{\delta} \ln r \le p \le 2p(\delta). \tag{7.2.40}$$

Assume that $\bar{y} \in Q$ solves (7.2.38) with relative accuracy $\delta$. Then, in view of (7.2.37) and (7.2.39), we have

$$\phi(\bar{y})/\phi_* \leq r^{\frac{1}{2p}} \cdot \sqrt{f_p(\bar{y})/f_p^*} \leq r^{\frac{1}{2p}} \cdot \sqrt{1+\delta}$$

$$\leq e^{\frac{\delta}{2(1+\delta)}} \cdot \sqrt{1+\delta} \leq 1+\delta.$$

Thus, we need to estimate the efficiency of method (6.1.19) as applied to the problem (7.2.38). Let us introduce the following norm

$$\| h \|_G = \langle Gh, h \rangle^{1/2}, \quad h \in \mathbb{R}^m.$$

Assuming that $p(\delta) \geq 1$ and using the estimate (6.3.8) and notation of Sect. 6.3.1, for any $y$ and $h$ from $\mathbb{R}^m$ we get

$$\begin{aligned} \langle \nabla^2 f_p(y)h, h \rangle &= \langle \nabla^2 F_p(A(y))A(h), A(h) \rangle_M \\ &\leq (2p-1) \| A(h) \|^2_{(2p)} \leq (2p-1) \| A(h) \|^2_{(2)} \\ &= (2p-1) \langle A(h), A(h) \rangle_M = (2p-1) \langle Gh, h \rangle \\ &= (2p-1) \| h \|^2_G. \end{aligned}$$

Thus, in view of Theorem 2.1.6 function $f_p(y)$ has Lipschitz continuous gradient on $\mathbb{R}^m$ with respect to the norm $\| \cdot \|_G$ with Lipschitz constant

$$L = 2p - 1 \leq 4p(\delta). \tag{7.2.41}$$

On the other hand, for any $X \in \mathbb{S}_n$ with rank $X \leq r$, and $p \geq 1$ we have

$$\tfrac{1}{r} \| X \|^2_{(2)} \leq \| X \|^2_{(\infty)} \leq \| X \|^2_{(2p)}.$$

Hence, $\frac{1}{2r} \| y \|^2_G \leq f_p(y)$ for any $y \in \mathbb{R}^m$. In particular,

$$\tfrac{1}{2r} \| y_p^* \|^2_G \leq f_p^*, \tag{7.2.42}$$

where $y_p^*$ is an optimal solution to (7.2.38).

Let $x_0 = \arg\min\limits_{y \in Q} \| y \|_G$. Since the norm $\| \cdot \|_G$ is Euclidean, and $Q$ is convex, in view of inequality (2.2.49), we have

$$\| y_p^* - x_0 \|^2_G \leq \| y_p^* \|^2_G - \| x_0 \|^2_G < \| y_p^* \|^2_G.$$

Combining this inequality with estimate (7.2.42), we get

$$\tfrac{1}{2}\|y_p^* - x_0\|_G^2 \leq \tfrac{1}{2}\|y_p^*\|_G^2 \ \leq \ rf_p^*. \tag{7.2.43}$$

In order to apply method (2.2.63) to problem (7.2.38), let us choose the following prox-function:

$$d(x) = \tfrac{1}{2}\|x - x_0\|_G^2. \tag{7.2.44}$$

Note that the convexity parameter of this function is equal to one. Hence, in view of bounds (7.2.41), (7.2.42), and (6.1.21), method (6.1.19) launched from the starting point $x_0$ converges as follows:

$$f_p(x_k) - f_p^* \leq \tfrac{16(1+\delta)r\ln r}{\delta\cdot k(k+1)} \cdot f_p^*. \tag{7.2.45}$$

Hence, in order to solve problem (7.2.38) with relative accuracy $\delta$ (and, therefore, solve (7.2.36) with the same relative accuracy), method (6.1.19) needs at most

$$\tfrac{4}{\delta}\sqrt{(1+\delta)r\ln r} \tag{7.2.46}$$

iterations. Note that this bound does not depend on the data size of the particular problem instance.

At each iteration of method (6.1.19) as applied to the problem (7.2.38) with $d(\cdot)$ defined by (7.2.44), it is necessary to compute a projection of a point onto the set $Q$ with respect to the Euclidean metric $\|\cdot\|_G$. This operation is easy in the following cases.

- The set $Q$ is an affine subspace in $\mathbb{R}^m$. Then the projection can be computed by inverting the matrix $G$. An important example of such a problem is as follows:

$$\min_{y\in\mathbb{R}^m} \left\{ \rho\left(\sum_{i=1}^m y^{(i)}A_i\right) : \ y^{(1)} = 1 \right\}.$$

- The matrix $G$ and the set $Q$ are both simple. For example, if $\langle A_i, A_j\rangle = 0$ for $i \neq j$, then $G$ is a diagonal matrix. In this case, a projection onto a box, for example, is easy to compute. Such a situation occurs when the matrix $A(y)$ is parameterized directly by its entries.

Finally, note that the computation of the value and the gradient of the function $f_p(\cdot)$ can be done without eigenvalue decomposition of the matrix $A(y)$. Indeed, let $p = 2^k$ satisfy condition (7.2.40). Consider the following of sequence of matrices:

$$\begin{aligned} &X_0 = A(y), \ Y_0 = I_n, \\ &X_i = X_{i-1}^2, \ Y_i = Y_{i-1}X_{i-1}, \ i = 1, \ldots, k. \end{aligned} \tag{7.2.47}$$

By induction, it is easy to see that $X_k = A^p(y)$ and $Y_k = A^{p-1}(y)$. Hence, in accordance with (6.3.3), (6.3.6), and the definition of the function $f_p(\cdot)$ in (7.2.38), we have:

$$f_p(y) = \tfrac{1}{2}\langle X_k, I_n \rangle_M^{2/p},$$

$$\nabla f_p(y)^{(i)} = \tfrac{2 f_p(y)}{\langle X_k, I_n \rangle_M} \cdot \langle Y_k, A_i \rangle_M, \quad i = 1, \dots, m.$$

Note that the complexity of computing the matrix $A(y)$ is of the order of $O(n^2 m)$ arithmetic operations. The auxiliary computation (7.2.47) takes

$$O(n^3 \ln p) = O\left(n^3 \ln \tfrac{\ln r}{\delta}\right)$$

operations. After that the vector $\nabla f_p(y)$ can be computed in $O(n^2 m)$ arithmetic operations. Clearly, the complexity of the first and the last computation is much lower if the matrices $A_i$ are sparse.

Note also that the computation (7.2.47) can be performed more efficiently if the matrix $A(y)$ is represented in the form

$$A(y) = UTU^T, \quad UU^T = I_n,$$

where $T$ is a tri-diagonal matrix. Computation of this representation needs $O(n^3)$ arithmetic operations.

## 7.3 Barrier Subgradient Method

(Smoothing by a self-concordant barrier; The barrier subgradient scheme; Relative accuracy and maximization of positive concave functions; Applications: The fractional covering problem, the maximal concurrent flow problem, the minimax problem with nonnegative components, Semidefinite relaxation of the Boolean quadratic problem; Online Optimization as an alternative to Stochastic Programming.)

### *7.3.1 Smoothing by a Self-Concordant Barrier*

In Nonlinear Optimization the performance of numerical methods strongly depends on our ability execute some auxiliary operations related to the convex sets involved in the problem's formulation. Usually, the optimization methods assume the feasibility of one of the following actions:

**L:** Maximization of a linear function $\langle c, x \rangle$ over a convex set $Q$.

**S:** Maximization of the function $\langle c, x\rangle - d(x)$ in $x \in Q$, where $d$ is a strongly convex prox-function of the set $Q$.

**B:** Computation of the value and first two derivatives of some self-concordant barrier at the interior points of the convex set $Q$.

Note that in Structural Optimization we can always consider the optimization problem posed in a primal-dual setting. The most important example of such a representation is a bilinear saddle point formulation:

$$\min_{x\in Q_p} \max_{w\in Q_d} \{\langle Ax, w\rangle + \langle c, x\rangle + \langle b, w\rangle\}, \tag{7.3.1}$$

where $Q_p$ and $Q_d$ are closed convex sets in corresponding spaces and $A$ is a linear operator. Since the structure of the primal and dual sets may be of different complexity, we have six possible combinations of the above mentioned auxiliary operations. Let us present the known results on their complexity.

- $\mathbf{L_p}\bigotimes\mathbf{L_d}$. The complexity of this combination is still not clear.
- $\mathbf{S_p}\bigotimes\mathbf{S_d}$. This case is treated by the smoothing technique (see Chap. 6). An $\epsilon$-solution of the problem (7.3.1) can be obtained in

  $$O\left(\tfrac{1}{\epsilon}\cdot\|A\|\cdot[D_1D_2]^{1/2}\right)$$

  gradient steps, where $D_1$ and $D_2$ are the *sizes* of the primal and dual sets, and the norm $\|A\|$ is defined by the norms of the primal and dual spaces.
- $\mathbf{B_p}\bigotimes\mathbf{B_d}$. In this situation, Interior-Point Methods provide an $\epsilon$-solution of the problem (7.3.1) in

  $$O\left(\sqrt{\nu}\cdot\ln\tfrac{\nu}{\epsilon}\right)$$

  Newton steps, where $\nu$ is the parameter of a self-concordant barrier for a primal-dual feasible set $Q_p\times Q_d$ (see Chap. 5).
- $\mathbf{S_p}\bigotimes\mathbf{L_d}$. This case is similar to the standard Black-Box Nonsmooth Minimization. Primal-dual subgradient methods provide an $\epsilon$-solution to (7.3.1) in

  $$O\left(\tfrac{1}{\epsilon^2}\cdot\|A\|^2\cdot D_1\cdot D_2\right)$$

  gradient steps (see Sect. 3.2).
- $\mathbf{B_p}\bigotimes\mathbf{S_d}$. The complexity of this combination is not known yet.
- $\mathbf{B_p}\bigotimes\mathbf{L_d}$. The last variant is studied in this section. From the viewpoint of Black-Box Optimization, it corresponds to the problem of minimizing nonsmooth convex function over a feasible set endowed with a self-concordant barrier.

Let us recall our notation. For a linear operator $A : \mathbb{E} \to \mathbb{H}^*$, we denote by $A^* : \mathbb{H} \to \mathbb{E}^*$ the adjoint operator:

$$\langle Ax, y\rangle_{\mathbb{H}} = \langle A^* y, x\rangle_{\mathbb{E}}, \quad x \in \mathbb{E},\ y \in \mathbb{H}.$$

If there is no ambiguity, the subscripts of scalar products are omitted. For a *concave* function $f$, we denote by $\nabla f(x)$ one of its subgradients at $x$:

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle, \quad y, x \in \mathrm{dom}\ f.$$

For a function of two vector variables $\Psi(u, x)$, the notation $\nabla_2 \Psi(u, x)$ is used to denote its subgradient with respect to the second argument.

Let $Q \subset \mathbb{E}$ be a closed convex set containing no straight lines. We assume that $Q$ is endowed with a $\nu$-self-concordant barrier $F$ (see Sect. 5.3). In view of Theorem 5.1.6, its Hessian is non-degenerate at all points of the domain.

Consider another closed convex set $\hat{P} \subseteq \mathbb{E}$. We are mainly interested in the set

$$P = \hat{P} \bigcap Q,$$

which we assume to be bounded. Denote by $x_0$ its *constrained analytic center*:

$$x_0 = \arg\min_{x \in P_0} F(x) \in P_0 \overset{\mathrm{def}}{=} \hat{P} \bigcap \mathrm{int}\ Q \subseteq P. \tag{7.3.2}$$

Thus, $F(x) \ge F(x_0)$ for all $x \in P$. Since $Q$ contains no straight lines, $x_0$ is well defined (see Theorem 5.1.6).

For the set $P$, we introduce the following smooth approximation of its support function:

$$U_\beta(s) = \max_{u \in \hat{P}} \{\langle s, u - x_0\rangle - \beta[F(u) - F(x_0)]\}, \quad s \in \mathbb{E}^*, \tag{7.3.3}$$

where $\beta > 0$ is a smoothing parameter. Denote by $u^\star_\beta(s)$ the unique solution of the maximization problem (7.3.3). Then, in view of relation (5.3.17) and Theorem 6.1.1, we have

$$\nabla U_\beta(s) = u^\star_\beta(s) - x_0, \quad s \in \mathbb{E}^*. \tag{7.3.4}$$

For any $x \in \mathrm{int}\ Q$, consider the following local norms:

$$\|h\|_x = \langle \nabla^2 F(x) h, h\rangle^{1/2}, \qquad h \in \mathbb{E},$$

$$\|s\|^*_x = \langle s, [\nabla^2 F(x)]^{-1} s\rangle^{1/2}, \ s \in \mathbb{E}^*.$$

Then, we can guarantee the following level of smoothness of the function $U_\beta(\cdot)$.

**Lemma 7.3.1** *Let $\beta > 0$, $s \in \mathbb{E}^*$ and $x = u^\star_\beta(s)$. Then for any $g \in \mathbb{E}^*$ with $\|g\|^*_x < \beta$ we have*

$$U_\beta(s+g) \leq U_\beta(s) + \langle g, \nabla U_\beta(s)\rangle + \beta\omega_*(\tfrac{1}{\beta}\|g\|^*_x), \tag{7.3.5}$$

*where $\omega_*(\tau) = -\tau - \ln(1-\tau) \overset{(5.1.24)}{\leq} \frac{\tau^2}{2(1-\tau)}$ for $\tau \in [0,1)$.*

*Proof* In view of definition (7.3.3) and Theorem 2.2.9, for any $y \in P_0$ we have

$$\langle s - \beta\nabla F(x), y - x\rangle \;\leq\; 0. \tag{7.3.6}$$

Moreover, since $F$ is a standard self-concordant function, at any point $y \in \text{int } Q$

$$F(y) \geq F(x) + \langle \nabla F(x), y - x\rangle + \omega(\|y - x\|_x), \tag{7.3.7}$$

where $\omega(t) = t - \ln(1+t)$ (see inequality (5.1.14)). Hence,

$$\begin{array}{rl} & U_\beta(s+g) - U_\beta(s) - \langle g, \nabla U_\beta(s)\rangle \\ \overset{(7.3.4)}{=} & \max\limits_{y\in P_0}\{\langle s+g, y-x_0\rangle - \beta[F(y)-F(x_0)] - \langle s+g, x-x_0\rangle \\ & +\beta[F(x)-F(x_0)]\} \\ = & \max\limits_{y\in P_0}\{\langle s+g, y-x\rangle - \beta[F(y)-F(x)]\} \\ \overset{(7.3.6)}{\leq} & \max\limits_{y\in P_0}\{\langle g, y-x\rangle + \beta[\langle \nabla F(x), y-x\rangle - F(y) + F(x)]\} \\ \overset{(7.3.7)}{\leq} & \max\limits_{y\in P_0}\{\langle g, y-x\rangle - \beta\omega(\|y-x\|_x)\} \;\leq\; \sup\limits_{\tau\geq 0}\{\,\tau\|g\|^*_x - \beta\omega(\tau)\,\}. \end{array}$$

If $\|g\|^*_x < \beta$, then the supremum in the right-hand side is equal to $\beta\omega_*(\frac{1}{\beta}\|g\|^*_x)$ (see Lemma 5.1.4). □

Consider now an affine function $\ell(x)$, $x \in P$. For $\beta \geq 0$ define

$$\ell^\star(\beta) \;=\; \max_{x\in P_0}\{\ell(x) - \beta[F(x) - F(x_0)]\} \;\geq\; \ell(x_0) \;\overset{\text{def}}{=}\; \ell_0. \tag{7.3.8}$$

Then $\ell^\star(0) = \max\limits_{x\in P} \ell(x) \;\overset{\text{def}}{=}\; \ell^\star$.

**Lemma 7.3.2** *For any* $\beta > 0$ *we have*

$$\ell^\star(\beta) \leq \ell^\star \;\leq\; \ell^\star(\beta) + \beta\nu\left(1 + \left[\ln \tfrac{\ell^\star - \ell_0}{\beta\nu}\right]_+\right), \tag{7.3.9}$$

*where* $[a]_+ = \max\{a, 0\}$. *Moreover,*

$$\ell^\star - \ell_0 \leq \left[\sqrt{\ell^\star(\beta) - \ell_0} + \sqrt{\beta\nu}\right]^2. \tag{7.3.10}$$

*Proof* The first part of inequality (7.3.9) follows from definitions (7.3.2) and (7.3.8). Let us prove the second part. Consider an arbitrary $y^\star \in \operatorname{Arg}\max\limits_{x\in P} \ell(x)$. Define

$$y(\alpha) = x_0 + \alpha(y^\star - x_0), \quad \alpha \in [0, 1].$$

In view of inequality (5.3.14), we have

$$F(y(\alpha)) \;\leq\; F(x_0) - \nu\ln(1-\alpha), \quad \alpha \in [0, 1).$$

Since $\ell(\cdot)$ is linear, this relation implies that

$$\begin{aligned} \ell^\star(\beta) &\geq \max_{\alpha\in[0,1)} \{\ell(y(\alpha)) - \beta[F(y(\alpha)) - F(x_0)]\} \\ &\geq (1-\alpha)\ell_0 + \alpha\ell^\star + \beta\nu\ln(1-\alpha), \quad \alpha \in [0, 1). \end{aligned} \tag{7.3.11}$$

The maximum in $\alpha$ of the latter expression is attained at $\alpha^\star = \left[1 - \frac{\beta\nu}{\ell^\star - \ell_0}\right]_+$. Thus, if $\frac{\ell^\star - \ell_0}{\beta\nu} \leq 1$ (that is $\alpha^\star = 0$), then $\ell^\star \leq \ell_0 + \beta\nu$, and (7.3.9) follows from (7.3.8). If $\alpha^\star > 0$, then we get (7.3.9) by direct substitution.

On the other hand, from (7.3.11) we have

$$\ell^\star - \ell_0 \leq \tfrac{1}{\alpha}\left[\ell^\star(\beta) - \ell_0 + \beta\nu\ln\left(1 + \tfrac{\alpha}{1-\alpha}\right)\right] \;\leq\; \tfrac{1}{\alpha}[\ell^\star(\beta) - \ell_0] + \tfrac{\beta\nu}{1-\alpha}.$$

Minimizing the latter expression in $\alpha$, we get (7.3.10). □

**Corollary 7.3.1** *For any* $\beta > 0$ *we have*

$$\ell^\star \leq \ell^\star(\beta) + \beta\nu\left[1 + 2\ln\left(1 + \sqrt{\tfrac{\ell^\star(\beta) - \ell_0}{\beta\nu}}\right)\right]. \tag{7.3.12}$$

### *7.3.2 The Barrier Subgradient Scheme*

In this section, we consider convex optimization problems in the following form:

$$\text{Find } f_\star \overset{\text{def}}{=} \max_x \{f(x) : \ x \in P\}, \tag{7.3.13}$$

where $f$ is a concave function and $P$ satisfies the structural assumptions specified at the beginning of Sect. 7.3.1. In the sequel, we assume $f$ to be subdifferentiable on $P_0$ and the set $P$ to be simple. The latter means that the auxiliary optimization problem (7.3.3) can be easily solved.

Consider now the generic scheme of the *Barrier Subgradient Method* (BSM).

```
Initialization: Set $s_0 = 0 \in \mathbb{E}^*$.

Iteration ($k \geq 0$):                                              (7.3.14)

    1. Choose $\beta_k > 0$ and compute $x_k = u^\star_{\beta_k}(s_k)$.

    2. Choose $\lambda_k > 0$ and set $s_{k+1} = s_k + \lambda_k \nabla f(x_k)$.
```

Recall that $u^\star_\beta(s)$ denotes the unique solution of the optimization problem (7.3.3). Thus, BSM is an *affine-invariant* scheme.

In order to analyze the performance of method (7.3.14), consider the following *gap functions*:

$$\ell_k(y) = \sum_{i=0}^{k} \lambda_i \langle \nabla f(x_i), y - x_i \rangle,$$

$$\ell^\star_k \overset{\text{def}}{=} \max_{y \in P} \ell_k(y), \quad k \geq 0.$$

**Theorem 7.3.1** *Assume that the parameters of scheme (7.3.14) satisfy the condition*

$$\lambda_k \|\nabla f(x_k)\|^*_{x_k} \leq \beta_k \leq \beta_{k+1}, \quad k \geq 0. \tag{7.3.15}$$

*Let* $S_k = \sum\limits_{i=0}^{k} \lambda_i$, *and* $A_k = \sum\limits_{i=0}^{k} \beta_i \omega_* \left( \frac{\lambda_i}{\beta_i} \|\nabla f(x_i)\|^*_{x_i} \right)$. *Then, for any* $k \geq 0$ *we have*

$$\ell^\star_k \leq A_k + \beta_{k+1}\nu \left[ 1 + 2 \ln \left( 1 + \sqrt{\frac{A_k}{\beta_{k+1}\nu} + 3 \frac{S_k}{\beta_{k+1}} \|\nabla f(x_0)\|^*_{x_0}} \right) \right]. \tag{7.3.16}$$

*Proof* Note that for any $k \geq 0$ we have

$$
\begin{aligned}
U_{\beta_{k+1}}(s_{k+1}) &\overset{(7.3.15)}{\leq} U_{\beta_k}(s_{k+1}) \\
&\overset{(7.3.5)}{\leq} U_{\beta_k}(s_k) + \lambda_k \langle \nabla f(x_k), u^\star_{\beta_k}(s_k) - x_0 \rangle + \beta_k \omega_* \left( \tfrac{\lambda_k}{\beta_k} \| \nabla f(x_k) \|^*_{x_k} \right).
\end{aligned}
$$

Since $U_{\beta_0}(0) = 0$, we conclude that

$$
\begin{aligned}
&\langle s_{k+1}, x_{k+1} - x_0 \rangle - \beta_{k+1}[F(x_{k+1}) - F(x_0)] \;=\; U_{\beta_{k+1}}(s_{k+1}) \\
&\leq \sum_{i=0}^{k} \lambda_i \langle \nabla f(x_i), x_i - x_0 \rangle + \sum_{i=0}^{k} \beta_i \omega_* \left( \tfrac{\lambda_i}{\beta_i} \| \nabla f(x_i) \|^*_{x_i} \right).
\end{aligned}
\tag{7.3.17}
$$

In view of the first-order optimality condition for (7.3.3), for all $y \in P_0$ we have

$$
\langle s_{k+1}, y - x_{k+1} \rangle \;\leq\; \beta_{k+1} \langle \nabla F(x_{k+1}), y - x_{k+1} \rangle. \tag{7.3.18}
$$

Note that $s_{k+1} = \sum\limits_{i=0}^{k} \lambda_i \nabla f(x_i)$. Therefore, for any $y \in P_0$ we obtain

$$
\begin{aligned}
\sum_{i=0}^{k} \lambda_i \langle \nabla f(x_i), y - x_i \rangle &\overset{(7.3.17)}{\leq} \langle s_{k+1}, y - x_{k+1} \rangle + \beta_{k+1}[F(x_{k+1}) - F(x_0)] + A_k \\
&\overset{(7.3.18)}{\leq} \beta_{k+1}[F(x_{k+1}) + \langle \nabla F(x_{k+1}), y - x_{k+1} \rangle - F(x_0)] \\
&\quad + A_k \\
&\leq \beta_{k+1}[F(y) - F(x_0)] + A_k.
\end{aligned}
$$

Hence, $\ell^\star_k(\beta_{k+1}) \leq A_k$. On the other hand, since $f$ is concave, we obtain

$$
\begin{aligned}
l_k(x_0) = \sum_{i=0}^{k} \lambda_i \langle \nabla f(x_i), x_0 - x_i \rangle \;&\geq\; \sum_{i=0}^{k} \lambda_i \langle \nabla f(x_0), x_0 - x_i \rangle \\
&\geq -\| \nabla f(x_0) \|^*_{x_0} \cdot \sum_{i=0}^{k} \lambda_i \| x_0 - x_i \|_{x_0}.
\end{aligned}
$$

In view of definition (7.3.2), we have $\langle \nabla F(x_0), x_i - x_0 \rangle \geq 0$. Hence, by Theorem 5.3.9, $\| x_i - x_0 \|_{x_0} \leq \nu + 2\sqrt{\nu} \leq 3\nu$ (recall that $\nu \geq 1$ by Lemma 5.4.1). Thus, we conclude that $\ell_k(x_0) \geq -3\nu S_k \| \nabla f(x_0) \|^*_{x_0}$. Using our observations and inequality (7.3.12), we obtain (7.3.16). □

Let us estimate now the rate of convergence of method (7.3.14) as applied to a specific problem class.

**Definition 7.3.1** We say that $f \in \mathscr{B}_M(P)$ if $\|\nabla f(x)\|_x^* \leq M$ for any $x \in P_0$.

For a function $f \in \mathscr{B}_M(P)$, we suggest the following values of parameters in (7.3.14):

$$\lambda_k = 1,\ k \geq 0, \qquad \beta_0 = \beta_1, \quad \beta_k = M \cdot \left(1 + \sqrt{\tfrac{k}{\nu}}\right),\ k \geq 1. \tag{7.3.19}$$

**Theorem 7.3.2** *Let problem (7.3.13) with $f \in \mathscr{B}_M(P)$ be solved by method (7.3.14) with parameters given by (7.3.19). Then for any $k \geq 0$ we have*

$$\tfrac{1}{S_k}\,\ell_k^\star \leq 2M \cdot \left(\sqrt{\tfrac{\nu}{k+1}} + \tfrac{\nu}{k+1}\right) \cdot \left(1 + \ln\left(2 + \tfrac{3}{2}\sqrt{\nu(k+1)}\right)\right). \tag{7.3.20}$$

*Proof* Define $\tau_k = \frac{1}{M}\beta_k > 1$. In view of the choice of parameters (7.3.19) and assumptions of the theorem, we have $S_k = k+1$, and

$$\begin{aligned} A_k &= \sum_{i=0}^{k} \beta_i \omega_* \left(\tfrac{\lambda_i}{\beta_i}\|\nabla f(x_i)\|_{x_i}^*\right) \;\leq\; M\sum_{i=0}^{k} \tau_i \omega_*\left(\tfrac{1}{\tau_i}\right) \;\leq\; \tfrac{1}{2}M\sum_{i=0}^{k}\tau_i\,\tfrac{\tau_i^{-2}}{1-\tau_i^{-1}} \\ &= \tfrac{1}{2}M\sum_{i=0}^{k}\tfrac{1}{\tau_i - 1} \;=\; \tfrac{\sqrt{\nu}}{2}M\left[1 + \sum_{i=1}^{k}\tfrac{1}{\sqrt{i}}\right] \;\leq\; \sqrt{\nu}M\left[\tfrac{1}{2} + \sqrt{k}\right]. \end{aligned} \tag{7.3.21}$$

(The last inequality can be easily justified by induction.) Furthermore,

$$\tfrac{S_k}{\beta_{k+1}}\|\nabla f(x_0)\|_{x_0}^* \leq \tfrac{k+1}{1+\sqrt{\frac{k+1}{\nu}}} \;\leq\; \sqrt{\nu(k+1)},$$

$$\tfrac{A_k}{\beta_{k+1}\nu} \leq \tfrac{\frac{1}{2}+\sqrt{k}}{\sqrt{\nu}+\sqrt{k+1}} \;\leq\; 1.$$

Thus, substituting the above estimates in inequality (7.3.16), we obtain

$$\begin{aligned} \tfrac{\ell_k^\star}{S_k} &\leq M\left[\tfrac{\sqrt{\nu}}{k+1}\left(\tfrac{1}{2}+\sqrt{k}\right) + \tfrac{\nu+\sqrt{\nu(k+1)}}{k+1}\left(1 + 2\ln\left(1+\sqrt{1+3\sqrt{\nu(k+1)}}\right)\right)\right] \\ &\leq 2M\cdot\left(\sqrt{\tfrac{\nu}{k+1}} + \tfrac{\nu}{k+1}\right)\cdot\left(1+\ln\left(2+\tfrac{3}{2}\sqrt{\nu(k+1)}\right)\right). \end{aligned}$$

In the last inequality we use the bound $\frac{\sqrt{\nu}}{k+1}\left(\frac{1}{2}+\sqrt{k}\right) \leq \sqrt{\frac{\nu}{k+1}} + \frac{\nu}{k+1}$. □

With parameters chosen by (7.3.19), the scheme of method (7.3.14) can be written in the following form:

$$x_{k+1} = \arg\max_{x\in P_0} \left\{ \tfrac{1}{k+1} \sum_{i=0}^{k} \langle \nabla f(x_i), x - x_i\rangle - M \tfrac{\sqrt{\nu}+\sqrt{k+1}}{\sqrt{\nu}(k+1)} [F(x) - F(x_0)] \right\}. \tag{7.3.22}$$

Since $f$ is a concave function,

$$\begin{aligned} \tfrac{1}{S_k} \ell_k^\star &= \tfrac{1}{S_k} \max_{y\in P} \sum_{i=0}^{k} \lambda_i \langle \nabla f(x_i), y - x_i\rangle \\ &\geq \tfrac{1}{S_k} \max_{y\in P} \sum_{i=0}^{k} \lambda_i [f(y) - f(x_i)] \;= f_\star - \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i f(x_i). \end{aligned}$$

Thus, the estimate (7.3.20) justifies the following rate of convergence for *primal* variables:

$$f_\star - \sum_{i=0}^{k} \tfrac{\lambda_i}{S_k} f(x_i) \leq 2M \cdot \left( \sqrt{\tfrac{\nu}{k+1}} + \tfrac{\nu}{k+1} \right) \cdot \left( 1 + \ln\left( 2 + \tfrac{3}{2}\sqrt{\nu(k+1)} \right) \right). \tag{7.3.23}$$

Note that the value $\ell_k^\star$ is computable. Hence, it can be used for terminating the process.

Let us show now that method (7.3.22) can also generate approximate solutions to the dual problem. For that, we need to employ the *internal structure* of our problem. Let us assume that it can be represented in a *saddle-point form*:

$$f(x) = \min_{w\in S} \Psi(x, w) \quad \to \quad \max_{x\in P}, \tag{7.3.24}$$

where $S \subset \mathbb{E}_1$ is a closed convex set, and the function $\Psi(x, w)$ is convex in $w \in S$ and concave and subdifferentiable in $x \in P$. Then, the dual problem is defined as

$$\begin{aligned} &\text{Find } f_\star = \min_{w\in S} \eta(w), \\ &\eta(w) = \max_{y\in P} \Psi(y, w). \end{aligned} \tag{7.3.25}$$

Since $P$ is bounded, the above problem is well defined. Without loss of generality, it is always possible to choose

$$\nabla f(x) = \nabla_1 \Psi(x, w(x)) \tag{7.3.26}$$

with some $w(x) \in \operatorname{Arg}\min_{w\in S} \Psi(x, w) \subseteq S$. Let us assume that $w(x)$ is computable for any $x \in P$.

**Lemma 7.3.3** *Define* $\bar{w}_k = \frac{1}{S_k}\sum\limits_{i=0}^{k} \lambda_i w(x_i)$, *and* $\bar{x}_k = \frac{1}{S_k}\sum\limits_{i=0}^{k} \lambda_i x_i$. *Then*

$$\eta(\bar{w}_k) - f(\bar{x}_k) \le \tfrac{1}{S_k}\,\ell_k^{\star}. \tag{7.3.27}$$

*Proof* Since $\Psi$ is concave in the first argument, for any $y \in P$ we have

$$\begin{aligned}\langle \nabla f(x_i), y - x_i\rangle &= \langle \nabla_1 \Psi(x_i, w(x_i)), y - x_i\rangle\\ &\ge \Psi(y, w(x_i)) - \Psi(x_i, w(x_i)) = \Psi(y, w(x_i)) - f(x_i).\end{aligned}$$

Hence,

$$\begin{aligned}\tfrac{1}{S_k}\,\ell_k^{\star} &= \tfrac{1}{S_k}\max_{y\in P}\sum_{i=0}^{k}\lambda_i\langle \nabla f(x_i), y - x_i\rangle \;\ge\; \tfrac{1}{S_k}\max_{y\in P}\sum_{i=0}^{k}\lambda_i[\Psi(y, w(x_i)) - f(x_i)]\\ &\ge \max_{y\in P}\Psi(y, \bar{w}_k) - \tfrac{1}{S_k}\sum_{i=0}^{k}\lambda_i f(x_i) \;=\; \eta(\bar{w}_k) - \tfrac{1}{S_k}\sum_{i=0}^{k}\lambda_i f(x_i)\\ &\ge \eta(\bar{w}_k) - f(\bar{x}_k). \qquad \square\end{aligned}$$

Thus, the scheme (7.3.22) can generate approximate primal-dual solutions:

$$\eta(\bar{w}_k) - f(\bar{x}_k) \le 2M\cdot\left(\sqrt{\tfrac{\nu}{k+1}} + \tfrac{\nu}{k+1}\right)\cdot\left(1 + \ln\left(2 + \tfrac{3}{2}\sqrt{\nu(k+1)}\right)\right). \tag{7.3.28}$$

### 7.3.3 Maximizing Positive Concave Functions

Consider now a convex optimization problem

$$\text{Find } \psi_\star \overset{\text{def}}{=} \max_x\{\psi(x) :\; x \in P\}, \tag{7.3.29}$$

where the set $P = \hat{P}\bigcap Q$ satisfies the assumptions introduced for problem (7.3.13). However, now we assume that the function $\psi$ is concave and *positive* on $\operatorname{int} Q$:

$$\psi(x) \;>\; 0, \quad \forall x \in \operatorname{int}\, Q. \tag{7.3.30}$$

**Lemma 7.3.4** *Let* $\psi$ *be concave and positive on int* $Q$. *Then for any* $x \in$ *int* $Q$ *we have*

$$\|\nabla\psi(x)\|_x^* \;\le\; \psi(x). \tag{7.3.31}$$

*Proof* Let us choose an arbitrary $x \in \operatorname{int} Q$ and $r \in [0, 1)$. Define

$$y = x - \tfrac{r}{\|\nabla\psi(x)\|_x^*}[\nabla^2 F(x)]^{-1}\nabla\psi(x).$$

In view of Item 1 of Theorem 5.1.5, $y \in \operatorname{int} Q$. Therefore,

$$0 \le \psi(y) \le \psi(x) + \langle \nabla\psi(x), y - x\rangle = \psi(x) - r\|\nabla\psi(x)\|_x^*.$$

Since $r$ is an arbitrary value from $[0, 1)$, we get (7.3.31). □

This result has an important corollary. Let us apply to the objective function of problem (7.3.29) a logarithmic transformation:

$$f(x) \overset{\text{def}}{=} \ln\psi(x). \tag{7.3.32}$$

**Lemma 7.3.5** *Let $\psi$ be concave and positive in the sense of (7.3.30). Then $f \in \mathscr{B}_1(Q)$, and it is concave on $Q$.*

*Proof* Indeed, it is well known that the logarithm of a concave function is a concave function too. It remains to note that $\nabla f(x) = \frac{1}{\psi(x)}\nabla\psi(x)$ and apply inequality (7.3.31). □

Thus, in order to solve problem (7.3.29), we can apply method (7.3.14) to problem (7.3.13) with the objective function defined by (7.3.32). The resulting optimization scheme is as follows:

$$x_{k+1} = \arg\max_{x\in P_0}\left\{\tfrac{1}{k+1}\sum_{i=0}^{k}\langle\tfrac{\nabla\psi(x_i)}{\psi(x_i)}, x - x_i\rangle - \tfrac{\sqrt{\nu}+\sqrt{k+1}}{\sqrt{\nu}(k+1)}[F(x) - F(x_0)]\right\}. \tag{7.3.33}$$

For scheme (7.3.33), we can guarantee a certain rate of convergence in *relative scale*.

**Theorem 7.3.3** *Let the sequence $\{x_k\}_{k=0}^{\infty}$ be generated by method (7.3.33) for problem (7.3.29). Then for any $k \ge 0$ we have*

$$\begin{aligned}
&\left[\prod_{i=0}^{k}\psi(x_i)\right]^{\frac{1}{k+1}}\\
&\ge \psi_\star\cdot\exp\left\{-2\left(\sqrt{\tfrac{\nu}{k+1}}+\tfrac{\nu}{k+1}\right)\left(1+\ln\left(2+\tfrac{3}{2}\sqrt{\nu(k+1)}\right)\right)\right\}\\
&\ge \psi_\star\cdot\left[1-2\left(\sqrt{\tfrac{\nu}{k+1}}+\tfrac{\nu}{k+1}\right)\left(1+\ln\left(2+\tfrac{3}{2}\sqrt{\nu(k+1)}\right)\right)\right].
\end{aligned} \tag{7.3.34}$$

*Proof* Indeed, we just apply method (7.3.22) to the function $f$ defined by (7.3.32). Since $f \in \mathscr{B}_1(Q) \subseteq \mathscr{B}_1(P)$, by (7.3.20) we conclude that

$$f_\star - \tfrac{1}{k+1}\sum_{i=0}^{k} f(x_i) \le \delta_k \overset{\text{def}}{=} 2\left(\sqrt{\tfrac{\nu}{k+1}} + \tfrac{\nu}{k+1}\right)\left(1 + \ln\left(2 + \tfrac{3}{2}\sqrt{\nu(k+1)}\right)\right).$$

Hence, $\left[\prod\limits_{i=0}^{k}\psi(x_i)\right]^{\frac{1}{k+1}} \ge \psi_\star \cdot e^{-\delta_k} \ \ge \ \psi_\star \cdot (1-\delta_k)$. This is exactly (7.3.34). □

Let us show how we can treat a problem dual to (7.3.29). For simplicity, assume that

$$\psi(x) \ = \ \min_{u\in\Omega} \Psi_0(u,x), \tag{7.3.35}$$

where $\Omega \subset \mathbb{E}_1$ is a closed convex set. In this case, condition (7.3.30) can be written as

$$\Psi_0(u,x) \ \ge 0, \quad u \in \Omega, \ x \in P. \tag{7.3.36}$$

Note that

$$\begin{aligned}
\max_{x\in P} \ln\psi(x) \ &= \ \max_{x\in P}\ \min_{\tau>0}\ \min_{u\in\Omega}\ [\,\tau\Psi_0(u,x) - \ln\tau - 1\,]\\
&= \ \max_{x\in P}\ \min_{\substack{v\in\tau\Omega,\\ \tau>0}} \left[\tau\Psi_0\left(\tfrac{1}{\tau}v, x\right) - \ln\tau - 1\right]\\
&\overset{(1.3.6)}{\le} \ \min_{\substack{v\in\tau\Omega,\\ \tau>0}} \left\{\eta(w) \equiv \eta(v,\tau) \overset{\text{def}}{=} -1 - \ln\tau + \tau\psi^\star\left(\tfrac{1}{\tau}v\right)\right\},
\end{aligned}$$

where $\psi^\star(u) = \max\limits_{x\in P}\Psi_0(u,x)$.

Denote by $u(x)$ a solution of the minimization problem (7.3.35). Then $w(x)$ is clearly defined as follows

$$w(x) = (v(x), \tau(x)), \quad v(x) = \tau(x)u(x), \quad \tau(x) = \tfrac{1}{\psi(x)}.$$

In accordance with Lemma 7.3.3, we can form $\bar{w}_k = (\bar{v}_k, \bar{\tau}_k)$ with

$$\bar{v}_k = \tfrac{1}{k+1}\sum_{i=0}^{k}\tfrac{u(x_i)}{\psi(x_i)}, \quad \bar{\tau}_k = \tfrac{1}{k+1}\sum_{i=0}^{k}\tfrac{1}{\psi(x_i)}.$$

Let $\bar{x}_k = \frac{1}{k+1} \sum\limits_{i=0}^{k} x_i$, and $\bar{u}_k = \frac{\bar{v}_k}{\bar{\tau}_k} = \sum\limits_{i=0}^{k} \frac{u(x_i)}{\psi(x_i)} \Big/ \left[ \sum\limits_{i=0}^{k} \frac{1}{\psi(x_i)} \right] \in \Omega$. Then, by (7.3.27) we get

$$\frac{1}{S_k} \ell_k^\star \geq \eta(\bar{w}_k) - \ln \psi(\bar{x}_k) \; = \; -1 - \ln \bar{\tau}_k + \bar{\tau}_k \psi^\star \left( \tfrac{1}{\bar{\tau}_k} \bar{v}_k \right) - \ln \psi(\bar{x}_k)$$

$$= -1 - \ln \bar{\tau}_k + \bar{\tau}_k \psi^\star \left( \bar{u}_k \right) - \ln \psi(\bar{x}_k) \; \geq \; \ln \frac{\psi^\star(\bar{u}_k)}{\psi(\bar{x}_k)}.$$

Hence,

$$\psi(\bar{x}_k) \geq \psi^\star(\bar{u}_k) \cdot \exp\left\{ -\frac{1}{S_k} \ell_k^\star \right\}. \tag{7.3.37}$$

Note that $\psi^\star(\bar{u}_k) \geq \psi_*$.

### *7.3.4 Applications*

In this section, we are going to consider examples of applications of method (7.3.33). It will be more convenient to use a slight modification of the usual notion of relative accuracy. We say that some value $\bar{\phi}$ is a $\delta$-approximation of the optimal value $\phi_\star > 0$ in *relative scale* if

$$\phi_\star \geq \bar{\phi} \geq \phi_\star \cdot e^{-\delta}, \quad \delta > 0.$$

In the complexity estimates, the short notation $\tilde{O}(\cdot)$ is used to indicate that some logarithmic factors are omitted. Since the rate of convergence (7.3.34) does not depend on the problem's data, our method is a so-called *fully polynomial-time approximation scheme*.

#### 7.3.4.1 The Fractional Covering Problem

Consider the following *fractional covering* problem:

$$\text{Find } \phi_\star \overset{\text{def}}{=} \min_y \{ \langle b, y \rangle : \; A^T y \geq c, \; y \geq 0 \in \mathbb{R}^m \}, \tag{7.3.38}$$

where $A = (a_1, \ldots, a_n)$ is an $(m \times n)$-matrix with non-negative coefficients, and vectors $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$ have positive coefficients. Define

$$\psi(y) = \min_{1 \leq i \leq n} \frac{1}{c(i)} \langle a_i, y \rangle.$$

Note that $\psi$ is concave and positively homogeneous of degree one. Therefore,

$$\begin{aligned}
\phi_\star &= \min_y \left\{ \tfrac{\langle b, y\rangle}{\psi(y)} : \; y \geq 0 \in \mathbb{R}^m \right\} \\
&= \left[ \max_y \left\{ \tfrac{\psi(y)}{\langle b, y\rangle} : \; y \geq 0 \in \mathbb{R}^m \right\} \right]^{-1} \\
&= \left[ \max_y \left\{ \psi(y) : \; \langle b, y\rangle = 1, \; y \geq 0 \in \mathbb{R}^m \right\} \right]^{-1}.
\end{aligned}$$

Thus, problem (7.3.38) can be written in the form (7.3.29) with $Q = \mathbb{R}^m_+$,

$$F(y) \; = \; -\sum_{j=1}^m \ln y^{(j)}, \quad \nu = m,$$

and $\hat{P} = \{y : \; \langle b, y\rangle = 1\}$. Hence, in accordance with the estimate (7.3.34) a $\delta$-approximation of $\phi_\star = \psi_\star^{-1}$ in relative scale can be found in $\tilde{O}(\frac{m}{\delta^2})$ iterations of method (7.3.33). Each iteration of the scheme needs $O(mn)$ operations to compute $\psi(y)$ and its subgradient, and essentially $O(m \ln m)$ operation to solve the auxiliary maximization problem in (7.3.33) (see Sect. A.2). Of course, this computational strategy is reasonable if $m << n$. Otherwise, it is better to solve the dual form of problem (7.3.38) by the smoothing technique (see Chap. 6).

#### 7.3.4.2 The Maximal Concurrent Flow Problem

Consider a network consisting of set of nodes $\mathscr{N}$, $|\mathscr{N}| = n$, and set of directed arcs

$$\mathscr{A} \; = \; \{\alpha = (i, j), \; i, j \in \mathscr{N}\}, \quad |\mathscr{A}| = m.$$

We assume that all arcs have bounded capacities. Formally, this means that the arc flow vector $f \in \mathbb{R}^m_+$ must satisfy the capacity constraint:

$$f \leq \bar{f}.$$

Let us introduce the set of origin-destination pairs

$$\mathscr{O}\mathscr{D} \; = \; \{(i, j), \; i, j \in \mathscr{N}\}.$$

Each pair $(i, j) \in \mathscr{O}\mathscr{D}$ generates for nodes $i$ and $j$ a directed flow $f_{i,j} \in \mathbb{R}^m_+$ of level $d_{i,j}$. Formally, this means that the vectors $f_{i,j}$ must satisfy the system of linear equations

$$B f_{i,j} \; = \; d_{i,j}(e_i - e_j), \quad (i, j) \in \mathscr{O}\mathscr{D},$$

where $B$ is the balance matrix of the network and $e_{(\cdot)}$ is the corresponding coordinate vectors in $\mathbb{R}^n$.

The *maximal concurrent flow* problem can be posed as follows:

$$\begin{aligned} \text{Find } \lambda_\star \overset{\text{def}}{=} \max_{\lambda, f_{i,j}} \{\lambda : Bf_{i,j} \; &= \; \lambda \cdot d_{i,j}(e_i - e_j), \\ f_{i,j} \; &\geq \; 0, \; (i,j) \in \mathscr{O}\mathscr{D}, \; \sum_{(i,j)\in\mathscr{O}\mathscr{D}} f_{i,j} \leq \bar{f} \;\}. \end{aligned} \tag{7.3.39}$$

Dualizing the flow capacity constraints by a vector of Lagrange multipliers $t \in \mathbb{R}^M_+$, we get the following dual problem:

$$\begin{aligned} \psi_\star \overset{\text{def}}{=} \lambda_\star^{-1} &= \max_t \{\psi(t) : \; \langle \bar{f}, t\rangle = 1, \; t \geq 0 \in \mathbb{R}^m\}, \\ \psi(t) &= \sum_{(i,j)\in\mathscr{O}\mathscr{D}} d_{i,j} \cdot SP_{i,j}(t), \end{aligned} \tag{7.3.40}$$

where the function $SP_{i,j}(t)$ is the *shortest path distance* between nodes $i$ and $j$ with respect to a non-negative arc travel time vector $t \in \mathbb{R}^m$.

Clearly the function $\psi$ in (7.3.40) satisfies all assumptions introduced for problem (7.3.29). Therefore (7.3.40) can be treated by method (7.3.33). In accordance with the estimate (7.3.34), a $\delta$-approximation of $\psi_\star$ in relative scale can be found in $\tilde{O}(\frac{m}{\delta^2})$ iterations. Each iteration of the scheme needs a computation of the shortest-path distances for all origin-destination pairs. The complexity of solving the auxiliary maximization problem in (7.3.33) is essentially $O(m \ln m)$ operations (see Sect. A.2). Note that we are also able to reconstruct the dual solutions (origin-destination flows) using the technique described at the end of Sect. 7.3.3.

#### 7.3.4.3 The Minimax Problem with Nonnegative Components

Consider the following minimax problem:

$$\text{Find } \psi_\star \overset{\text{def}}{=} \min_{x\in S} \max_{1\leq i\leq m} f_i(x), \tag{7.3.41}$$

where $S$ is a closed convex set and all functions $f_i(\cdot)$ are convex and non-negative on $S$. We assume that the function

$$\psi(y) = \min_{x\in S} \sum_{i=1}^{m} y^{(i)} f_i(x)$$

is well defined for any $y \geq 0 \in \mathbb{R}^m$. Moreover, let us assume that the values of this function and its subgradients are easily computable.

Then we can rewrite problem (7.3.41) in the dual form

$$\psi_\star \;=\; \max_y \left\{ \psi(y) : \; \langle \bar e_m, y\rangle = 1, \;\; y \geq 0 \in \mathbb{R}^m \right\}, \tag{7.3.42}$$

where $\bar e_m \in \mathbb{R}^m$ is the vector of all ones.

Note that (7.3.42) satisfies all assumption of problem (7.3.29). Therefore, in accordance with the estimate (7.3.34), a $\delta$-approximation of $\psi_\star$ in relative scale can be found by method (7.3.33) in $\tilde O\left(\frac{m}{\delta^2}\right)$ iterations. Each iteration of the scheme results in a minimization of a weighted sum of functions $f_i$ and the barrier function $F$.

#### 7.3.4.4 Semidefinite Relaxation of the Boolean Quadratic Problem

Consider the following maximization problem:

$$\text{Find } f_\star \overset{\mathrm{def}}{=} \max_x \{ \langle Ax, x\rangle : \; x^{(i)} = \pm 1, \; i = 1, \ldots, n\}, \tag{7.3.43}$$

where $A$ is a symmetric positive definite $(n \times n)$-matrix. It is well known that this problem is NP-hard. However, it appears that its optimal value can be approximated in polynomial time with a certain dimension-independent relative accuracy. Namely, define

$$\psi_\star \;=\; \min_y \{ \langle \bar e_n, y\rangle : \; D(y) \succeq A\}, \tag{7.3.44}$$

where $D(y)$ is a diagonal $(n \times n)$-matrix with vector $y$ on the diagonal. Then it can be proved that

$$\tfrac{2}{\pi}\psi_\star \leq f_\star \;\leq\; \psi_\star.$$

Usually the problem (7.3.44) is treated by Interior-Point Methods. However, note that quite often it is useless to compute an approximation to $\psi_\star$ with a high relative accuracy. Therefore it seems reasonable to solve it by a cheap gradient scheme.

Let us justify another representation for $\psi_\star$.

**Lemma 7.3.6** *Let $A = L^T L$. Then*

$$\psi_\star = \max_X \left\{ \psi(X) \overset{\mathrm{def}}{=} \left[ \sum_{i=1}^n \langle X q_i, q_i\rangle^{1/2} \right]^2 : \; \langle I_n, X\rangle_F = 1, \;\; X \succeq 0 \right\}, \tag{7.3.45}$$

*where $q_i$ are the columns of matrix $L$, $I_n$ is the identity matrix, and the scalar product in the space of symmetric matrices is defined in a natural way.*

*Proof* Indeed, since $A \succ 0$, we have

$$\begin{aligned}\psi_\star &= \min_u \left\{ \sum_{i=1}^n \tfrac{1}{u^{(i)}} : A^{-1} \succeq D(u) \right\}\\ &= \min_u \quad \max_{Y \succeq 0} \quad \left\{ \sum_{i=1}^n \tfrac{1}{u^{(i)}} + \langle Y, D(u) - A^{-1} \rangle_M \right\}\\ &= \max_{Y \succeq 0} \quad \min_u \quad \left\{ \sum_{i=1}^n \left( \tfrac{1}{u^{(i)}} + Y^{(i,i)}\, u^{(i)} \right) - \langle Y, A^{-1} \rangle_M \right\}.\end{aligned}$$

Thus, $\psi_\star = \max\limits_{Y \succeq 0} \left\{ 2 \sum\limits_{i=1}^n \left[ Y^{(i,i)} \right]^{1/2} - \langle Y, A^{-1} \rangle_M \right\}$. Maximizing the objective function in this problem along a fixed direction $Y \succeq 0$, we obtain

$$\psi_\star = \max_{Y \succeq 0} \left\{ \tfrac{1}{\langle Y, A^{-1} \rangle_M} \left[ \sum_{i=1}^n \left[ Y^{(i,i)} \right]^{1/2} \right]^2 \right\}.$$

Choosing in this problem new variables $X = L^{-T} Y L^{-1}$, we obtain representation (7.3.45). □

Note that the function $\psi$ in (7.3.45) is concave. Moreover, it is differentiable and positive at any $X \succ 0$. In our case, $Q$ is the cone of positive-semidefinite matrices with

$$F(X) \;=\; -\ln \det X, \qquad \nu = n.$$

Hence, (5.8) satisfies the conditions of the problem (7.3.29). Consequently, $\psi_\star$ can be approximated by (7.3.33) in $\tilde{O}\left(\frac{n}{\delta^2}\right)$ iterations, where $\delta$ is the desired relative accuracy. In our case, each iteration of the scheme (7.3.33) requires a representation of an $(n \times n)$-matrix in the form $UTU^T$, where $U$ is an orthogonal matrix, and the matrix $T$ is tri-diagonal. After that, we can apply the efficient search procedure described at the end of Sect. A.2.

## 7.3.5 *Online Optimization as an Alternative to Stochastic Programming*

### 7.3.5.1 A Decision-Making Process in an Uncertain Environment

Consider a repeatable decision-making process with uncertain income. Assume we have $N + 1$ periods of time, each of which corresponds to a *full* production cycle.

In the beginning of the $k$th period, we choose a production strategy

$$x_k \in P, \quad k = 0, \dots, N,$$

where the structure of $P$ satisfies the assumptions of Sect. 7.3.1. The results of different economic activities in this period are given by a *production function*

$$\psi_k(x) \geq 0, \quad x \in P.$$

The value $\psi_k(x)$ is equal to the *rate of growth* of the capital invested at the beginning of period $k$ in accordance with production strategy $x \in P$. The function $\psi_k(\cdot)$ becomes known only at the end of the period $k$. So, it can be used for choosing the production strategies of the next periods.

Assume for a moment that we know in advance all production functions

$$\psi_k(x), \quad k = 0, \dots, N.$$

However, for certain reasons, we are obliged to apply in all these periods the same strategy $x \in P$. In this case, of course, it is reasonable to use

$$x_N^\star \overset{\text{def}}{=} \arg\max_{x \in P} \prod_{k=0}^{N} \psi_k(x).$$

Then, the average efficiency of this static strategy is given by

$$\psi_N^\star = \left[ \prod_{k=0}^{N} \psi_k(x^\star) \right]^{\frac{1}{N+1}}.$$

However, usually the future is unknown. Instead, often we have the freedom to choose for each period $k$ a specific production strategy $x_k \in P$. Let us look at its possible efficiency.

Suppose we know a $\nu$-self-concordant barrier $F(\cdot)$ for the set $Q$. Then, we could apply the following variant of method (7.3.33):

$$x_{k+1} = \arg\max_{x \in P} \left\{ \frac{1}{k+1} \sum_{i=0}^{k} \langle \frac{\nabla \psi_i(x_i)}{\psi_i(x_i)}, x - x_i \rangle - \frac{\sqrt{\nu}+\sqrt{k+1}}{\sqrt{\nu}(k+1)} [F(x) - F(x_0)] \right\}. \tag{7.3.46}$$

In this case, after $N + 1$ periods, the average rate of growth is given by

$$\Psi_N \overset{\text{def}}{=} \left[ \prod_{k=0}^{N} \psi_k(x_k) \right]^{\frac{1}{N+1}}.$$

**Theorem 7.3.4** *For any $N \geq 0$ we have $\Psi_N \geq \psi_N^\star \cdot e^{-\delta_N}$ with*

$$\delta_N = 2\left(\sqrt{\tfrac{\nu}{N+1}} + \tfrac{\nu}{N+1}\right) \cdot \left(1 + \ln\left(2 + \tfrac{3}{2}\sqrt{\nu(N+1)}\right)\right) \;\to\; 0$$

*as $N \to \infty$.*

*Proof* The proof is very similar to the proofs of Theorems 7.3.1 and 7.3.2. Define

$$f_k(x) = \ln \psi_k(x), \quad f(x) = \tfrac{1}{N+1}\sum_{k=0}^{N} f_k(x), \quad s_k = \sum_{i=0}^{k} \nabla f_i(x_i) = \sum_{i=0}^{k} \tfrac{\nabla \psi_i(x_i)}{\psi_i(x_i)}.$$

Note that method (7.3.46) can be seen as an application of scheme (7.3.14), (7.3.19) to a changing objective function.

For any $k \geq 0$, we have

$$\begin{aligned} U_{\beta_{k+1}}(s_{k+1}) &\leq U_{\beta_k}(s_{k+1}) \\ &\overset{(7.3.5)}{\leq} U_{\beta_k}(s_k) + \langle \nabla f_k(x_k), u^\star_{\beta_k}(s_k) - x_0\rangle + \beta_k \omega_*\left(\tfrac{1}{\beta_k}\|\nabla f_k(x_k)\|^*_{x_k}\right) \\ &\overset{(7.3.31)}{\leq} U_{\beta_k}(s_k) + \langle \nabla f_k(x_k), u^\star_{\beta_k}(s_k) - x_0\rangle + \beta_k \omega_*\left(\tfrac{1}{\beta_k}\right). \end{aligned}$$

Since $U_{\beta_0}(0) = 0$, we conclude that

$$\begin{aligned} &\langle s_{N+1}, x_{N+1} - x_0\rangle - \beta_{N+1}[F(x_{N+1}) - F(x_0)] \\ &= U_{\beta_{N+1}}(s_{N+1}) \leq \sum_{i=0}^{N}\langle \nabla f_i(x_i), x_i - x_0\rangle + \sum_{i=0}^{N}\beta_i\omega_*\left(\tfrac{1}{\beta_i}\right) \qquad (7.3.47) \\ &\overset{(7.3.21)}{\leq} \sum_{i=0}^{N}\langle \nabla f_i(x_i), x_i - x_0\rangle + \sqrt{\nu}\left[\tfrac{1}{2} + \sqrt{N}\right]. \end{aligned}$$

In view of the first-order optimality condition for (7.3.3), for all $y \in P_0$ we have

$$\langle s_{N+1}, y - x_{N+1}\rangle \;\leq\; \beta_{N+1}\langle \nabla F(x_{N+1}), y - x_{N+1}\rangle. \qquad (7.3.48)$$

Therefore, using the concavity of all functions $f_i$, for any $y \in P$ we get

$$\begin{aligned} \ell_N(y) &\overset{\text{def}}{=} \sum_{i=0}^{N}\langle \nabla f_i(x_i), y - x_i\rangle \\ &\overset{(7.3.47)}{\leq} \langle s_{N+1}, y - x_{N+1}\rangle + \beta_{N+1}[F(x_{N+1}) - F(x_0)] + \sqrt{\nu}\left[\frac{1}{2} + \sqrt{N}\right] \end{aligned}$$

$$\begin{aligned}
&\overset{(7.3.48)}{\le} \quad \beta_{N+1}[F(x_{N+1}) + \langle \nabla F(x_{N+1}), y - x_{N+1}\rangle - F(x_0)] \\
&\qquad +\sqrt{\nu}\left[\frac{1}{2} + \sqrt{N}\right] \\
&\le \quad \beta_{N+1}[F(y) - F(x_0)] + \sqrt{\nu}\left[\frac{1}{2} + \sqrt{N}\right].
\end{aligned}$$

Hence, $\ell_N^\star(\beta_{N+1}) \le \sqrt{\nu}\left[\frac{1}{2} + \sqrt{N}\right]$. On the other hand, applying the same arguments as in the end of the proof of Theorem 7.3.1, we obtain

$$\begin{aligned}
\ell_N(x_0) &= \sum_{i=0}^{N} \langle \nabla f_i(x_i), x_0 - x_i\rangle \ \ge \ \sum_{i=0}^{N} \langle \nabla f_i(x_0), x_0 - x_i\rangle \\
&\ge -3\nu \cdot (N+1).
\end{aligned}$$

Thus, $\ell_N^\star(\beta_{N+1}) - \ell_N(x_0) \le \sqrt{\nu}\left(\frac{1}{2} + \sqrt{N}\right) + 3\nu \cdot (N+1)$. Since $\beta_{N+1} = 1 + \sqrt{\frac{N+1}{\nu}}$, by (7.3.12) we have:

$$\begin{aligned}
\frac{\ell_N^\star}{N+1} &\le \frac{\sqrt{\nu}}{N+1}\left(\frac{1}{2} + \sqrt{N}\right) \\
&\qquad + \frac{\nu + \sqrt{\nu(N+1)}}{N+1}\left[1 + 2\ln\left(1 + \sqrt{\frac{\sqrt{\nu}\left(\frac{1}{2} + \sqrt{N}\right) + 3\nu\cdot(N+1)}{\nu + \sqrt{\nu(N+1)}}}\ \right)\right] \\
&\le \frac{\sqrt{\nu}}{N+1}\left(\frac{1}{2} + \sqrt{N}\right) + \frac{\nu + \sqrt{\nu(N+1)}}{N+1}\left[1 + 2\ln\left(1 + \sqrt{1 + 3\sqrt{\nu(N+1)}}\ \right)\right] \\
&\le \delta_N
\end{aligned}$$

(see the arguments used at the end of the proof of Theorem 7.3.2). On the other hand,

$$\begin{aligned}
\frac{1}{N+1}\,\ell_N^\star &= \frac{1}{N+1}\max_{y\in P}\left\{\sum_{i=0}^{N} \langle \nabla f_i(x_i), y - x_i\rangle\right\} \ \ge \ \frac{1}{N+1}\max_{y\in P}\left\{\sum_{i=0}^{N} [f_i(y) - f_i(x_i)]\right\} \\
&= \ln \psi_N^\star - \ln \Psi_N. \qquad \Box
\end{aligned}$$

Let us now look at several applications of this theorem.

#### 7.3.5.2 Portfolio Management

Let $x \in \Delta_n$ be the structure of our portfolio. Denote by $c_k^{(i)} \geq 0$, $i = 1, \dots, n$, the growth coefficient for the price of stock $i$ during day $k \geq 0$. Then the optimal portfolio with *constant sharing* is defined as

$$x_N^\star = \arg\max_{x \in P} \prod_{k=0}^{N} \langle c_k, x \rangle, \quad \psi_N^\star = \left[ \prod_{k=0}^{N} \langle c_k, x_N^\star \rangle \right]^{1/(N+1)}.$$

For the set $Q = \mathbb{R}^n_+$, we can apply the standard $n$-self-concordant barrier

$$F(x) = -\sum_{i=1}^{n} \ln x^{(i)}.$$

Then, we can use the following variant of method (7.3.46):

$$x_{k+1} = \arg\max_{x \in P} \left\{ \frac{1}{k+1} \sum_{i=0}^{k} \frac{\langle c_i, x - x_i \rangle}{\langle c_i, x_i \rangle} - \frac{\sqrt{\nu} + \sqrt{k+1}}{\sqrt{\nu}(k+1)} [F(x) - F(x_0)] \right\}, \quad k \geq 0. \tag{7.3.49}$$

In this case, after $N + 1$ periods, the average rate of growth of our portfolio is given by

$$\Psi_N \overset{\text{def}}{=} \left[ \prod_{k=0}^{N} \langle c_k, x_k \rangle \right]^{\frac{1}{N+1}}.$$

In view of Theorem 7.3.4, we have $\Psi_N \geq \psi_N^\star \cdot e^{-\delta_N}$. Note that each step of the algorithm (7.3.49) is implementable in $O(n \ln n)$ arithmetic operations (see Sect. A.2).

#### 7.3.5.3 Processes with Full Production Cycles

Assume that in our economy there are $n$ elastic production processes. At the beginning of the $k$th period, we know the cost $a_k^{(i)} > 0$ of producing one unit of product $i$, $i = 1, \dots, n$. This cost is derived from the prices of raw materials, labor, equipment, etc. However, the price $b_k^{(i)} \geq 0$ of the unit of product $i$ becomes known only at the end of period $k$, when we sell it. It may depend on competition in the market, uncertain preferences of the consumers, etc. Denoting by $x^{(i)}$ the fraction of the capital invested in the process $i$, we come to the following model:

$$\begin{aligned} \psi_k(x) &= \sum_{i=1}^{n} \frac{b_k^{(i)}}{a_k^{(i)}} \cdot x^{(i)}, \\ x &= (x^{(1)}, \dots, x^{(n)})^T \in Q \overset{\text{def}}{=} \mathbb{R}^n_+, \\ \hat{P} &= \Delta_n. \end{aligned} \tag{7.3.50}$$

Then we can apply method (7.3.46) with

$$F(x) = -\sum_{i=1}^{n} \ln x^{(i)}, \quad \nu = n.$$

In this situation, the complexity of solving the auxiliary maximization problem in (7.3.46) is again $O(n \ln n)$ arithmetic operations (see Sect. A.2).

#### 7.3.5.4 Discussion

Theorem 7.3.4, being applied in an uncertain environment, delivers an *absolute and risk-free guarantee* for a certain level of efficiency of online optimization strategy (7.3.46). To obtain such a result, we do not need to introduce the standard machinery related to random events, risk measures, stochastic or robust optimization. Note that in Theorem 7.3.4 we compare the efficiency of a *dynamic* adjustment strategy with a *static* one. Hence, our arguments may not be too convincing. However, let us look at the standard one-stage stochastic programming problem

$$x_\star = \arg\max_{x \in P} \mathscr{E}_\zeta[f(x, \zeta)], \tag{7.3.51}$$

where $\mathscr{E}_\zeta[\cdot]$ denotes the expectation with respect to a random vector $\zeta$. The optimal strategy $x_\star$ must be *static* by its origin (otherwise, maximization of *expectation* does not make sense). At the same time, the quality of the model $f(x, \xi)$, constructed by an analysis of the *past*, can hardly be comparable with the quality of the static model based on *exact* knowledge of *future*. Thus, by transitivity, we can hope that our online adjustment strategy gives much better results than the standard Stochastic Programming approach. Of course, it can be applied only in the situations when the dynamic adjustments of the decision variables are implementable.

The main drawback of online optimization strategy (7.3.46) is its low rate of convergence. Therefore, it is efficient only for the processes where the average gain is big as compared to the number of iterations and the parameter of the barrier function. Interesting applications of this technique can be found most probably in long-run production planning and management than in stock market activity.

## 7.4 Optimization with Mixed Accuracy

(Strictly positive functions; The Quasi-Newton Method; Approximate solutions; Mixed accuracy.)

### 7.4.1 Strictly Positive Functions

In the previous chapters, we considered different approaches for finding approximate solutions of optimization problems with absolute and relative accuracy. In all cases, the type of desired accuracy was very important for the definition of the problem class, and consequently for the development of the corresponding numerical schemes. In this section, we proceed in a converse way. Firstly, we define a class of functions with favorable properties. Only after that will we try to understand what kind of theory can be developed for corresponding optimization problems.

Consider a closed convex function $f$ with $\mathrm{dom}\, f \subseteq \mathbb{R}^n$. Let $Q \subseteq \mathrm{dom}\, f$ be a closed convex set. We assume that $\partial f(x) \neq \emptyset$ for all $x \in Q$.

**Definition 7.4.1** A convex function $f$ is called *strictly positive* on $Q$ if for any $x, y$ from $Q$ and $g \in \partial f(x)$ we have

$$f(y) + f(x) + \langle g, y - x \rangle \geq 0. \tag{7.4.1}$$

Since $f$ is convex, this inequality can be written in a more appealing form:

$$f(y) \geq |f(x) + \langle g, y - x \rangle|, \quad x, y \in Q, \; g \in \partial f(x). \tag{7.4.2}$$

Clearly, strong positivity is an *affine-invariant* property.

**Lemma 7.4.1** *Let $f$ be strictly positive on $Q_x \subseteq \mathbb{R}^n$ and let $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^n$. Then the function $\phi(y) = f(Ay + b)$ is strictly positive on the set*

$$Q_y = \{ y \in \mathbb{R}^m : \; Ay + b \in Q_x \}.$$

*Proof* Indeed, in view of Lemma 3.1.11, for $x = Ay + b$ we have

$$g_y = A^T g_x \in \partial \phi(y), \quad \forall g_x \in \partial f(x).$$

For two arbitrary points $y_1, y_2 \in Q_y$ let $x_i = Ay_i + b$, $i = 1, 2$. Then

$$\phi(y_2) + \phi(y_1) + \langle g_{y_1}, y_2 - y_1 \rangle = f(x_2) + f(x_1) + \langle A^T g_{x_1}, y_1 - y_2 \rangle$$

$$= f(x_2) + f(x_1) + \langle g_{x_1}, x_1 - x_2 \rangle \overset{(7.4.1)}{\geq} 0. \quad \square$$

Let us give some important examples of strictly positive functions and mention their main properties.

1. Any positive constant is a strictly positive function.
2. Let us look at convex homogeneous functions of degree one.

**Lemma 7.4.2** *Let* $f(x) = \max\limits_{x \in S} \langle s, x\rangle$, *where the set* $S$ *is bounded, closed and centrally symmetric. Then the function* $f$ *is strictly positive.*

*Proof* For any $x \in \mathbb{R}^n$ and $g_x \in \partial f(x)$, we have $f(x) \overset{(3.1.40)}{=} \langle g_x, x\rangle$ and $-g_x \in S$. Therefore,

$$f(y) \overset{(3.1.23)}{\geq} \langle -g_x, y\rangle \overset{(3.1.40)}{=} -f(x) - \langle g_x, y - x\rangle. \qquad \square$$

3. Thus, the simplest nontrivial examples of strictly positive functions are *norms*.

Let us look now at operations preserving strong positivity.

**Lemma 7.4.3** *The class of strictly positive functions is a convex cone: if* $f_1$ *and* $f_2$ *are strictly positive on* $Q$, *and* $\alpha_1, \alpha_2 \geq 0$, *then* $f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x)$ *is strictly positive on* $Q$.

*Proof* Indeed, the characteristic inequality (7.4.1) is convex in $f$. □

**Lemma 7.4.4** *Let the functions* $f_1(\cdot)$ *and* $f_2(\cdot)$ *be strictly positive on* $Q$. *Then the function* $f(x) = \max\{f_1(x), f_2(x)\}$ *is also strictly positive.*

*Proof* Let us fix an arbitrary $x \in Q$. Assume that $f_1(x) > f_2(x)$. Then, for $y \in Q$ and $g_1 \in \partial f_1(x)$ we have

$$f(y) \geq f_1(y) \ \geq \ -f_1(x) - \langle g_1, y - x\rangle \ = \ -f(x) - \langle \nabla f(x), y - x\rangle.$$

The case $f_1(x) < f_2(x)$ and $f_1(x) = f_2(x)$ can be justified in a similar way (see Lemma 3.1.13). □

Thus, the functions below are strictly positive on $\mathbb{R}^n$:

$$f_1(x) = \sum_{i=1}^{m} \|A_i x - b_i\|, \quad f_2(x) \ = \ \max_{1 \leq i \leq m} \|A_i x - b_i\|,$$

where $A_i \in \mathbb{R}^{m \times n}$, and $b_i \in \mathbb{R}^m$, $i = 1 \ldots n$.

At the same time, the class of strictly positive functions contains functions with quite a general shape of epigraph. Let us fix a norm $\|\cdot\|$ for measuring distances in $\mathbb{R}^n$, and define the corresponding dual norm $\|\cdot\|_*$ in the standard way (7.1.3).

**Theorem 7.4.1** *Let the function $\phi$ be convex on $Q$ and all its subgradients be uniformly bounded:*

$$\|g_x\|_* \leq L, \quad x \in Q, \; g_x \in \partial f(x). \tag{7.4.3}$$

*Then the function $f(x) = \max\{\phi(x), L\|x\|\}$ is strictly positive on $Q$.*

*Proof* Let us fix an arbitrary $x \in Q$. Assume first, that $\phi(x) < L\|x\|$. Let us choose $s \in \mathbb{R}^n$ with $\|s\|_* = 1$, such that $\langle s, x\rangle = \|x\|$. Note that any $g_x \in \partial f(x)$ coincides with one of the vectors $Ls$ (see Lemma 3.1.15). Hence, for any $y \in E$ we have

$$f(y) + f(x) + \langle g_x, y - x\rangle \geq L\|y\| + L\|x\| + \langle Ls, y - x\rangle = L\|y\| + L\langle s, y\rangle \geq 0.$$

Further, if $\phi(x) > L\|x\|$, then $\partial f(x) = \partial\phi(x)$ and therefore for any $g_x \in \partial f(x)$ we have

$$\begin{aligned} f(y) + f(x) + \langle g_x, y - x\rangle &\geq L\|y\| + L\|x\| + \langle g_x, y - x\rangle \\ &\overset{(7.4.3)}{\geq} L\|y\| + L\|x\| - L\|y - x\| \geq 0. \end{aligned}$$

Finally, for the case $\phi(x) = L\|x\|$ we can apply a convex combination of the above inequalities. □

Using this result, we can endow a general minimization problem

$$\text{Find } \phi^* = \min_{x \in Q} \phi(x) \tag{7.4.4}$$

with a strictly positive objective function. Denote by $x^* \in Q$ its optimal solution.

**Corollary 7.4.1** *Let the function $\phi$ satisfy condition (7.4.3). Then for any $x_0 \in Q$ the function*

$$f(x) = \max\{\phi(x) - \phi(x_0) + 2LR, L\|x - x_0\|\}$$

*is strictly positive on $Q$. Moreover, for all $x$ with $\|x - x_0\| \leq R$ we have*

$$f(x) = \phi(x) - \phi(x_0) + 2LR. \tag{7.4.5}$$

*If $\|x_0 - x^*\| \leq R$, then problem (7.4.4) is equivalent to the problem*

$$f^* = \min_{x \in Q} f(x),$$

*with optimal value satisfying the following bounds:*

$$LR \leq f^* \leq 2LR. \tag{7.4.6}$$

*Proof* Indeed, $f$ is strictly positive on $Q$ in view of Theorem 7.4.1. If $\|x - x_0\| \leq R$, then

$$\phi(x) - \phi(x_0) + 2LR \overset{(7.4.3)}{\geq} 2LR - L\|x - x_0\| \ \geq \ L\|x - x_0\|,$$

and we obtain representation (7.4.5). Further, $f^* \leq f(x_0) = 2LR$. Finally,

$$f(x) \overset{(7.4.3)}{\geq} \max\{2LR - L\|x - x_0\|, L\|x - x_0\|\} \ \geq \ LR. \qquad \square$$

### 7.4.2 The Quasi-Newton Method

Consider the following minimization problem:

$$\min_{x \in Q} f(x), \tag{7.4.7}$$

where $Q$ is a closed convex set in $\mathbb{R}^n$, and the function $f$ is strictly positive on $Q$. Denote by $x^*$ the optimal solution of this problem. It will be convenient to work with another objective function:

$$\begin{aligned} \hat{f}(x) &= \tfrac{1}{2} f^2(x), \\ \hat{g}(x) &= f(x) \cdot g(x) \overset{\text{Lm } 3.1.8}{\in} \partial \hat{f}(x), \quad g(x) \in \partial f(x). \end{aligned} \tag{7.4.8}$$

Since the function $f$ is nonnegative, problem (7.4.7) can be rewritten in the equivalent form

$$\min_{x \in Q} \ \hat{f}(x). \tag{7.4.9}$$

The most unusual feature of the function $\hat{f}$ is the existence of nonlinear lower support functions.

**Lemma 7.4.5** *Let the function $f$ be strictly positive on $Q$. Then for any $x$ and $y \in Q$ we have*

$$\hat{f}(y) \geq \hat{f}(x) + \langle \hat{g}(x), y - x \rangle + \tfrac{1}{2} \langle g(x), y - x \rangle^2. \tag{7.4.10}$$

*Proof* Indeed,

$$\begin{aligned} \hat{f}(y) &\overset{(7.4.8)}{=} \tfrac{1}{2} f^2(y) \overset{(7.4.2)}{\geq} \tfrac{1}{2}[f(x) + \langle g(x), y - x \rangle]^2 \\ &\overset{(7.4.8)}{=} \hat{f}(x) + \langle \hat{g}(x), y - x \rangle + \tfrac{1}{2} \langle g(x), y - x \rangle^2. \qquad \square \end{aligned}$$

We will use inequality (7.4.10) in the framework of estimating sequences (see Sects. 2.2.1, 4.2.4, and 6.1.3). Let us fix a symmetric $n \times n$-matrix $G_0 \succ 0$, and a starting point $x_0 \in Q$. Define the primal and dual norms:

$$\|x\|_{G_0} = \langle G_0 x, x\rangle^{1/2}, \quad \|g\|^*_{G_0} = \langle g, G_0^{-1} g\rangle^{1/2}, \quad x, g \in \mathbb{R}^n.$$

We assume that $\|x_0 - x^*\|_{G_0} \leq R$. Define the initial function for the estimating sequence as follows:

$$\psi_0(x) = \tfrac{1}{2}\|x - x_0\|^2_{G_0}.$$

Let us fix an accuracy parameter $\delta \in (0, 1)$. Assuming that $g(x_k) \neq 0$, $k \geq 0$, define

$$a_k = \tfrac{\delta}{1-\delta} \cdot \tfrac{1}{(\|g(x_k)\|^*_{G_k})^2}, \quad A_k = \sum_{i=0}^{k-1} a_i, \quad k \geq 0. \tag{7.4.11}$$

Thus, $A_0 = 0$. For $k \geq 0$, consider the following process:

$$x_k = \arg\min_{x \in Q} \psi_k(x),$$

$$\psi_{k+1}(x) = \psi_k(x) + a_k \cdot \left[\hat{f}(x_k) + \langle \hat{g}(x_k), x - x_k\rangle + \tfrac{1}{2}\langle g(x_k), x - x_k\rangle^2\right]. \tag{7.4.12}$$

Clearly, in view of inequality (7.4.10), we have

$$\psi_k(x) \leq A_k \hat{f}(x) + \psi_0(x), \quad x \in Q. \tag{7.4.13}$$

On the other hand, $\psi_k(\cdot)$ is a quadratic function with Hessian $G_k \succ 0$ updated by the following rule

$$G_{k+1} = G_k + a_k \cdot g(x_k) g^T(x_k) \overset{(7.4.11)}{=} G_k + \tfrac{\delta}{1-\delta} \cdot \tfrac{g(x_k) g^T(x_k)}{(\|g(x_k)\|^*_{G_k})^2}, \quad k \geq 0. \tag{7.4.14}$$

Therefore, by the Sherman–Morrison–Woodbury rule, we have

$$G_{k+1}^{-1} = G_k^{-1} - \delta \cdot \tfrac{G_k^{-1} g(x_k) g^T(x_k) G_k^{-1}}{(\|g(x_k)\|^*_{G_k})^2}.$$

Thus, we conclude that

$$\begin{aligned}
\tfrac{1}{2}a_k^2(\|\hat{g}(x_k)\|^*_{G_{k+1}})^2 &\overset{(7.4.8)}{=} a_k^2 \cdot \hat{f}(x_k) \cdot (\|g(x_k)\|^*_{G_{k+1}})^2 \\
&= a_k^2 \cdot \hat{f}(x_k) \cdot (1-\delta) \cdot (\|g(x_k)\|^*_{G_k})^2 \qquad (7.4.15) \\
&\overset{(7.4.11)}{=} \delta \cdot a_k \cdot \hat{f}(x_k).
\end{aligned}$$

**Lemma 7.4.6** *For any* $k \geq 0$ *we have*

$$\psi_k^* \overset{\text{def}}{=} \min_{x \in Q} \psi_k(x) \;\geq\; (1-\delta)\sum_{i=0}^{k-1} a_i \hat{f}(x_i). \qquad (7.4.16)$$

*Proof* Let us prove inequality (7.4.16) by induction. For $k = 0$ it is true. Let us assume that it is true for some $k \geq 0$. Since $\psi_k(\cdot)$ is a quadratic function, it is strongly convex in the norm $\|\cdot\|_{G_k}$ with convexity parameter one. Thus, for any $x \in Q$ the first-order optimality condition implies

$$\psi_k(x) = \psi_k^* + \langle \psi_k'(x_k), x - x_k\rangle + \tfrac{1}{2}\|x - x_k\|^2_{G_k} \overset{(2.2.40)}{\geq} \psi_k^* + \tfrac{1}{2}\|x - x_k\|^2_{G_k}.$$

Therefore,

$$\begin{aligned}
\psi_{k+1}^* &\geq \psi_k^* + \min_{x\in Q}\Big\{\tfrac{1}{2}\|x-x_k\|^2_{G_k} + a_k[\hat{f}(x_k) + \langle \hat{g}(x_k), x - x_k\rangle \\
&\qquad + \tfrac{1}{2}\langle g(x_k), x - x_k\rangle^2]\Big\} \\
&\overset{(7.4.14)}{=} \psi_k^* + a_k\hat{f}(x_k) + \min_{x\in Q}\Big\{\tfrac{1}{2}\|x-x_k\|^2_{G_{k+1}} + a_k\langle \hat{g}(x_k), x - x_k\rangle\Big\} \\
&\geq \psi_k^* + a_k\hat{f}(x_k) - \tfrac{1}{2}a_k^2\|\hat{g}(x_k)\|^2_{G_{k+1}} \\
&\overset{(7.4.15)}{=} \psi_k^* + (1-\delta)\cdot a_k\hat{f}(x_k).
\end{aligned}$$

□

We can now estimate the rate of convergence of method (7.4.12). Define

$$x_k^* = \arg\min_x\{f(x) : \; x = x_0, \ldots, x_k\}, \quad \tilde{x}_k = \tfrac{1}{A_k}\sum_{i=0}^{k-1} a_i x_i.$$

**Theorem 7.4.2** *Let us assume that a strictly positive function* $f$ *has uniformly bounded subgradients:*

$$\|g(x)\|^*_{G_0} \leq L, \quad x \in Q. \qquad (7.4.17)$$

*Then, for any $k \geq 0$ we have*

$$(1-\delta)\hat{f}(x_k^*) \;\leq\; \hat{f}(x^*) + \frac{L^2 R^2}{2n\left[e^{\delta(k+1)/n} - 1\right]}. \tag{7.4.18}$$

*This estimate is also valid for the value $\hat{f}(\tilde{x}_{k+1})$.*

*Proof* In view of inequalities (7.4.13) and (7.4.16),

$$(1-\delta)\hat{f}(x_k^*) \leq \hat{f}(x^*) + \tfrac{1}{2A_{k+1}}\|x_0 - x^*\|_{G_0}^2.$$

Let us estimate the rate of growth of the coefficients $A_k$. Let $\bar{G}_k = G_0^{-1/2} G_k G_0^{-1/2}$, $k \geq 0$. Since $\det G_{k+1} \overset{(7.4.14)}{=} \frac{1}{1-\delta} \det G_k$, we have

$$\det \bar{G}_k = \tfrac{1}{(1-\delta)^k}, \quad k \geq 0. \tag{7.4.19}$$

It remains to note that

$$A_k \overset{(7.4.11)}{=} \sum_{i=0}^{k-1} a_i \overset{(7.4.17)}{\geq} \tfrac{1}{L^2} \sum_{i=0}^{k-1} a_i (\|g(x_i)\|_{G_0}^*)^2 \overset{(7.4.11)}{=} \tfrac{1}{L^2}\left[\text{Trace } \bar{G}_k - n\right]$$

$$\overset{(7.4.19)}{\geq} \tfrac{n}{L^2}\left[\tfrac{1}{(1-\delta)^{k/n}} - 1\right] \geq \tfrac{n}{L^2}\left[e^{\delta k/n} - 1\right]. \qquad \square$$

### 7.4.3 Interpretation of Approximate Solutions

Note that the quality of point $x_k^*$ as an approximate solution to problem (7.4.9) is characterized by inequality (7.4.18) in a nonstandard way. Let us introduce a new definition.

**Definition 7.4.2** We say that a point $\bar{x} \in Q$ is an approximate solution to problem (7.4.9) with *mixed* $(\epsilon, \delta)$-accuracy if

$$(1-\delta)\hat{f}(\bar{x}) \leq \hat{f}(x^*) + \epsilon.$$

In this definition, $\epsilon > 0$ serves as an absolute accuracy, and $\delta \in (0, 1)$ represents the relative accuracy of the point $\bar{x}$. Thus, in view of (7.4.18), the mixed $(\epsilon, \delta)$-accuracy can be reached by the Quasi-Newton Method (7.4.12) in

$$N_n(\epsilon, \delta) \overset{\text{def}}{=} \tfrac{n}{\delta} \ln\left(1 + \tfrac{L^2 R^2}{2n\,\epsilon}\right) \tag{7.4.20}$$

iterations.

Thus, it is not difficult to reach a high absolute accuracy. A high level of relative accuracy is much more expensive. Nevertheless, despite to the non-smoothness of the objective function in (7.4.9), the number of iterations of method (7.4.12) is proportional to $\frac{1}{\delta}$. This is, of course, a consequence of the finite dimension of the space of variables. Note that we have the following uniform upper bound for our estimate of the number of iterations:

$$N_n(\epsilon, \delta) < N_\infty(\epsilon, \delta) \overset{\mathrm{def}}{=} \frac{L^2R^2}{2\epsilon\delta}. \tag{7.4.21}$$

It is easy to see that the bound $N_n(\epsilon, \delta)$ is a monotonically increasing function of dimension $n$.

Let us discuss now the ability of method (7.4.12) to generate approximate solutions in the standard accuracy scales.

#### 7.4.3.1 Relative Accuracy

Consider our initial problem (7.4.7). Assume that our goal is to generate an approximate solution $\bar{x} \in Q$ to this problem with relative accuracy $\delta \in (0, \frac{1}{2})$:

$$f(\bar{x}) \le (1+\delta) f^*. \tag{7.4.22}$$

After $k$ iterations of method (7.4.12), we have

$$\begin{aligned} (1-\delta)(f(x_k^*) - f^*)f^* &\overset{(7.4.8)}{\le} (1-\delta)(\hat{f}(x_k^*) - \hat{f}(x^*)) \\ &\overset{(7.4.18)}{\le} \delta \hat{f}(x^*) + \frac{L^2R^2}{2n[e^{\delta(k+1)/n} - 1]}. \end{aligned} \tag{7.4.23}$$

In order to have the point $\bar{x} = x_k^*$ satisfying inequality (7.4.22), we need to ensure that the right-hand side of the latter inequality does not exceed $\delta(1-\delta)(f^*)^2$. Thus, for $\delta \in (0, \frac{1}{2})$ we need

$$k = R_n(\delta) \overset{\mathrm{def}}{=} \frac{n}{\delta} \ln\left(1 + \frac{L^2R^2}{n\delta(1-2\delta)(f^*)^2}\right) \tag{7.4.24}$$

iterations. Note that the main factor $\frac{n}{\delta}$ in this complexity bound does not depend on the data of the problem. Thus, for problem (7.4.7), we get a *fully polynomial-time approximation scheme*. Its dependence on $n$ is the same as that of optimal methods for nonsmooth convex minimization in finite dimensions. However, each iteration of method (7.4.12) is very simple, of the same order as in the Ellipsoid Method. Note that for problem (7.4.7) the Ellipsoid Method has complexity bound $O(n^2 \ln \frac{LR}{\delta f^*})$ iterations (see, Sect. 3.2.8). Thus, for a moderate relative accuracy, method (7.4.12) is faster. It is important that the right-hand side of inequality (7.4.24) is uniformly bounded as $n \to \infty$:

$$R_n(\delta) < R_\infty(\delta) \overset{\mathrm{def}}{=} \tfrac{L^2 R^2}{\delta^2(1-2\delta)(f^*)^2}.$$

#### 7.4.3.2 Absolute Accuracy

Consider now the general minimization problem (7.4.4), which we want to solve with absolute accuracy $\epsilon > 0$:

$$\phi(\bar{x}) \le \phi^* + \epsilon, \quad \bar{x} \in Q. \tag{7.4.25}$$

We assume that $\phi$ satisfies condition (7.4.3) and the constants $L$ and $R$ are known. Moreover, for the sake of simplicity, we assume that

$$\|x - x_0\| \le R \quad \forall x \in Q. \tag{7.4.26}$$

Defining now a new strictly positive objective function $f(\cdot)$ by equation (7.4.5), we get

$$f(x) = \phi(x) - \phi(x_0) + 2LR \quad \forall x \in Q. \tag{7.4.27}$$

Let us choose some $\delta \in (0, 1)$ and apply method (7.4.12) to the corresponding problem (7.4.7) (by solving (7.4.9), of course). After $k$ iterations of this scheme, we have

$$\phi(x_k^*) - \phi^* \overset{(7.4.27)}{=} f(x_k^*) - f^* \overset{(7.4.23)}{\le} \tfrac{\delta f^*}{2(1-\delta)} + \tfrac{L^2R^2}{2n\left[e^{\delta(k+1)/n}-1\right]\cdot(1-\delta)f^*}$$

$$\overset{(7.4.6)}{\le} LR\left[\tfrac{\delta}{1-\delta} + \tfrac{1}{2n\left[e^{\delta(k+1)/n}-1\right]\cdot(1-\delta)}\right].$$

Thus, to obtain accuracy $\epsilon > 0$, we can find $\delta = \delta(\epsilon)$ from the equation

$$\tfrac{\delta}{1-\delta} = \tfrac{\epsilon}{2LR} \quad \Rightarrow \quad \delta(\epsilon) = \tfrac{\epsilon}{\epsilon+2LR}.$$

Then, we need at most

$$\begin{aligned} k = T_n(\epsilon) &\overset{\mathrm{def}}{=} \tfrac{n}{\delta(\epsilon)} \ln\left(1 + \tfrac{LR}{n\epsilon(1-\delta(\epsilon))}\right) \\ &= n\left(1 + 2\tfrac{LR}{\epsilon}\right) \cdot \ln\left(1 + \tfrac{\epsilon+2LR}{2n\epsilon}\right) \end{aligned} \tag{7.4.28}$$

iterations of method (7.4.12). Note that

$$T_n(\epsilon) < T_\infty(\epsilon) \ = \ \tfrac{1}{2}\left(1 + 2\tfrac{LR}{\epsilon}\right)^2.$$

Thus, in finite dimensions the worst-case complexity bound of the Quasi-Newton Method (7.4.12) is always better than the bound of the standard subgradient scheme (see Sect. 3.2.3).

# Appendix A
# Solving Some Auxiliary Optimization Problems

## A.1 Newton's Method for Univariate Minimization

Let us show that Newton's Method is very efficient in finding the maximal root of increasing convex univariate functions. Consider a univariate function $f$ such that

$$f(\tau_*) = 0, \quad f(\tau) \ > \ 0, \text{ for } \tau > \tau_*, \tag{A.1.1}$$

and it is convex for $\tau \geq \tau_*$. Let us choose $\tau_0 > \tau_*$. Consider the following Newton process:

$$\tau_{k+1} = \tau_k - \tfrac{f(\tau_k)}{g_k}, \tag{A.1.2}$$

where $g_k \in \partial f(\tau_k)$. Thus, we do not assume $f$ to be differentiable for $\tau \geq \tau_*$.

**Theorem A.1.1** *Method (A.1.2) is well defined. For any $k \geq 0$ we have*

$$f(\tau_{k+1})g_{k+1} \leq \tfrac{1}{4} f(\tau_k) g_k. \tag{A.1.3}$$

*Thus, $f(x_k) \leq \left(\frac{1}{2}\right)^k g_0(\tau_0 - \tau_*)$.*

*Proof* Let $f_k = f(\tau_k)$. Let us assume that $f_k > 0$ for all $k \geq 0$. Since $f$ is convex for $\tau \geq \tau_*$, $0 = f(\tau_*) \geq f_k + g_k(\tau_* - \tau_k)$. Thus,

$$g_k(\tau_k - \tau_*) \geq f_k \ > \ 0. \tag{A.1.4}$$

This means that $g_k > 0$ and $\tau_{k+1} \in [\tau_*, \tau_k)$. In particular, we conclude that

$$\tau_k - \tau_* \leq \tau_0 - \tau_*. \tag{A.1.5}$$

Y. Nesterov, *Lectures on Convex Optimization*, Springer Optimization
and Its Applications 137, https://doi.org/10.1007/978-3-319-91578-4

Further, for any $k \geq 0$ we have:

$$f_k \geq f_{k+1} + g_{k+1}(\tau_k - \tau_{k+1}) \overset{(A.1.2)}{=} f_{k+1} + \frac{f_k g_{k+1}}{g_k}.$$

Thus, $1 \geq \frac{f_{k+1}}{f_k} + \frac{g_{k+1}}{g_k} \geq 2\sqrt{\frac{f_{k+1}g_{k+1}}{f_k g_k}}$, and this is (A.1.3). Finally, since $f$ is convex for $\tau \geq \tau_*$, we have

$$g_0 \overset{(A.1.4)}{\geq} \sqrt{\frac{f_0 g_0}{\tau_0 - \tau_*}} \overset{(A.1.3)}{\geq} 2^k \sqrt{\frac{f_k g_k}{\tau_0 - \tau_*}} \overset{(A.1.4)}{\geq} 2^k \sqrt{\frac{f_k^2}{(\tau_0 - \tau_*)(\tau_k - \tau_*)}}$$

$$\overset{(A.1.5)}{\geq} 2^k \frac{f_k}{\tau_0 - \tau_*}. \qquad \square$$

Thus, we have seen that method (A.1.2) has linear rate of convergence, which does not depend on the particular properties of the function $f$. Let us show that in a non-degenerate situation this method has local quadratic convergence.

**Theorem A.1.2** *Let a convex function $f$ be twice differentiable. Assume that it satisfies the conditions (A.1.1) and its second derivative increases for $\tau \geq \tau_*$. Then for any $k \geq 0$ we have*

$$f(\tau_{k+1}) \leq \frac{f''(\tau_k)}{2(f'(\tau_k))^2} \cdot f^2(\tau_k). \tag{A.1.6}$$

*If the root $\tau_*$ is non-degenerate:*

$$f'(\tau_*) > 0, \tag{A.1.7}$$

*then $f(\tau_{k+1}) \leq \frac{f''(\tau_0)}{2(f'(\tau_*))^2} \cdot f^2(\tau_k)$.*

*Proof* In view of conditions of the theorem, $f''(\tau) \leq f''(\tau_k)$ for all $\tau \in [\tau_{k+1}, \tau_k]$. Therefore,

$$\begin{aligned} f(\tau_{k+1}) &\leq f(\tau_k) + f'(\tau_k)(\tau_{k+1} - \tau_k) + \tfrac{1}{2} f''(\tau_k)(\tau_{k+1} - \tau_k)^2 \\ &\overset{(A.1.2)}{=} \tfrac{1}{2} f''(\tau_k) \frac{f^2(\tau_k)}{(f'(\tau_k))^2}. \end{aligned}$$

To prove the last statement, it remains to note that $f''(\tau_k) \leq f''(\tau_0)$ and $f'(\tau_k) \geq f'(\tau_*)$. $\square$

## A.2 Barrier Projection onto a Simplex

In the case $K = \mathbb{R}^n_+$, we can take

$$F(x) \;=\; -\sum_{i=1}^{n} \ln x^{(i)}, \quad \nu = n.$$

Consider $\hat{P} = \{x \in \mathbb{R}^n_+ : \; \langle \bar{e}_n, x\rangle = 1\}$. Then, at each iteration of method (7.3.14) we need to solve the following problem:

$$\phi^* \overset{\mathrm{def}}{=} \max_x \left\{ \langle s, x\rangle + \sum_{i=1}^{n} \ln x^{(i)} : \; \sum_{i=1}^{n} x^{(i)} = 1 \right\}. \tag{A.2.1}$$

Let us show that its complexity does not depend on the size of particular data (that is, the coefficients of the vector $s \in \mathbb{R}^n$).

Consider the following Lagrangian:

$$\mathscr{L}(x,\lambda) = \langle s, x\rangle + \sum_{i=1}^{n} \ln x^{(i)} + \lambda \cdot \left[ 1 - \sum_{i=1}^{n} x^{(i)} \right], \quad x \in \mathbb{R}^n, \;\lambda \in \mathbb{R}.$$

The dual function

$$\phi(\lambda) = \max_x \left\{ \mathscr{L}(x,\lambda) : \; \sum_{i=1}^{n} x^{(i)} = 1 \right\} \overset{\mathrm{def}}{=} \mathscr{L}(x(\lambda),\lambda)$$

is defined by the vector $x(\lambda) : \; x^{(i)}(\lambda) = \frac{1}{\lambda - s^{(i)}}$, $i = 1,\dots,n$. Thus,

$$\begin{aligned} \phi(\lambda) &= -n + \lambda - \sum_{i=1}^{n} \ln\left(\lambda - s^{(i)}\right), \\ \phi_* &= \min_\lambda \left\{ \phi(\lambda) : \; \lambda > \max_{1\le i\le n} s^{(i)} \right\}. \end{aligned} \tag{A.2.2}$$

Note that $\phi(\cdot)$ is a standard self-concordant function. Therefore we can apply to its minimization the intermediate Newton's Method (5.2.1), Item C), which converges quadratically starting from any $\lambda$ from the region

$$\mathscr{Q}(s) \;=\; \{\lambda : \; 4(\phi'(\lambda))^2 \le \phi''(\lambda)\}$$

(see Theorem 5.2.2). Let us show that the complexity of finding a starting point from this set does not depend on the initial data.

Consider the function $\psi(\lambda) \;=\; -\phi'(\lambda) \;=\; \sum_{i=1}^{n} \frac{1}{\lambda - s^{(i)}} - 1$. Clearly, the problem (A.2.2) is equivalent to finding the largest root $\lambda_*$ of the equation

$$\psi(\lambda) \;=\; 0. \tag{A.2.3}$$

Let $\lambda_0 = 1 + \max\limits_{1 \le i \le n} s^{(i)}$. Then $\psi(\lambda_0) \ge 0$ and therefore $\lambda_0 \le \lambda_*$. Consider the following process:

$$\lambda_{k+1} = \lambda_k - \tfrac{\psi(\lambda_k)}{\psi'(\lambda_k)}, \quad k \ge 0. \tag{A.2.4}$$

This is a standard Newton's method for solving the Eq. (A.2.3), which can be also interpreted as a Newton's method for the minimization problem (A.2.2).

**Lemma A.2.1** *For any $k \ge 0$ we have $(\phi'(\lambda_k))^2 \le n^7 \cdot \left(\frac{1}{16}\right)^k \phi''(\lambda_k)$.*

*Proof* Note that function $\psi$ is decreasing and strictly convex. Therefore, for any $k \ge 0$ we have

$$\lambda_k \; < \; \lambda_{k+1} \; < \lambda_*, \quad \psi'(\lambda_k) \; < 0 \;, \quad \psi(\lambda_k) \; > \; 0.$$

Since $\psi(\lambda_k) \; \ge \; \psi(\lambda_{k+1}) + \psi'(\lambda_{k+1})(\lambda_k - \lambda_{k+1}) \; = \; \psi(\lambda_{k+1}) + \frac{\psi'(\lambda_{k+1})}{\psi'(\lambda_k)}\psi(\lambda_k)$, we obtain[1]

$$1 \; \ge \; \tfrac{\psi(\lambda_{k+1})}{\psi(\lambda_k)} + \tfrac{\psi'(\lambda_{k+1})}{\psi'(\lambda_k)} \; \ge \; 2\sqrt{\tfrac{\psi(\lambda_{k+1})\psi'(\lambda_{k+1})}{\psi(\lambda_k)\psi'(\lambda_k)}}.$$

Thus, for any $k \ge 0$ we get

$$\phi''(\lambda_k) \cdot |\phi'(\lambda_k)| \le \left(\tfrac{1}{4}\right)^k \phi''(\lambda_0) \cdot |\phi'(\lambda_0)|. \tag{A.2.5}$$

Further, in view of the choice of $\lambda_0$ we have

$$|\phi'(\lambda_0)| = \psi(\lambda_0) \; = \; \sum_{i=1}^{n} \tfrac{1}{\lambda_0 - s^{(i)}} - 1 \; < \; n - 1,$$

$$\phi''(\lambda_0) = \sum_{i=1}^{n} \tfrac{1}{(\lambda_0 - s^{(i)})^2} \; \le \; n.$$

Finally, since $0 \le \psi(\lambda_k) = \sum\limits_{i=1}^{n} \frac{1}{\lambda_k - s^{(i)}} - 1$, we conclude that

$$\phi''(\lambda_k) = \sum_{i=1}^{n} \tfrac{1}{(\lambda_k - s^{(i)})^2} \; \ge \; \tfrac{1}{n}.$$

[1] We use the same arguments as in the proof of Theorem A.1.1, but for a decreasing univariate function.

Using these bounds in (A.2.5), we obtain

$$\tfrac{1}{\phi''(\lambda_k)}(\phi'(\lambda_k))^2 \le \left(\tfrac{1}{16}\right)^k \tfrac{(\phi''(\lambda_0))^2(\phi'(\lambda_0))^2}{(\phi''(\lambda_k))^3} \le \left(\tfrac{1}{16}\right)^k \cdot n^7. \qquad \Box$$

Comparing the statement of Lemma A.2.1 with the definition of $\mathscr{Q}(s)$, we conclude that the process (A.2.4) arrives at the region of quadratic convergence at most after

$$\left\lceil \tfrac{1}{4}(2 + 7\log_2 n) \right\rceil \tag{A.2.6}$$

iterations. Each such iteration takes $O(n)$ arithmetic operations.

A similar technique can be used for finding the barrier projection in the cone of positive-semidefinite matrices:

$$\max_X \{\langle S, X\rangle + \ln\det X : \ \langle I_n, X\rangle = 1\}.$$

The most straightforward strategy consists in finding an eigenvalue decomposition of the matrix $S$ and solving the problem (A.2.1) with $s$ being the spectrum of the matrix. In a more efficient strategy, we transform $S$ into tri-diagonal form by an orthogonal transformation, compute its maximal eigenvalue and apply the Newton's method to the corresponding dual function.

# Bibliographical Comments

In the past few decades, numerical methods for Convex Optimization have become widely studied in the monographic literature. The reader interested in engineering applications can benefit from the introductory exposition by Polyak [55], excellent course by Boyd and Vandenberghe [6], and lecture notes by Ben-Tal and Nemirovski [5]. Mathematical aspects are described in detail in the older lectures by A. Nemirovski (see [33] for the Internet version) and in the original versions of the theory for Interior-Point Methods by Renegar [57], Roos et al. [59], and Ye [63]. Recent theoretical highlights can be found in the monographs by Beck [3] and Bubeck [7]. In our book, we have tried to be more balanced, combining the comprehensive mathematical theory with many examples of practical applications, sometimes supported by numerical experiments.

## Chapter 1: Nonlinear Optimization

*Section 1.1* The complexity theory for black-box optimization schemes was developed in [34], where the reader can find different examples of resisting oracles and lower complexity bounds similar to that of Theorem 1.1.2.

*Sections 1.2 and 1.3* There exist several classical monographs [11, 12, 30, 53] treating different aspects of Nonlinear Optimization. For understanding Sequential Unconstrained Minimization, the best source is still [14]. Some facts in Sect. 1.3, related to conditions for zero duality gap, are probably new.

Y. Nesterov, *Lectures on Convex Optimization*, Springer Optimization
and Its Applications 137, https://doi.org/10.1007/978-3-319-91578-4

## Chapter 2: Smooth Convex Optimization

*Section 2.1* The original lower complexity bounds for smooth convex and strongly convex functions can be found in [34]. The proof used in this section was first published in [39].

*Section 2.2* Gradient mapping was introduced in [34]. The first optimal method for smooth and strongly convex functions was proposed in [35]. The constrained variant of this scheme is taken from [37]. However, the framework of estimating sequences was suggested for the first time in [39]. A discussion of different approaches for generating points with small norm of the gradient can be found in [48].

*Section 2.3* Optimal methods for discrete minimax problems were developed in [37]. The approach of Sect. 2.3.5 was first described in [39].

## Chapter 3: Nonsmooth Convex Optimization

*Section 3.1* A comprehensive treatment of different topics of Convex Analysis can be found in [24]. However, the classical monograph [58] is still very useful.

*Section 3.2* Lower complexity bounds for nonsmooth minimization problems can be found in [34]. The framework of Sect. 3.2.2 was suggested in [36]. For detailed bibliographical comments on the early history of Nonsmooth Minimization see [55, 56].

*Section 3.3* The example of a difficult function for Kelley's method is taken from [34]. The presentation of the Level Method in this section is close to [28].

## Chapter 4: Second-Order Methods

*Section 4.1* Starting from the seminal papers of Bennet [4] and Kantorovich [26], Newton's Method became an important tool for numerous applied problems. In the last 50 years, the number of different suggestions for improving the scheme is extremely large (see, for example, [11, 12, 15, 21, 29, 31]). The reader can consult an exhaustive bibliography in [11].

Most probably, the natural idea of using cubic regularization to improve the stability of the Newton scheme was first analyzed in [22]. However, the author was very sceptical about the complexity of solving the auxiliary minimization problem in the case of nonconvex quadratic approximation (and indeed, it can have an exponential number of local minima). As a result, this paper was never published. Twenty five years later, in an independent paper [52] this idea was checked again, and it was shown that this problem is solvable by standard techniques of Linear Algebra. The authors also developed global worst-case complexity bounds for different problem classes. This paper forms the basis of Sect. 4.1. The interested reader can also consult the complementary approach [8, 9], where cubic regularization is coupled with a line search along the gradient direction. However, note that this feature, though improving somewhat the numerical stability, forces the algorithm to stop at saddle points. A historical exposition of the development in this field with recent results, including lower complexity bounds for gradient norm minimization, can be found in [10].

*Section 4.2* This section is based on the paper [45].

*Section 4.3* This section is based on very recent and partially unpublished results. The first lower complexity bounds for second-order methods were obtained in [2]. At the same time, one of the second-order schemes in [32] achieves the rate of convergence $\tilde{O}\left(\frac{1}{k^{7/2}}\right)$, which is optimal. However, each iteration of this method needs an expensive search procedure based on additional calls of oracle. So, its practical efficiency is questionable.

In our presentation, we use a simpler derivation of the lower complexity bounds and a simpler conceptual version of the "optimal" second-order scheme, based on iteration of the Cubic Newton Method.

*Section 4.4* Methods for solving systems of nonlinear equations have attracted a lot of attention (see [11, 12, 53, 54]). However, we have not been able to find any global worst-case efficiency estimates for them in the literature. Our presentation follows the paper [43].

## Chapter 5: Polynomial-Time Interior-Point Methods

This chapter contains an adaptation of the main concepts from [51]. We added several useful inequalities and a slightly simplified presentation of the path-following scheme. We refer the reader to [5] for numerous applications of interior-point methods, and to [57, 59, 62] and [63] for a detailed treatment of different theoretical aspects.

*Section 5.1* In this section, we introduce the definition of a self-concordant function and study its properties. As compared with Section 4.1 in [39], we add Fenchel duality and the Implicit Function Theorem. The main novelty is an explicit treatment of the constant of self-concordance. However, most of the material can be found in [51].

*Section 5.2* In this new section, we analyze different methods for minimizing self-concordant functions. We propose a new step-size rule for the Newton scheme (*intermediate step*), which gives better constants for the path-following approach. Complexity estimates for a path-following scheme, as applied to a self-concordant function, were obtained only recently [13].

*Section 5.3* In this section we study the properties of a self-concordant barrier and give the complexity analysis for the path-following method. This is an adaptation of Section 4.2 in [39].

*Section 5.4* In this section, we give examples of self-concordant barriers and related applications. This is an extension of Section 4.3 in [39] by the results of [49].

## Chapter 6: The Primal-Dual Model of an Objective Function

This is the first attempt at presenting in the monographic literature the fast primal-dual gradient methods based on an explicit minimax model of the objective function. In the first three sections we present different aspects of the smoothing technique, following the papers [40, 41], and [42]. It seems that the Fast Gradient Method in the form of the Method of Similar Triangles (6.1.19) was published for the first time only recently (see [20]).

The last Sect. 6.4 is devoted to the new analysis of the old Conditional Gradient Method (or, the *Frank–Wolfe algorithm* [16, 18, 19, 23, 25]). Our presentation follows the paper [50], which is close in spirit to [17].

## Chapter 7: Optimization in Relative Scale

The presentation in this new chapter is based on the papers [44, 46], and [47]. Some examples of application were analyzed in [5], however, from the viewpoint of the applicability of Interior-Point Methods. Algorithms for computing the rounding ellipsoids are studied in [1, 27, 61], and in the recent book [60]. Constant quality of semidefinite relaxation for Boolean quadratic maximization with general matrix was proved in [38]. The material of Sect. 7.4 is new.

# References

1. K.M. Anstreicher, Ellipsoidal approximations of convex sets based on the volumetric barrier. CORE Discussion Paper 9745, 1997
2. Y. Arjevani, O. Shamir, R. Shiff, Oracle complexity of second-order methods for smooth convex optimization. arXiv:1705.07260v2 (2017)
3. A. Beck, *First-Order Methods in Optimization* (SIAM, Philadelphia, 2017)
4. A.A. Bennet, Newton's method in general analysis. Proc. Natl. Acad. Sci. U. S. A. **2**(10), 592–598 (1916)
5. A. Ben-Tal, A. Nemirovskii, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications* (SIAM, Philadelphia, 2001)
6. S. Boyd, L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004)
7. S. Bubeck, *Convex Optimization: Algorithms and Complexity* (Now Publishers, LP Breda, 2015). arXiv:1405.4980
8. C. Cartis, N.I.M. Gould, P.L. Toint, Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. Math. Program. **127**(2), 245–295 (2011)
9. C. Cartis, N.I.M. Gould, P.L. Toint, Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity. Math. Program. **130**(2), 295–319 (2011)
10. C. Cartis, N.I.M. Gould, P.L. Toint, How much patience do you have? a worst-case perspective on smooth nonconvex optimization. Optima **88**, 1–10 (2012)
11. A.B. Conn, N.I.M. Gould, P.L. Toint. *Trust Region Methods* (SIAM, Philadelphia, 2000)
12. J.E. Dennis, R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, 2nd edn. (SIAM, Philadelphia, 1996)
13. P. Dvurechensky, Yu. Nesterov, Global performance guarantees of second-order methods for unconstrained convex minimization, CORE Discussion Paper, 2018
14. A.V. Fiacco, G.P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques* (Wiley, New York, 1968)
15. R. Fletcher, *Practical Methods of Optimization, Vol. 1, Unconstrained Minimization* (Wiley, New York, 1980)
16. M. Frank, P. Wolfe, An algorithm for quadratic programming. Nav. Res. Logist. Q. **3**, 149–154 (1956)
17. R.M. Freund, P. Grigas, New analysis and results for the Frank–Wolfe method. Math. Program. **155**, 199–230 (2014). https://doi.org/10.1007/s10107-014-0841-6
18. D. Garber, E. Hazan, A linearly convergent conditional gradient algorithm with application to online and stochastic optimization. arXiv: 1301.4666v5 (2013)

Y. Nesterov, *Lectures on Convex Optimization*, Springer Optimization
and Its Applications 137, https://doi.org/10.1007/978-3-319-91578-4

19. D. Garber, E. Hazan, Faster rates for the Frank–Wolfe method over strongly convex sets. arXiv:1406.1305v2 (2015)
20. A. Gasnikov, Yu. Nesterov, Universal method for problems of stochastic composite minimization. Comput. Math. Math. Phys. **58**(1), 48–64 (2018)
21. S. Goldfeld, R. Quandt, H. Trotter, Maximization by quadratic hill climbing. Econometrica **34**, 541–551 (1966)
22. A. Griewank, The modification of Newton's method for unconstrained optimization by bounding cubic terms, Technical Report NA/12 (1981), Department of Applied Mathematics and Theoretical Physics, University of Cambridge, United Kingdom, 1981
23. Z. Harchaoui, A. Juditsky, A. Nemirovski, Conditional gradient algorithms for norm-regularized smooth convex optimization. Math. Program. **152**, 75–112 (2014). https://doi.org/10.1007/s10107-014-0778-9
24. J.-B. Hiriart-Urruty, C. Lemarechal, *Convex Analysis and Minimization Algorithms. Part 1*. A Series of Comprehensive Studies in Mathematics (Springer, Berlin, 1993)
25. M. Jaggi, Revisiting Frank–Wolfe: projection-free sparse convex optimization, in *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, Georgia (2013)
26. L.V. Kantorovich, Functional analysis and applied mathematics. Uspehi Mat. Nauk **3**(1), 89–185 (1948) (in Russian). Translated as N.B.S. Report 1509, Washington D.C., 1952
27. L.G. Khachiyan, Rounding of polytopes in the real number model of computation. Math. Oper. Res. **21**(2), 307–320 (1996)
28. C. Lemarechal, A. Nemirovskii, Yu. Nesterov, New variants of bundle methods. Math. Program. **69**, 111–148 (1995)
29. K. Levenberg. A method for the solution of certain problems in least squares. Q. Appl. Math. **2**, 164–168 (1944)
30. D.G. Luenberger, *Linear and Nonlinear Programming*, 2nd edn. (Addison Wesley, Boston, 1984)
31. D. Marquardt, An algorithm for least-squares estimation of nonlinear parameters. SIAM J. Appl. Math. **11**, 431–441 (1963)
32. R. Monteiro, B. Svaiter, An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. SIAM J. Optim. **23**(2), 1092–1125 (2013)
33. A. Nemirovski, Interior-point polynomial-time methods in convex programming (1996), https://www2.isye.gatech.edu/~nemirovs/LectIPM.pdf
34. A.S. Nemirovskij, D.B. Yudin, *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics (A Wiley-Interscience Publication/Wiley, New York, 1983)
35. Yu. Nesterov, A method for unconstrained convex minimization problem with the rate of convergence $O(\frac{1}{k^2})$. Doklady AN SSSR **269**, 543–547 (1983) (In Russian; translated as Soviet Math. Docl.)
36. Yu. Nesterov, Minimization methods for nonsmooth convex and quasiconvex functions. *Ekonomika i Mat. Metody* **11**(3), 519–531 (1984) (In Russian; translated in *MatEcon*.)
37. Yu. Nesterov, *Efficient Methods in Nonlinear Programming* (Radio i Sviaz, Moscow, 1989) (In Russian.)
38. Yu. Nesterov, Semidefinite relaxation and nonconvex quadratic optimization. Optim. Methods Softw. **9**, 141–160 (1998)
39. Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course* (Kluwer, Boston, 2004)
40. Yu. Nesterov, Smooth minimization of non-smooth functions. Math. Program. (A) **103**(1), 127–152 (2005)
41. Yu. Nesterov, Excessive gap technique in non-smooth convex minimizarion. SIAM J. Optim. **16** (1), 235–249 (2005)
42. Yu. Nesterov, Smoothing technique and its applications in semidefinite optimization. Math. Program. **110**(2), 245–259 (2007)

43. Yu. Nesterov, Modified Gauss–Newton scheme with worst-case guarantees for its global performance. Optim. Methods Softw. **22**(3), 469–483 (2007)
44. Yu. Nesterov, Rounding of convex sets and efficient gradient methods for linear programming problems. Optim. Methods Softw. **23**(1), 109–128 (2008)
45. Yu. Nesterov, Accelerating the cubic regularization of Newton's method on convex problems. Math. Program. **112**(1), 159–181 (2008)
46. Yu. Nesterov, Unconstrained convex minimization in relative scale. Math. Oper. Res. **34**(1), 180–193 (2009)
47. Yu. Nesterov, Barrier subgradient method. Math. Program. **127**(1), 31–56 (2011)
48. Yu. Nesterov, How to make the gradients small. Optima **88**, 10–11 (2012)
49. Yu. Nesterov, Towards non-symmetric conic optimization. Optim. Methods Softw. **27**(4–5), 893–918 (2012)
50. Yu. Nesterov, Complexity bounds for primal-dual methods minimizing the model of objective function. Math. Program. (2017). https://doi.org/10.1007/s10107-017-1188-6
51. Yu. Nesterov, A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming* (SIAM, Philadelphia, 1994)
52. Yu. Nesterov, B. Polyak, Cubic regularization of Newton's method and its global performance. Math. Program. **108**(1), 177–205 (2006)
53. J. Nocedal, S.J. Wright, *Numerical Optimization* (Springer, New York, 1999)
54. J. Ortega, W. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables* (Academic Press, New York, 1970)
55. B.T. Polyak, *Introduction to Optimization* (Optimization Software, Publications Division, New York, 1987)
56. B.T. Polyak, History of mathematical programming in the USSR: analyzing the phenomenon. Math. Program. **91**(3), 401–416 (2002)
57. J. Renegar, *A Mathematical View of Interior-Point Methods in Convex Optimization*. MPS-SIAM Series on Optimization (SIAM, Philadelphia, 2001)
58. R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1970)
59. C. Roos, T. Terlaky, J.-Ph. Vial, *Theory and Algorithms for Linear Optimization: An Interior Point Approach* (Wiley, Chichester, 1997)
60. M. Todd, *Minimum-Volume Ellipsoids: Theory and Algorithms*. MOS-SIAM Series on Optimization (SIAM, philadelphia, 2016)
61. M.J. Todd, E.A.Yildirim, On Khachiyan's algorithm for the computation of minimum volume enclosing ellipsoids, Technical Report, TR 1435, School of Operations Research and Industrial Engineering, Cornell University, 2005
62. S.J. Wright, *Primal-Dual Interior Point Methods* (SIAM, Philadelphia, 1996)
63. Y. Ye, *Interior Point Algorithms: Theory and Analysis* (Wiley, Hoboken, 1997)

# Index

Y. Nesterov, *Lectures on Convex Optimization*, Springer Optimization
and Its Applications 137, https://doi.org/10.1007/978-3-319-91578-4