An Introduction to Optimization:
Foundations and Fundamental Algorithms

Niclas Andréasson, Anton Evgrafov, and Michael Patriksson

Draft from February 14, 2005

Preface

The present book has been developed from course notes, continuously updated and used in optimization courses during the past several years at Chalmers University of Technology, Göteborg (Gothenburg), Sweden.

A note to the instructor: The book serves to provide lecture and exercise material in a first course on optimization for second to fourth year students at the university. (Computer exercises and projects are provided at course home pages on the local web site.) The book's focus lies on providing a solid basis for the analysis of optimization models and of candidate optimal solutions, especially for continuous optimization models. The main part of the mathematical material therefore concerns the analysis and algebra that underlie the workings of convexity and duality, and necessary/sufficient local/global optimality conditions for unconstrained and constrained optimization. Natural and most often classic algorithms are then developed from these principles, and their convergence characteristics analyzed. The book answers many more questions of the form "Why/why not?" than "How?".

This choice of focus is in contrast to books mainly providing numerical guidelines as to how these optimization problems should be solved. The number of algorithms for linear and nonlinear optimization problems—the two main topics covered in this book—are kept quite low; those that are discussed are considered classical, and serve to illustrate the basic principles for solving such classes of optimization problems and their links to the fundamental theory of optimality. Any course based on this book therefore should add project work on concrete optimization problems, including their modelling, analysis, solution, and interpretation.

A note to the student: The material assumes some familiarity with algebra, real analysis, and logic. In algebra, we assume an active knowledge of bases, norms, and matrix algebra and calculus. In real analysis, we assume an active knowledge of sequences, the basic topology of sets, real- and vector-valued functions and their calculus of differentiation. We also assume a familiarity with basic predicate logic, especially because proofs are based on it. A summary of the most important background topics is found in Chapter 2, which also serves as an introduction to the mathematical notation. The student is advised to refresh any unfamiliar or forgotten material of this chapter before reading the rest of the book.

A detailed road map of the contents of the book's chapters, and didactic statements as well, are provided at the end of Chapter 1. Each chapter ends with a selected number of exercises which either illustrate the theory and algorithms with numerical examples or develop the theory slightly further. In Appendix B solutions are given to most of them, in a few cases in detail. (Those exercises marked "exam" together with a date are examples of exam questions given in the course "Applied optimization" at Göteborg University and Chalmers University of Technology since 1997.) Sections with supplementary (but nevertheless important) material are marked with an asterisk.

In our work on this book we have benefited from discussions with Dr. Ann-Brith Strömberg, presently at the Fraunhofer–Chalmers Research Centre for Industrial Mathematics (FCC), Göteborg, and formerly at mathematics at Chalmers University of Technology. We thank the heads of undergraduate studies at mathematics, Göteborg University and Chalmers University of Technology, Jan-Erik Andersson and Sven Järner respectively, for reducing our teaching duties while preparing this book.

Göteborg, XX 2005

Anton Evgrafov
Niclas Andréasson
Michael Patriksson

Contents

I Introduction 1

1 Modelling and classification 3
- 1.1 Modelling of optimization problems . . . 3
  - 1.1.1 What does it mean to optimize? . . . 3
  - 1.1.2 Application examples . . . 5
- 1.2 A quick glance at optimization history . . . 9
- 1.3 Classification of optimization models . . . 11
- 1.4 Conventions . . . 14
- 1.5 Applications and modelling examples . . . 16
- 1.6 Defining the field . . . 16
- 1.7 Soft and hard constraints . . . 17
  - 1.7.1 Definitions . . . 17
  - 1.7.2 A derivation of the exterior penalty function . . . 18
- 1.8 A road map through the material . . . 19
- 1.9 On the background of this book and a didactics statement 25
- 1.10 Notes and further reading . . . 26
- 1.11 Exercises . . . 27

II Fundamentals 31

2 Analysis and algebra—A summary 33
- 2.1 Reductio ad absurdum . . . 33
- 2.2 Linear algebra . . . 34
- 2.3 Analysis . . . 37

3 Convex analysis 41
- 3.1 Convexity of sets . . . 41
- 3.2 Polyhedral theory . . . 42
  - 3.2.1 Convex hulls . . . 42

3.2.2 Polytopes 45
3.2.3 Polyhedra 47
3.2.4 The Separation Theorem and Farkas' Lemma . . . 52
3.3 Convex functions 57
3.4 Application: the projection of a vector onto a convex set . 66
3.5 Notes and further reading 67
3.6 Exercises 68

III Optimality Conditions 73

4 An introduction to optimality conditions 75
4.1 Local and global optimality 75
4.2 Existence of optimal solutions 78
4.3 Optimality in unconstrained optimization 84
4.4 Optimality for optimization over convex sets 88
4.5 Near-optimality in convex optimization 95
4.6 Applications 96
4.6.1 *Continuity of convex functions 96
4.6.2 The Separation Theorem 98
4.6.3 Euclidean projection 99
4.6.4 Fixed point theorems 100
4.7 Notes and further reading 106
4.8 Exercises 106

5 Optimality conditions 111
5.1 Relations between optimality conditions (OCs) and CQs at a glance 111
5.2 A note of caution 112
5.3 Geometric optimality conditions 114
5.4 The Fritz–John conditions 118
5.5 The Karush–Kuhn–Tucker conditions 124
5.6 Proper treatment of equality constraints 128
5.7 Constraint qualifications 130
5.7.1 Mangasarian–Fromovitz CQ (MFCQ) 130
5.7.2 Slater CQ 131
5.7.3 Linear independence CQ (LICQ) 131
5.7.4 Affine constraints 132
5.8 Sufficiency of KKT–conditions under convexity 132
5.9 Applications and examples 134
5.10 Notes and further reading 136
5.11 Exercises 137

6 Lagrangian duality **141**
6.1 The relaxation theorem 141
6.2 Lagrangian duality 142
6.2.1 Lagrangian relaxation and the dual problem 142
6.2.2 Global optimality conditions 146
6.2.3 Strong duality for convex programs 148
6.2.4 Strong duality for linear and quadratic programs . 153
6.3 Illustrative examples 155
6.3.1 Two numerical examples 155
6.3.2 An application to combinatorial optimization . . . 157
6.4 *Differentiability properties of the dual function 162
6.4.1 Sub-differentiability of convex functions 162
6.4.2 Differentiability of the Lagrangian dual function . 164
6.5 Subgradient optimization methods 167
6.5.1 Convex problems 168
6.5.2 Application to the Lagrangian dual problem 174
6.6 *Obtaining a primal solution 177
6.6.1 Differentiability at the optimal solution 177
6.6.2 Everett's Theorem 179
6.7 *Sensitivity analysis 180
6.7.1 Analysis for convex problems 180
6.7.2 Analysis for differentiable problems 182
6.8 Notes and further reading 184
6.9 Exercises 185

IV Linear Optimization **191**

7 Linear programming: An introduction **193**
7.1 The manufacturing problem 193
7.2 A linear programming model 194
7.3 Graphical solution 195
7.4 Sensitivity analysis 195
7.4.1 An increase in the number of large pieces available 196
7.4.2 An increase in the number of small pieces available 197
7.4.3 A decrease in the price of the tables 197
7.5 The dual of the manufacturing problem 198
7.5.1 A competitor 198
7.5.2 A dual problem 199
7.5.3 Interpretations of the dual optimal solution 199

8 Linear programming models 201
8.1 Linear programming modelling . . . 201
8.2 The geometry of linear programming . . . 206
8.2.1 Standard form . . . 207
8.2.2 Basic feasible solutions and the Representation Theorem . . . 210
8.2.3 Adjacent extreme points . . . 216
8.3 Notes and further reading . . . 218
8.4 Exercises . . . 218

9 The simplex method 221
9.1 The algorithm . . . 221
9.1.1 A BFS is known . . . 222
9.1.2 A BFS is not known: Phase I & II . . . 229
9.1.3 Alternative optimal solutions . . . 233
9.2 Termination . . . 233
9.3 Computational complexity . . . 234
9.4 Notes and further reading . . . 235
9.5 Exercises . . . 236

10 LP duality and sensitivity analysis 239
10.1 Introduction . . . 239
10.2 The linear programming dual . . . 240
10.2.1 Canonical form . . . 241
10.2.2 Constructing the dual . . . 241
10.3 Linear programming duality theory . . . 245
10.3.1 Weak and strong duality . . . 245
10.3.2 Complementary slackness . . . 249
10.4 The Dual Simplex method . . . 252
10.5 Sensitivity analysis . . . 256
10.5.1 Perturbations in the objective function . . . 256
10.5.2 Perturbations in the right-hand side coefficients . . . 257
10.6 Notes and further reading . . . 259
10.7 Exercises . . . 259

V Optimization over Convex Sets 265

11 Unconstrained optimization 267
11.1 Introduction . . . 267
11.2 Descent directions . . . 269
11.2.1 Basic ideas . . . 269
11.2.2 Less basic ideas . . . 272

11.3 Line searches 277
11.3.1 Introduction 277
11.3.2 Approximate line search strategies 278
11.4 Convergent algorithms 280
11.4.1 Basic convergence results 280
11.5 Finite termination criteria 283
11.6 A comment on non-differentiability 284
11.7 Trust region methods 286
11.8 Conjugate gradient methods 287
11.8.1 Conjugate directions 288
11.8.2 Conjugate direction methods 289
11.8.3 Generating conjugate directions 290
11.8.4 Conjugate gradient methods 291
11.8.5 Extension to non-quadratic problems 294
11.9 A quasi-Newton method 295
11.9.1 Introduction 295
11.9.2 The Davidon–Fletcher–Powell method 295
11.10 Convergence rates 298
11.11 Implicit functions 298
11.12 Notes and further reading 299
11.13 Exercises 300

12 Optimization over convex sets 307
12.1 Feasible direction methods 307
12.2 The Frank–Wolfe method 309
12.3 The simplicial decomposition method 312
12.4 The gradient projection algorithm 315
12.4.1 The algorithm and its convergence 315
12.4.2 A method for the projection problem 319
12.5 Notes and further reading 321
12.6 Exercises 322

VI Optimization over General Sets 325

13 Constrained optimization 327
13.1 Penalty methods 327
13.1.1 Exterior penalty methods 328
13.1.2 Interior penalty methods 332
13.1.3 Computational considerations 335
13.1.4 Applications and examples 335
13.2 Sequential quadratic programming 340
13.2.1 Introduction 340

13.2.2 A penalty-function based SQP algorithm 342
13.2.3 A numerical example on the MSQP algorithm . . . 346
13.2.4 On recent developments in SQP algorithms 347
13.3 A summary and comparison 348
13.4 Notes and further reading 349
13.5 Exercises 350

VII Appendices 353

A Introduction to LP using LEGO 355
A.1 The manufacturing problem 355
A.2 Solving the model using LEGO 355
A.3 Sensitivity analysis using LEGO 356
A.4 Geometric solution of the model 357
A.4.1 Geometric sensitivity analysis 357
A.5 The Simplex method and sensitivity analysis in LP 357
A.5.1 Slack, dependent, independent variables and extreme points 357
A.5.2 The simplex method 361
A.5.3 The example problem 362
A.5.4 Sensitivity analysis 363
A.6 Linear programming duality 364
A.6.1 A competitor 364
A.6.2 A dual problem 365
A.6.3 Interpretations of the dual optimal solution 365
A.6.4 Dual problems 366
A.6.5 Duality theory 366
A.6.6 Farkas' Lemma and the strong duality theorem . . 367

B Answers to the exercises 371
Chapter 1: Modelling and classification 371
Chapter 3: Convexity 374
Chapter 4: An introduction to optimality conditions 377
Chapter 5: Optimality conditions 378
Chapter 6: Lagrangian duality 380
Chapter 8: Linear programming models 381
Chapter 9: The simplex method 383
Chapter 10: LP duality and sensitivity analysis 384
Chapter 11: Unconstrained optimization 386
Chapter 12: Optimization over convex sets 387
Chapter 13: Constrained optimization 388

References **391**

Index **401**

Part I

Introduction

Modelling and classification

I

1.1 Modelling of optimization problems

1.1.1 What does it mean to optimize?

The word "optimum" is Latin, and means "the ultimate ideal;" similarly, "optimus" means "the best." Therefore, to *optimize* refers to trying to bring whatever we are dealing with towards its ultimate state. Let us take a closer look at what that means in terms of an example, and at the same time bring the definition of the term *optimization* forward, as the scientific field understands and uses it.

Example 1.1 (a staff planning problem) Consider a hospital ward which operates 24 hours a day. At different times of day, the staff requirement differs. Table 1.1 shows the demand for reserve wardens during six work shifts.

Shift	1	2	3	4	5	6
Hours	0–4	4–8	8–12	12–16	16–20	20–24
Demand	8	10	12	10	8	6

Table 1.1: Staff requirements at a hospital ward.

Each member of staff works in 8 hour shifts. The goal is to fulfill the demand with the least total number of reserve wardens. ■

Consider now the following interpretation of the term "to optimize:"

To optimize = to do something as well as is possible.

We utilize this description to identify the mathematical problem associated with Example 1.1; in other words, we create a *mathematical model* of the above problem.

Do something We identify, in the decision problem, activities which we can control and influence. Each such activity is associated with a *variable* whose value (or, activity level) is to be decided upon (that is, optimized). The remaining quantities are constants in the problem.

Well How good a vector of activity levels is is measured by a real-valued function of the variable values. This quantity is to be given a highest or lowest value, that is, we minimize or maximize, depending on our goal; this defines the *objective function*.

Possible Normally, the activity levels cannot be arbitrarily large, since an activity often is associated with the utilization of resources (time, money, raw materials, labour, etcetera) that are limited; there may also be requirements of a least activity level, resulting from a demand. Some variables must also fulfill technical/logical restrictions, and/or relationships among themselves. The former can be associated with a variable necessarily being integer-valued or non-negative, by definition. The latter is the case when products are blended, a task is performed for several types of products, or a process requires the input from more than one source. These restrictions on activities form *constraints* on the possible choices of the variable values.

Looking again at the problem described in Example 1.1, this is then our declaration of a mathematical model thereof:

Variables We define

$$\begin{aligned} x_j &:= \text{number of reserve wardens whose first shift is } j, \\ &\qquad j = 1, 2, \dots, 6. \end{aligned}$$

Objective function We wish to minimize the total number of reserve wardens, that is, the objective function, which we call f, is to

$$\text{minimize } f(\boldsymbol{x}) := x_1 + x_2 + \cdots + x_6 = \sum_{j=1}^{6} x_j.$$

Constraints There are two types of constraints:

Demand The demand for wardens during the different shifts can be written as the following inequality constraints:

$$\begin{aligned}
x_6 + x_1 &\geq 8, \\
x_1 + x_2 &\geq 10, \\
x_2 + x_3 &\geq 12, \\
x_3 + x_4 &\geq 10, \\
x_4 + x_5 &\geq 8, \\
x_5 + x_6 &\geq 6.
\end{aligned}$$

Logical There are two physical/logical constraints:

Sign $x_j \geq 0, \qquad j = 1, \dots, 6.$

Integer x_j integer, $\qquad j = 1, \dots, 6.$

Summarizing, we have defined our first mathematical optimization model, namely, that to

$$\begin{aligned}
\underset{\boldsymbol{x}}{\text{minimize}} \quad & f(\boldsymbol{x}) := \sum_{j=1}^{6} x_j, \\
\text{subject to} \quad & x_1 + x_6 \geq 8, && \text{(last shift: 1)} \\
& x_1 + x_2 \geq 10, && \text{(last shift: 2)} \\
& x_2 + x_3 \geq 12, && \text{(last shift: 3)} \\
& x_3 + x_4 \geq 10, && \text{(last shift: 4)} \\
& x_4 + x_5 \geq 8, && \text{(last shift: 5)} \\
& x_5 + x_6 \geq 6, && \text{(last shift: 6)} \\
& x_j \geq 0, && j = 1, \dots, 6, \\
& x_j \text{ integer}, && j = 1, \dots, 6.
\end{aligned}$$

This problem has an *optimal solution*, which we denote by $\boldsymbol{x}^*$, that is, a vector of decision variable values which gives the objective function its minimal value among the *feasible solutions* (that is, the vectors $\boldsymbol{x}$ that satisfy all the constraints). In fact, the problem has at least two optimal solutions: $\boldsymbol{x}^* = (4, 6, 6, 4, 4, 4)^{\mathrm{T}}$ and $\boldsymbol{x}^* = (8, 2, 10, 0, 8, 0)^{\mathrm{T}}$; the *optimal value* is $f(\boldsymbol{x}^*) = 28$. (The reader is asked to verify that they are indeed optimal.)

1.1.2 Application examples

The above model is of course a crude simplification of any real application. In practice, we would have to add requirements on the individual's competence as well as other more detailed restrictions, the planning horizon is usually longer, employment rules and other conditions apply, etcetera, which all contribute to a more complex model. We mention a few successful applications of staffing problems below.

Example 1.2 (applications of staffing optimization problems) (a) It has been reported that a 1990 staffing problem application for the Montreal municipality bus company, employing 3,000 bus drivers and 1,000 metro drivers and ticket salespersons and guards, saved some 4 million Canadian dollars per year.

(b) Together with the San Francisco police department a group of operations research scientists developed in 1989 a planning tool based on a heuristic solution of the staff planning and police vehicle allocation problem. It has been reported that it gave a 20% faster planning and savings in the order of 11 million US dollars per year.

(c) In an application from 1986, scientists collaborating with United Airlines considered their crew scheduling problem. This is a very complex problem, where the time horizon is long (typically, 30 minute intervals during 7 days), and the constraints that define a feasible pattern of allocating staff to airplanes are defined by, among others, complicated work regulations. The savings reported then was 6 million US dollars per year. The company Carmen Systems AB in Gothenburg develops and markets such a tool; buyers include American Airlines, Lufthansa, SAS, and SJ; this company has one of the largest concentrations of optimizers in Sweden. ∎

Remark 1.3 (on the complexity of the variable definition) The variables x_j defined in Example 1.1 are *decision variables*; we say that, since the selection of the values of these variables are immediately connected to the decisions to be made in the decision problem, and they also contain, within their very definition, a substantial amount of information about the problem at hand (such as shifts being eight hours long).

In the application examples discussed in Example 1.2 the variable definitions are much more complex than in our simple example. A typical decision variable arising in a crew scheduling problem is associated with a specific staff member, his/her home base, information about the crew team he/she works with, a current position in time and space, a flight leg specified by flight number(s), additional information about the staff member's previous work schedule and work contract, and so on. The number of possible combinations of work schedules for a given staff member is nowadays so huge that not all variables in a crew scheduling problem can even be defined! (That is, the complete problem we wish to solve cannot be written down.) The philosophy in solving a crew scheduling problem is instead to algorithmically *generate* variables that one believes may receive a non-zero optimal value, and most of the computational effort lies in defining and solving good variable generation problems, whose result is (part of) a feasible work schedule for given staff member. The term *column generation* is the operations researcher's name for this process of generating variables in a decision problem. ■

Remark 1.4 (non-decision variables) Not all variables in a mathematical optimization model are decision variables:

In linear programming, we will utilize *slack variables* whose role is to take on the difference between the left-hand and the right-hand side of an inequality constraint; the slack variable thereby aids in the transformation of the inequality constraint to an equality constraint, which is more appropriate to work with in linear programming.

Other variables can be introduced into a mathematical model simply in order to make the model more easy to state or interpret, or to improve upon the properties of the model. As an example of the latter, consider the following simple problem: we wish to minimize over $\mathbb{R}$ the special one-variable function $f(x) := \text{maximum}\,\{x^2, x+2\}$. (Plot the function to see where the optimum is.) This is an example of a non-differentiable function: at $x = -2$, for example, both the functions $f_1(x) := x^2$ and $f_2(x) := x+2$ define the value of the function f, but they have different derivatives there. One way to turn this problem into a differentiable one is by introducing an additional variable. We let z take on the value of the largest of $f_1(x)$ and $f_2(x)$ for a given value of x, and instead write the problem as that to minimize z, subject to $z \in \mathbb{R}$, $x \in \mathbb{R}$, and the additional constraints that $x^2 \leq z$ and $x + 2 \leq z$. Convince yourself that this transformation is equivalent to the original problem in terms of the set of optimal solutions in x, and that the transformed problem is differentiable. ■

Figure 1.1 illustrates several issues in the modelling process, which are forthwith discussed.

The decision problem faced in the "fluffy" reality is turned into an optimization model, through a process with several stages. By communicating with those who have raised the issue of solving the problem in the first place, one reaches an understanding about the problem to be solved. In order to identify and describe the components of a mathematical model which is also tractable, it is often necessary to simplify and also limit the problem somewhat, and to quantify any remaining qualitative statements.

The modelling process does not come without difficulties. The communication can often be difficult, simply because the two parties speak different languages in terms of describing the problem. The optimization

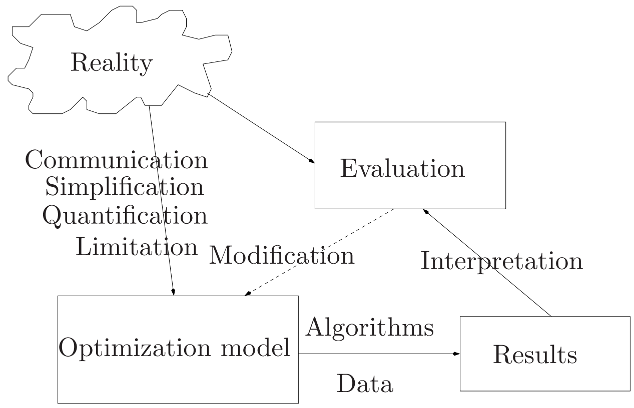


Figure 1.1: Flow chart of the modelling process

problem quite often have uncertainties in the data, which moreover are not always easy to collect or to quantify. Perhaps the uncertainties are there for a purpose (such as in financial decision problems), but it may be that the data is uncertain because not enough effort has been put into providing a good enough accuracy. Further, there is often a conflict between problem solvability and problem realism.

The problem actually solved through the use of an optimization methodology must be supplied with data, providing model constants and parameters in functions describing the objective function and perhaps also some of the constraints. For this optimization problem, an optimization algorithm then yields a result in the form of an optimal value and/or optimal solution, if an optimal solution exists. This result is then interpreted and evaluated, which may lead to alterations of the model, and certainly to questions regarding the applicability of the optimal solution. The optimization model can also be altered slightly on purpose in order to answer "what if?" type questions, for example sensitivity analysis questions concerning the effect of small variations in data.

The final problems that we will mention come at this stage: it is crucial that the interpretation of the result makes sense to those who wants to use the solution, and, finally, it must be possible to transfer the solution back into the "fluffy" world where the problem came from.

The art of forming *good* optimization models is as much an art as a science, and an optimization course can only really cover the latter. On the other hand, this part of the modelling process should not be glossed over; it is often possible to construct more than one form of an mathematical model that represents the same problem equally accurately, and the computational complexity can differ substantially between them. Forming a good model is in fact as crucial to the success of the application as the modelling exercise itself.

Optimization problems can be grouped together in classes, according to their properties. According to this classification, the staffing problem is a *linear integer optimization problem*. In Section 1.3 we present the major distinguishing factors between different problem classes.

1.2 A quick glance at optimization history

At Chalmers, the courses in optimization are mainly given at the mathematics department. "Mainly" is the important word here, because courses that have a substantial content of optimization theory and/or methodology can be found also at other departments, such as computer science, the mechanical, industrial and chemical engineering departments, and at the Gothenburg School of Economics. The reason is that optimization is so broad in its applications.

From the mathematical standpoint, optimization, or *mathematical programming* as it is sometimes called, rests on several legs: analysis, topology, algebra, discrete mathematics, etcetera, build the foundation of the theory, and applied mathematics subjects such as numerical analysis and mathematical parts of computer science build the bridge to the algorithmic side of the subject. On the other side, then, with optimization we solve problems in a huge variety of areas, in the technical, natural, life and engineering sciences, and in economics.

Before moving on, we would just like to point out that the term "program" has nothing to do with "computer program;" a program is understood to be a "decision program," that is, a strategy or decision rule. A "mathematical program" therefore is a mathematical problem designed to produce a decision program.

The history of optimization is also very long. Many very often geometrical or mechanical problems (and quite often related to warfare!) that Archimedes, Euclid, Heron, and other masters from antiquity formulated and also solved, are optimization problems. For example, we mention the problem of maximizing the volume of a closed three-dimensional object (such as a sphere or a cylinder) built from a two-dimensional sheet of metal with a given area.

The masters of two millenia later, like Bernoulli, Lagrange, Euler, and Weierstrass developed variational calculus, studying problems in applied physics (and still often with a mind towards warfare!) such as how to find the best trajectory for a flying object.

The notion of *optimality* and especially how to *characterize* an optimal solution, began to be developed at the same time. Characterizations of various forms of optimal solutions are indeed a crucial part of any basic optimization course.

The scientific subject *operations research* refers to the study of decision problems regarding operations, in the sense of controlling complex systems and phenomena. The term was coined in the 1940s at the height of World War 2 (WW2), when the US and British military commands hired scientists from several disciplines in order to try to solve complex problems regarding the best way to construct convoys in order to avoid, or protect the cargo ships from, enemy (read: German) submarines, how to best cover the British isles with radar equipment given the scarce availability of radar systems, and so on. The multi-disciplinarity of these questions, and the common topic of maximizing or minimizing some objective subject to constraints, can be seen as being the defining moment of the scientific field. A better term than operations research is *decision science*, which better reflects the scope of the problems that can be, and are, attacked using optimization methods.

Among the scientists that took part in the WW2 effort in the US and Great Britain, some were the great pioneers in placing optimization on the map after WW2. Among them, we find several researchers in mathematics, physics, and economics, who contributed greatly to the foundations of the field as we now know it. We mention just a few here. George W. Dantzig invented the *simplex method* for solving linear optimization problems during his WW2 efforts at Pentagon, as well as the whole machinery of modelling such problems.[1] Dantzig was originally a statistician and famously, as a young Ph.D. student, provided solutions to some then unsolved problems in mathematical statistics that he found on the blackboard when he arrived late to a lecture, believing they were home work assignments in the course. Building on the knowledge of duality in the theory of two-person zero-sum games, which was developed by the world-famous mathematician John von Neumann in the 1920s, Dantzig was very much involved in developing the theory of duality in linear programming, together with the various characterizations of an optimal solution that is brought out from that theory. A large part of the duality theory was developed in collaboration with the mathematician Albert W. Tucker.

[1] As Dantzig explains in [Dan57], linear programming formulations in fact can first be found in the work of the first theoretical economists in France, such as F. Quesnay in 1760; they explained the relationships between the landlord, the peasant and the artisan. The first practical linear programming problem solved with the simplex method was the famous diet problem.

Several researchers interested in national economics studied transportation models at the same time, modelling them as special linear optimization problems. Two of them, the mathematician Leonid W. Kantorovich and the statistician Tjalling C. Koopmans received The Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel in 1975 "for their contributions to the theory of optimum allocation of resources." They had, in fact, both worked out some of the basics of linear programming, independently of Dantzig, at roughly the same time. (Dantzig stands out among the three especially for creating an efficient algorithm for solving such problems.)[2]

1.3 Classification of optimization models

We here develop a subset of problem classes that can be set up by contrasting certain aspects of a general optimization problem. We let

$$\begin{aligned} \boldsymbol{x} \in \mathbb{R}^n &: \text{vector of decision variables } x_j, \quad j = 1, 2, \ldots, n; \\ f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\} &: \text{objective function;} \\ X \subseteq \mathbb{R}^n &: \text{ground set defined logically/physically;} \\ g_i : \mathbb{R}^n \to \mathbb{R} &: \text{constraint function defining restriction on } \boldsymbol{x}: \end{aligned}$$

$$\begin{aligned} g_i(\boldsymbol{x}) &\geq b_i, \qquad i \in \mathcal{I}; \qquad \text{(inequality constraints)} \\ g_i(\boldsymbol{x}) &= d_i, \qquad i \in \mathcal{E}. \qquad \text{(equality constraints)} \end{aligned}$$

We let $b_i \in \mathbb{R}$ $(i \in \mathcal{I})$ and $d_i \in \mathbb{R}$ $(i \in \mathcal{E})$ denote the right-hand sides of these constraints; without loss of generality, we could actually let them all be equal to zero, as any constants can be incorporated into the definitions of the functions g_i $(i \in \mathcal{I} \cup \mathcal{E})$.

The optimization problem then is to

$$\begin{aligned} \operatorname*{minimize}_{\boldsymbol{x}} \quad & f(\boldsymbol{x}), && (1.1a) \\ \text{subject to} \quad & g_i(\boldsymbol{x}) \geq b_i, \qquad i \in \mathcal{I}, && (1.1b) \\ & g_i(\boldsymbol{x}) = d_i, \qquad i \in \mathcal{E}, && (1.1c) \\ & \boldsymbol{x} \in X. && (1.1d) \end{aligned}$$

(If it is really a maximization problem, then we change the sign of f.)

[2]Incidentally, several other laureates in economics have worked with the tools of optimization: Paul A. Samuelson (1970, linear programming), Kenneth J. Arrow (1972, game theory), Wassily Leontief (1973, linear transportation models), Gerard Debreu (1983, game theory), Harry M. Markowitz (1990, quadratic programming in finance), John F. Nash Jr. (1994, game theory), William Vickrey (1996, econometrics), and Daniel L. McFadden (2000, microeconomics).

The problem type depends on the nature of the functions f and g_i, and the set X. Let us look at some examples.

(LP) Linear programming Objective function linear: $f(\boldsymbol{x}) = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} = \sum_{j=1}^{n} c_j x_j$ ($\boldsymbol{c} \in \mathbb{R}^n$); constraint functions affine: $g_i(\boldsymbol{x}) = \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} - b_i$ ($\boldsymbol{a}_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$, $i \in \mathcal{I} \cup \mathcal{E}$)); $X = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid x_j \geq 0, \quad j = 1, 2, \dots, n \,\}$.

(NLP) Nonlinear programming Some function(s) f, g_i ($i \in \mathcal{I} \cup \mathcal{E}$) are nonlinear.

Continuous optimization f, g_i ($i \in \mathcal{I} \cup \mathcal{E}$) are continuous on an open set containing X; X is closed and convex.

(IP) Integer programming $X \subseteq \{0, 1\}^n$ (binary) or $X \subseteq \mathbb{Z}^n$ (integer).

Unconstrained optimization $\mathcal{I} \cup \mathcal{E} = \emptyset$; $X = \mathbb{R}^n$.

Constrained optimization $\mathcal{I} \cup \mathcal{E} \neq \emptyset$ and/or $X \subset \mathbb{R}^n$.

Differentiable optimization f, g_i ($i \in \mathcal{I} \cup \mathcal{E}$) are at least once continuously differentiable on an open set containing X (that is, "in C^1 on X," which means that ∇f and ∇g_i ($i \in \mathcal{I} \cup \mathcal{E}$) exist there and the gradients are continuous); further, X is closed and convex.

Non-differentiable optimization At least one of f, g_i ($i \in \mathcal{I} \cup \mathcal{E}$) is non-differentiable.

(CP) Convex programming f is convex; g_i ($i \in \mathcal{I}$) are concave; g_i ($i \in \mathcal{E}$) are affine; and X is closed and convex.

Non-convex programming The complement of the above

In Figure 1.2 we show how the problem types NLP, IP, and LP are related.

That LP is a special case of NLP is clear by the fact that a linear function is a special kind of nonlinear function; that IP is a special case

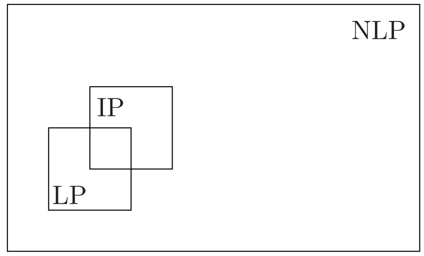


Figure 1.2: The relations among NLP, IP, and LP.

of NLP can be illustrated by the fact that the constraint $x_j \in \{0, 1\}$ can be written as the nonlinear constraint $x_j(1 - x_j) = 0$.[3]

Last, there is a subclass of IP that is equivalent to LP, that is, a class of problems for which there exists at least one optimal solution which *automatically* is integer valued even without imposing any integrality constraints, provided of course that the problem has any optimal solutions at all. We say that such problems have the *integrality property*. An important example problem belonging to this category is the linear single-commodity network flow problem with integer data; this class of problems in turn includes as special cases such important problems as the linear versions of the assignment problem, the transportation problem, the maximum flow problem, and the shortest route problem.

Among the above list of problem classes, we distinguish, roughly only, between two of the most important ones, as follows:

LP Linear programming $\approx$ applied linear algebra. LP is "easy," because there exist algorithms that can solve every LP problem instance efficiently in practice.

NLP Nonlinear programming $\approx$ applied analysis in several variables. NLP is "hard," because there does *not* exist an algorithm that can solve every NLP problem instance efficiently in practice. NLP is such a large problem area that it contains very hard problems as well as very easy problems. The largest class of NLP problems that are solvable with some algorithm in reasonable time is CP (of which LP is a special case).

Our problem formulation (1.1) does not cover the following:

[3] If a non-negative integer variable x_j is upper bounded by the integer M, it is also possible to write $\prod_{k=0}^{M}(x_j - k) = (x_j - 0)(x_j - 1) \cdots (x_j - M) = 0$, by which we restrict a *continuous* variable x_j to be integer-valued.

- infinite-dimensional problems (that is, problems formulated in functional spaces rather than vector spaces);
- implicit functions f and/or g_i ($i \in \mathcal{I} \cup \mathcal{E}$): then, no explicit formula can then be written down; this is typical in engineering applications, where the value of, say, $f(\boldsymbol{x})$ can be the result of a simulation;
- multiple-objective optimization:

$$\text{“minimize } \{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \dots, f_p(\boldsymbol{x})\}\text{”;}$$

- optimization under uncertainty, or, stochastic programming (that is, where some of f, g_i ($i \in \mathcal{I} \cup \mathcal{E}$) are only known probabilistically).

1.4 Conventions

Let us denote the set of vectors satisfying the constraints (1.1b)–(1.1d) by $S \subseteq \mathbb{R}^n$, that is, the set of *feasible solutions* to the problem (1.1). What exactly do we mean by solving the problem to

$$\underset{\boldsymbol{x} \in S}{\text{minimize}}\ f(\boldsymbol{x})? \tag{1.2}$$

Since there is no explicit operation involved here, the question is warranted. The following two operations are however well-defined:

$$f^* := \underset{\boldsymbol{x} \in S}{\text{infimum}}\ f(\boldsymbol{x})$$

denotes the infimum value of the function f over the set S; if and only if the infimum value is attained at some point $\boldsymbol{x}^*$ in S (and then both f^* and $\boldsymbol{x}^*$ necessarily are finite) we can write that

$$f^* := \underset{\boldsymbol{x} \in S}{\text{minimum}}\ f(\boldsymbol{x}), \tag{1.3}$$

and then we of course have that $f(\boldsymbol{x}^*) = f^*$. (When considering maximization problems, we obtain the analogous definitions of the supremum and the maximum.)

The second operation defines the set of optimal solutions to the problem at hand:

$$S^* := \arg \underset{\boldsymbol{x} \in S}{\text{minimum}}\ f(\boldsymbol{x});$$

the set $S^* \subseteq S$ is nonempty if and only if the infimum value f^* is attained. Finding at least one optimal solution,

$$\boldsymbol{x}^* \in \arg \underset{\boldsymbol{x} \in S}{\text{minimum}}\ f(\boldsymbol{x}), \tag{1.4}$$

is a special case which moreover defines an often much more simple task.

As an example, consider the problem instance where $S = \{ x \in \mathbb{R} \mid x \geq 0 \}$ and

$$f(x) = \begin{cases} 1/x, & \text{if } x > 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

For this problem $f^* = 0$ but $S^* = \emptyset$, because the value 0 is not attained for a finite value of x—the problem has a finite infimum value but not an optimal solution.

These examples lead to our convention in reading the problem (1.2): the statement "solve the problem (1.2)" means "find f^* and an $\boldsymbol{x}^* \in S^*$, or conclude that $S^* = \emptyset$."

Hence, it is *implicit* in the formulation that we are interested both in the infimum value and in (at least) one optimal solution if one exists. Whenever we are certain that only one of the two is of interest then we will state so explicitly. We are aware that the formulation has, in the past, been considered "vague" since no operation is visible; so, to summarize and clarify our convention, it in fact includes two operations, (1.3) and (1.4).

There is a second reason for stating the optimization problem (1.1) in the way it is, a reason which is computational. To solve the problem, we almost always need to solve a sequence of relaxations/simplifications of the original problem in order to eventually reach a solution. (These problem manipulations include Lagrangian relaxation, penalization, and objective function linearization, which will be developed later on.) When describing the particular relaxation/simplification utilized, having access to constraint identifiers [such as (1.1c)] certainly makes the presentation easier and clearer. That will become especially valuable when dealing with various forms of duality, when (subsets of) the constraints are relaxed.

A last comment on conventions: as it is stated prior to the problem formulation (1.1) the objective function f can in general take on both $\pm\infty$ as values. Since we are generally going to study minimization problem, we will only be interested in objective functions f having the properties that (a) $f(\boldsymbol{x}) \neq -\infty$ for every feasible vector $\boldsymbol{x}$, and (b) $f(\boldsymbol{x}) < +\infty$ for at least one feasible vector $\boldsymbol{x}$. Such functions are known as *proper* functions (which makes sense, as it is impossible to perform a proper optimization unless these two properties hold). We will some times refer to these properties, in particular by stating explicitly when f can take on the value $+\infty$, but we will assume throughout that f does *not* take on the value $-\infty$. So, in effect then, *we assume implicitly that the objective function f is proper*.

1.5 Applications and modelling examples

To give but a quick view of the scope of applications of optimization, here is a subset of the past few years of applied master's or doctoral projects, performed either at Linköping University or at Chalmers University of Technology:

- Planning routes for snow removal machines
- Planning routes for disabled persons transportation
- Planning of production of energy in power plants
- Scheduling production and distribution of electricity
- Scheduling of empty freight cars in railways
- Scheduling log cutting in forests
- Optimizing paper production in paper mills
- Scheduling paper cutting in paper mills
- Optimization of engine performance for aircraft, boats, and cars
- Portfolio optimization under uncertainty for pension funds
- Analysis of investment in future energy systems
- Network design for mobile telecommunication, optical and internet protocol networks
- Optimal wave-length and routing in optimal networks
- Scheduling of production of circuit boards
- Scheduling of time tables in schools
- Optimal packing and distribution of gas
- Bin packing of objects in freight cars, truck, and cargo ships
- Routing of vehicles for road carriers
- Optimal congestion pricing in urban traffic networks

1.6 Defining the field

To define what the subject area of optimization encompasses is difficult, given that it is connected to so many scientific areas in the natural and technical sciences.

An obvious distinguishing factor is that an optimization model always has an objective function and a group of constraints. On the other hand by letting $f \equiv 0$ and $\mathcal{E} = \emptyset$ then the generic problem (1.1) is that of a *feasibility problem* for equality constraints, and by instead letting $\mathcal{I} \cup \mathcal{E} = \emptyset$ we obtain an *unconstrained optimization* problem. Both these special cases are classic problems in *numerical analysis*, which most often deal with the solution of a linear or non-linear system of equations.

We can here identify a distinguishing element between optimization and numerical analysis—that an optimization problem often involve *inequality constraints* while a problem in numerical analysis does not. Why does that make a difference? The reason is that while in the latter case the analysis is performed on a manifold—possibly even a linear subspace—the analysis of an optimization problem must deal with the fact that there are feasible regions residing in different dimensions because of the nature of inequality constraints being either active or inactive. As a result, there will always be some kind of *non-differentiabilities* present in some associated functionals, while numerical analysis typically is "smooth."

As an illustration, although this is beyond the scope of this book, we ask the reader to ask herself what the proper extension of the famous *Implicit Function Theorem* is when we replace the system $\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) = \mathbf{0}^\ell$ with, say, $\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) \leq \mathbf{0}^m$?

1.7 Soft and hard constraints

1.7.1 Definitions

So far, we have not discussed much about the role of different types of constraints. In the *set covering problem*, for example, the constraints are of the form $\sum_{j=1}^n a_{ij} x_j \geq 1$, $i = 1, 2, \dots, m$, where $a_{ij} \in \{0, 1\}$. These, as well as constraints of the form $x_j \geq 0$ and $x_j \in \{0, 1\}$ are *hard constraints*, meaning that if they are violated then the solution does not make much sense. Typical such constraints are technological ones; for example, if x_j is associated with the level of production, then a negative value has no meaning, and therefore a negative value is never acceptable. A binary variable, $x_j \in \{0, 1\}$, is often logical, associated with the choice between something being "on" or "off," such as a production facility, a city being visited by a traveling salesman, and so on; again, a fractional value like 0.7 makes no sense, and binary restrictions almost always are hard.

Consider now a collection of constraints that are associated with the capacity of production, and suppose it has the form $\sum_{j=1}^n u_{ij} x_{ij} \leq c_i$, $i = 1, 2, \dots, m$, where x_{ij} denotes the level of production of an item/product j using a production process i, u_{ij} is a positive number associated with the use of a resource (man hours, hours before inspection of the machine, etcetera) per unit of production of the item, and c_i is the available capacity of this resource in the production process. In some circumstances, it is not unnatural to allow for the left-hand side to become larger than the capacity, because that production plan might still be feasible, provided however that additional resources are made available. We consider two types of ways to allow for this violation, and which give rise to two different types of solution.

The first, which we are not quite ready to discuss here from a technical standpoint, is connected to the *Lagrangian relaxation* of the capacity constraints. If, when solving the corresponding Lagrangian dual optimization problem, we terminate the solution process prematurely, we will typically have a terminal primal vector that violates some of the capacity constraints slightly. Since the capacity constraints are soft, this solution may be acceptable.[4] See Chapter 6 for further details on Lagrangian duality.

Since it is however natural that additional resources come only at an additional cost, an increase in the violation of this *soft constraint* should have the effect of an additional, increasing cost in the objective function. In other words, violating a constraint should come with a *penalty*. Given a measure of the cost of violating the constraints, that is, the unit cost of additional resource, we may transform the resulting problem to an unconstrained problem with a *penalty function* representing the original constraint.

Below, we relate soft constraints to exterior penalties.

1.7.2 A derivation of the exterior penalty function

Consider the standard nonlinear programming problem to

$$\underset{\boldsymbol{x}}{\operatorname{minimize}} \ f(\boldsymbol{x}), \tag{1.5a}$$

$$\text{subject to } \ g_i(\boldsymbol{x}) \geq 0, \qquad i = 1, \dots, m, \tag{1.5b}$$

where f and g_i $(i = 1, \dots, m)$ are real-valued functions.

Consider the following relaxation of (1.5), where $\rho > 0$:

$$\underset{(\boldsymbol{x},\boldsymbol{s})}{\operatorname{minimize}} \ f(\boldsymbol{x}) + \rho \sum_{i=1}^{m} s_i, \tag{1.6a}$$

$$\text{subject to } \ g_i(\boldsymbol{x}) \geq -s_i, \qquad i = 1, \dots, m, \tag{1.6b}$$

$$s_i \geq 0, \qquad i = 1, \dots, m. \tag{1.6c}$$

We interpret this problem as follows: by allowing the variable s_i to become positive, we allow for extra slack in the constraint, at a positive cost, ρs_i, proportional to the violation.

[4] One interesting application arises when making capacity expansion decisions in production and work force planning problems (e.g., Johnson and Montgomery [JoM74, Example 4-14]) and in forest management scheduling (Hauer and Hoganson [HaH96]).

How do we solve this problem, for a given value of $\rho > 0$? What we will develop below is a specialization of the following result (see, for example, [RoW97, Proposition 1.35]): for a function $\phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ one has in terms of $p(\boldsymbol{s}) = \operatorname{infimum}_{\boldsymbol{x}} \phi(\boldsymbol{x}, \boldsymbol{s})$ and $q(\boldsymbol{x}) = \operatorname{infimum}_{\boldsymbol{s}} \phi(\boldsymbol{x}, \boldsymbol{s})$ that

$$\underset{(\boldsymbol{x},\boldsymbol{s})}{\operatorname{infimum}}\ \phi(\boldsymbol{x}, \boldsymbol{s}) = \underset{\boldsymbol{x}}{\operatorname{infimum}}\ q(\boldsymbol{x}) = \underset{\boldsymbol{s}}{\operatorname{infimum}}\ p(\boldsymbol{s}).$$

In other words, we can solve an optimization problem in two types of variables $\boldsymbol{x}$ and $\boldsymbol{s}$ by "eliminating" one of them (in our case, $\boldsymbol{s}$) through optimization, and then determine the best value of the remaining one.

Suppose then that we for a moment keep $\boldsymbol{x}$ fixed to an arbitrary value. The above problem (1.6) then reduces to that to

$$\begin{aligned} \underset{\boldsymbol{s}}{\text{minimize}}\ & \rho \sum_{i=1}^{m} s_i, && \text{(1.7a)} \\ \text{subject to}\ & s_i \geq -g_i(\boldsymbol{x}), \qquad i = 1, \dots, m, && \text{(1.7b)} \\ & s_i \geq 0, \qquad i = 1, \dots, m, && \text{(1.7c)} \end{aligned}$$

which clearly separates into the m independent problems to

$$\begin{aligned} \underset{s_i}{\text{minimize}}\ & \rho s_i, && \text{(1.8a)} \\ \text{subject to}\ & s_i \geq -g_i(\boldsymbol{x}), && \text{(1.8b)} \\ & s_i \geq 0. && \text{(1.8c)} \end{aligned}$$

This problem is trivially solvable: $s_i := \text{maximum}\,\{0, -g_i(\boldsymbol{x})\}$, that is, s_i takes on the role of a slack variable for the constraint. Replacing s_i with this expression in $\boldsymbol{x}$ in the problem (1.6) we finally obtain the problem to

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}}\ & f(\boldsymbol{x}) + \rho \sum_{i=1}^{m} \text{maximum}\,\{0, -g_i(\boldsymbol{x})\}, && \text{(1.9a)} \\ \text{subject to}\ & \boldsymbol{x} \in \mathbb{R}^n. && \text{(1.9b)} \end{aligned}$$

If the constraints instead are of the form $g_i(\boldsymbol{x}) \leq 0$, then the resulting penalty function is of the form $\rho \sum_{i=1}^{m} \text{maximum}\,\{0, g_i(\boldsymbol{x})\}$.

See Section 13.1 for a thorough discussion on and analysis of penalty functions and methods.

1.8 A road map through the material

Chapter 2 gives a short overview of some basic material from calculus and linear algebra that is used throughout the book. Familiarity with these topics is therefore very important.

Chapter 3 is devoted to the study of convexity, a subject known as *convex analysis*. We characterize the convexity of sets and real-valued functions and show their relations. We provide an overview of the special convex sets called polyhedra, which can be described by linear constraints. Parts of the theory covered, such as the Representation Theorem, Farkas' Lemma and the Separation Theorem, build the foundation of the study of optimality conditions in Chapter 5, the theory of strong duality in Chapter 6 and of linear programming in Chapters 7–10.

Chapter 4 gives a gentle overview of topics associated with optimality, including the very important result that locally optimal solutions are globally optimal solution in a convex problem. We establish basic results regarding the existence of optimal solutions, including the famous Weierstrass Theorem, and establish basic logical relationships between locally optimal solutions and characterizations in terms of conditions of "stationarity". The latter includes the standard result in differentiable, unconstrained optimization that says that a locally optimal solution must have the property that the gradient of the objective function there is zero. Along the way, we define important concepts such as the normal cone, the variational inequality, and the Euclidean projection of a vector onto a convex set, and outline fixed point theorems and their applications.

Chapter 5 collects results leading up to the central Karush–Kuhn–Tucker (KKT) Theorem on the necessary conditions for the local optimality of a feasible point in a constrained optimization problem. Essentially, these conditions state that a given feasible vector $\boldsymbol{x}$ can only be a local minimum if it is feasible in the problem and if there is no descent direction at $\boldsymbol{x}$ which simultaneously is a feasible direction. In order to state the KKT conditions in algebraic terms such that it can be checked in practice and such that as few interesting vectors $\boldsymbol{x}$ as possible satisfy them, we must restrict our study to problems and vectors satisfying some regularity properties. These properties are called constraint qualifications (CQs); among them, the classic one is that "the active constraints are linearly independent" which is familiar from the Lagrange Multiplier Theorem in differential calculus. Our treatment however is more general and covers weaker (that is, better) CQs as well. The chapter begins with a schematic road map for these results to further help in the study of this material.

Chapter 6 presents a rather broad picture of the theory of Lagrangian duality. Associated with the KKT conditions in the previous chapter is a vector, known as the Lagrange multiplier vector, denoted $\boldsymbol{\mu}$ ($\boldsymbol{\lambda}$) for inequality (equality) constraints. The Lagrange multipliers are associated with an optimization problem which is referred to as the Lagrangian dual, or simply dual, problem.[5] The role of the dual problem is to define a largest lower bound on the primal value f^* of the primal (original) problem. This chapter establishes the basic properties of this dual problem. In particular, it is always a convex problem. It is therefore an appealing problem to solve in order to extract the optimal solution to the primal problem. This chapter is in fact almost entirely devoted to the topic of analyzing when it is possible to generate, from an optimal dual solution $\boldsymbol{\mu}^*$, in a rather simple manner an optimal primal solution $\boldsymbol{x}^*$. The most important term in this context then is "strong duality" which refers to the occasion when the optimal values in the two problems are equal—only then is the "translation" relatively easy. Some of the results established here are immediately transferable to the important case of linear programming, so the link between this chapter and Chapter 10 is very strong. The main difference is that in the present chapter we must work with more general tools, while for linear programming we have access to a more specialized analysis; therefore, proof techniques, for example in establishing the Strong Duality Theorem, will be quite different. Additional topics include an analysis of optimization algorithms for the solution of the Lagrangian dual problem, and sensitivity analysis with respect to changes in the right-hand sides of inequality constraints.

Chapters 7–10 are devoted to the study of linear programming (LP) models and methods. Its importance is unquestionable: it has been stated that in the 1980s LP problems was the scientific problem that ate the most computing power in the world. While the efficiency of LP solvers have multiplied since then, so has the speed of computers, and LP models still define the most important problem area in optimization in practice. (Partly, this is also due to the fact that integer programming models, where some, or all, variables are required to take on integer values, use LP techniques.) It is not only for this reason however that we devote special chapters to this topic. Their optimal solutions can be found using quite special techniques that are not common to nonlinear programming. As was shown in Chapter 3 linear programs have optimal solutions at the extreme points of the polyhedral feasible set. This fact, together with the linearity of the objective function and the constraints, means that a feasible-direction (descent) method can be very cleverly devised. Since we know that only extreme points are of interest, we start at one extreme point, and then only consider as candidate search

[5]The dual problem was first discovered in the study of (linear) matrix games by John von Neumann in the 1920s, but had for a long time implicitly been used also for nonlinear optimization problems before it was properly stated and studied by Arrow, Hurwicz, Uzawa, Everett, Falk, Rockafellar, etcetera, starting in earnest in the 1950s. By the way, the original problem is then referred to as the primal problem, a name given by George Dantzig's father, a Greek scholar.

directions those that point towards another (in fact, adjacent) extreme point. We can generate such directions extremely efficiently by using a basis representation of the extreme points, and the move from one extreme point to the other is then associated with a very simple basis change. This special procedure is known as the Simplex method, which was invented by George Dantzig in the 1940s.

In Chapter 7 a simple manufacturing problem is used to illustrate the basics of linear programming. The problem is graphically solved and it turns out that the optimal solution is an extreme point. We investigate how the optimal solution changes if the data of the problem is changed, and the linear programming dual to the manufacturing problem is derived by using economical arguments.

Chapter 8 begins with a presentation of the axioms underlying the use of LP models, and a general modelling technique is discussed. The rest of the chapter deals with the geometry of LP models. It is shown that every linear program can be transformed into the *standard form* which is the form that the Simplex method uses. We introduce the concept of *basic feasible solution* and discuss its connection to extreme points. A version of the Representation Theorem adapted to the standard form is presented, and we show that if there exists an optimal solution to a linear program in standard form, then there exists an optimal solution among the basic feasible solutions. Finally, we define adjacency between extreme points and give an algebraic characterization of adjacency which actually proves that the Simplex method at each iteration step moves from one extreme point to an adjacent one.

Chapter 9 presents the Simplex method. First it is assumed that a basic feasible solution (BFS) is known at the start of the algorithm, and then we describe what to do when a BFS is not known from the beginning. Termination characteristics of the algorithm is discussed and it is shown that if all the BFSs of the problem are non-degenerate, then the basic algorithm terminates. However, if there exist degenerate BFSs there is a possibility that the basic algorithm cycles between degenerate BFSs and hence never terminates. We give a simple rule, called Bland's rule, that eliminates cycling. We close the chapter by discussing the computational complexity of the Simplex algorithm.

In Chapter 10 linear programming duality is studied. We discuss how to construct the linear programming dual to a general linear program and present duality theory, such as weak and strong duality and complementary slackness. The dual simplex method is developed, and we discuss how the optimal solution of a linear program changes if the right-hand side or the objective function coefficients are modified.

Chapter 11 presents basic algorithms for differentiable, unconstrained optimization problems. The typical optimization algorithm is iterative, which means that a solution is approached through a sequence of trial vectors, typically such that each consecutive objective value is strictly lower than the previous one in a minimization problem. This improvement is possible because we can generate improving search directions—descent (ascent) directions in a minimization (maximization) problem—by means of solving an approximation of the original problem or the optimality conditions. This approximate problem (for example, the system of Newton equations) is then combined with a line search, which approximately solve the original problem over the line segment defined by the current iterate and the search direction. This idea of combining approximation (or, relaxation) with a line search (or, coordination) is the basic methodology also for constrained optimization problems. Also, while our opinion is that the subject of differentiable unconstrained optimization largely is a subject within numerical analysis rather than within the optimization field, its understanding is important because the approximations/relaxations that we utilize in constrained optimization often result in (essentially) unconstrained optimization subproblems. We develop a class of quasi-Newton methods in detail, to illustrate a classic analysis.

Chapter 12 presents some natural algorithms for differentiable nonlinear optimization over polyhedral sets, which utilize LP techniques when searching for an improving direction. The basic algorithm is known as the Frank–Wolfe algorithm, or the conditional gradient method; it utilizes $\nabla f(\boldsymbol{x}_k)$ as the linear cost vector at iteration k, and the direction towards any optimal extreme point $\boldsymbol{y}_k$ has already in Chapter 4 been shown to be a feasible direction of descent whenever $\boldsymbol{x}_k$ is not stationary. A line search in the line segment $[\boldsymbol{x}_k, \boldsymbol{y}_k]$ completes an iteration. Because of the work involved in repeatedly solving LPs a natural improvement of this algorithm is to keep in memory all, or some of, the previously generated extreme points $\boldsymbol{y}_0, \boldsymbol{y}_1, \dots, \boldsymbol{y}_{k-1}$, and to generate the next iteration point as the optimal solution within the convex hull of the union of them, the current iterate $\boldsymbol{x}_k$ and the new extreme point $\boldsymbol{y}_k$. The gradient projection method extends the steepest descent method for unconstrained optimization problem in a natural manner. The subproblems here are Euclidean projection problems which in this case are strictly convex quadratic programming problems that can be solved efficiently for some types of polyhedral sets. The convergence results reached show that convexity of the problem is crucial in reaching good convergence results—not only regarding the global optimality of limit points but regarding the nature of the set of limit points as well.

Chapter 13 begins by describing natural approaches to nonlinearly constrained optimization problems, wherein all (or, a subset of) the constraints are replaced by penalties. The resulting penalized problem is then possible to solve by using techniques for unconstrained problems or problems with convex feasible sets, like those we have presented in Chapters 11 and 12. In order to force the penalized problems to more and more resemble the original one, the penalties are more and more strictly enforced. There are essentially two types of penalty functions, exterior and interior penalties. Exterior penalty methods were devised mainly in the 1960s, and are perhaps the most natural ones; they are valid for almost every type of explicit constraint, and is therefore amenable to solving also non-convex problems. The penalty terms are gradually enforced by letting larger and larger weights be associated with the constraints in comparison with the objective function. Under some circumstances, one can show that a finite value of the these penalty parameters are needed, but in general they must tend to infinity. Therefore, these algorithms are often burdened by numerical accuracy problems, which however, in some cases can be limited when Newton methods are used for the subproblems. Interior penalty methods are also amenable to the solution of non-convex problems, but are perhaps most naturally associated with convex problems, where they are quite effective. In particular, the best methods for linear programming in terms of their worst-case complexity are interior point methods which are based on interior penalty functions. In this type of method, the interior penalties are asymptotes with respect to the constraint boundaries; a decreasing value of the penalty parameters then allow for the boundaries to be approached at the same time as the original objective function come more and more into play. For both types of methods, we reach convergence results on the convergence to KKT points in the general case—including estimates of the Lagrange multipliers—, and global convergence results in the convex case.

Chapter 13 continues by describing a basic and quite popular class of algorithms for general nonlinear programming problems with twice differentiable objective and constraint functions. It is called Sequential Quadratic Programming (SQP) and is, essentially, Newton's method applied to the KKT conditions of the problem; there are, however, some modifications necessary. For example, because of the linearization of the constraints, it is in general difficult to maintain feasibility in the process, and therefore convergence cannot merely be based on line searches in the objective function; instead one must devise a measure of "goodness" that take constraint violation into account. The classic approach is to utilize a penalty function so that a constraint violation comes with a price, and as such the SQP method ties in with the penalty methods above. Another approach which is gaining popularity is to use a type of bi-criterion method where a new iterate is "accepted" based both on its objective value and its constraint violation; this is referred to as a *filter SQP* method. In any case, in this type of method one strives for feasibility and optimality simultaneously, like Lagrangian relaxation methods do; in fact, there are strong relationships between the methods in this chapter and Lagrangian methods.

Each chapter ends with exercises on its contents, through numerical examples or extensions of the theory developed; we have also included a few previous exam questions from the course Applied Optimization tought at Chalmers and Gothenburg University.

1.9 On the background of this book and a didactics statement

This book's foundation is the collection of lecture notes written by the third author and used in basic optimization courses for about ten years at Linköping University, Chalmers University of Technology, and Gothenburg University. With the same lecturer the course Applied Optimization has been given at Chalmers University of Technology and Gothenburg University since 1997, and the lecture notes have developed more and more from one based on algorithms to one that mainly covers the fundamentals of optimization. With the addition of the first two authors has come a further development of these fundamentals into the present book, in which also our didactic wishes has begun to come true.

The third author's main inspiration in shaping the lecture notes and the book came from the excellent text book by Bazaraa, Sherali, and Shetty [BSS93]. The authors separate the basic theory (convexity, polyhedral theory, separation, optimality, etcetera) from the algorithms devised for solving nonlinear optimization problems, and they develop the theory based on first principles, in a natural order. (The book is however too advanced to be used in a first optimization course, it does not cover linear programming, and the algorithmic part is getting old in some parts.)

In writing the book we have also made a few additional didactic developments. In almost every text book on optimization the topic of linear optimization is developed before that of nonlinear and convex optimization, and linear programming duality is developed before Lagrangian duality. Teaching in this order may however feel unnatural both for the lecturer and for the students: since Lagrangian duality is more general, but similar, to linear programming duality, the feeling is that more or less the same material is repeated, or, which is even worse, the feeling is that linear programming is a rather strange special case that we develop because we must, but not because it is an interesting topic. We have developed the material in this book such that linear programming emerges as a natural special case of general convex programming, having a duality theory which is even richer than that of general convex programming duality.

In keeping with this idea of developing nonlinear programming before linear programming, we should also have covered the simplex method last in the book. This is a possibly conflicting situation, because we believe that the simplex method should not be described merely as a feasible-direction method; its combinatorial nature is important, and the subject of degeneracy is more naturally treated and understood by developing the simplex method immediately following the development of the connections between the geometry and algebra of linear programming. This has been our choice, and we have consequently also decided that iterative algorithms for general nonlinear optimization over convex sets, especially polyhedra, should be developed before those for more general constraints, the reason being that linear programming is an important basis for these algorithms.

1.10 Notes and further reading

Extensive collections of optimization applications and models can be found in several basic text books in operations research, such as [Wag75, BHM77, Mur95, Rar98, Tah03]. The optimization modelling book by Williams [Wil99] is a classic, now in its fourth edition. Modelling books also exist for certain categories of applications; for example, the book [EHL01] concerns the mathematical modelling and solution of optimization problem arising in chemical engineering applications.

Several accounts have been written during the past few years on the origins of operations research and mathematical programming, the reasons being that we recently celebrated the 50^{th} anniversary of the simplex method (1997), the 80^{th} birthday of its inventor George Dantzig (1994), the 50^{th} anniversary of the creation of ORSA (the Operations Research Society of America) (2002), and the 50^{th} anniversary of the Operational Research Society (2003). The special issue of the journal *Operations Research*, vol. 50, no. 1 (2002), is filled with historical anecdotes, as is the book *History of Mathematical Programming* ([LRS91]).

1.11 Exercises

Exercise 1.1 (modelling, exam 980819) A new producer of perfume wish to get a break into a lucrative market. An exclusive fragrance, Chinelle, is to be produced and marketed for maximum profit. With the equipment available it is possible to produce the perfume using two alternative processes, and the company also consider utilizing the services of a famous model when launching it. In order to simplify the problem, let us assume that the perfume is manufactured by the use of two main ingredients—the first a secret substance called MO and the second a more well-known mixture of ingredients. The first of the two processes available provides three grams of perfume for every unit of MO and two units of the standard substance, while the other process gives five grams of perfume for every two (respectively, three) units of the two main ingredients. The company has at its disposal manufacturing processes that can produce at most 20,000 units of MO during the planning period and 35,000 units of the standard mixture. Every unit of MO costs three EUR (it is manufactured in France) to produce, and the other mixture only two EUR per unit. One gram of the new perfume will cost fifty EUR. Even without any advertising the company thinks they can sell 1000 grams of the perfume, simply because of the news value. A famous model can be contracted for commercials, costing 5,000 EUR per photo session (which takes half an hour), and the company thinks that a campaign using his image can raise the demand by about 200 grams per half hour of his time, but not exceeding three hours (he has too many other offers).

Formulate the problem of choosing the best production strategy as an LP problem. ■

Exercise 1.2 (modelling) A computer company has estimated the number of service hours needed during the next five months, according to Table 1.2.

Month	# Service hours
January	6000
February	7000
March	8000
April	9500
May	11,500

Table 1.2: Number of service hours per month; Exercise 1.2.

The service is performed by hired technicians; their number is 50 at the beginning of January. Each technician can work up to 160 hours per month. In order to cover the future demand of technicians new ones must be hired. Before a technician is hired he/she undergoes a period of training, which takes a month and requires 50 hours of supervision by a trained technician. A trained technician has a salary of 15,000 SEK per month (regardless of the number of working hours) and a trainee has a monthly salary of 7500 SEK. At the end of each month on average 5% of the technicians quit to work for another company.

Formulate an LP problem whose optimal solution will minimize the total salary costs during the given time period, given that the number of available service hours are enough to cover the demand. ■

Exercise 1.3 (modelling, exam 010821) The advertising agency ZAP (Zetterström, Anderson, and Pettersson) is designing their new office with an open office space. The office is rectangular, with length l meters and width b meters. Somewhat simplified, we may assume that each working space requires a circle of diameter d and that the working spaces must not overlap. In addition, each working space must be connected to the telecom and computer network at one of the two possible connection points in the office. As the three telephones have limited cable lengths (the agency is concerned with the possible radiation danger associated with hands-free phones and therefore do not use cordless phones)—a_i meters, respectively, $i = 1, \dots, 3$—the work spaces must be placed quite near the connection points.[6] See Figure 1.3 for a simple picture of the office.

For simplicity we assume that the phone is placed at the center of the work place. One of the office's walls is a large panorama window and the three partners all want to sit as close as possible to it. Therefore, they decide to try to minimize the distance to the window for the workplace that is the furthest away from it.

Formulate the problem of placing the three work places so that the maximum distance to the panorama window is minimized, subject to all the necessary constraints. ■

Exercise 1.4 (modelling, exam 010523) A large chain of department stores wants to build a number of distribution centers (warehouses) which will supply 30 department stores with goods. They have 10 possible locations to choose between. To build a warehouse at location i ($i = 1, \dots, 10$) costs c_i MEUR and the capacity of a warehouse at that location would be k_i volume units per week. Department store j has a demand of e_j volume units per week. The distance between warehouse i and department

[6] All the money went to other interior designs of the office space, so there is no money left to buy more cable.

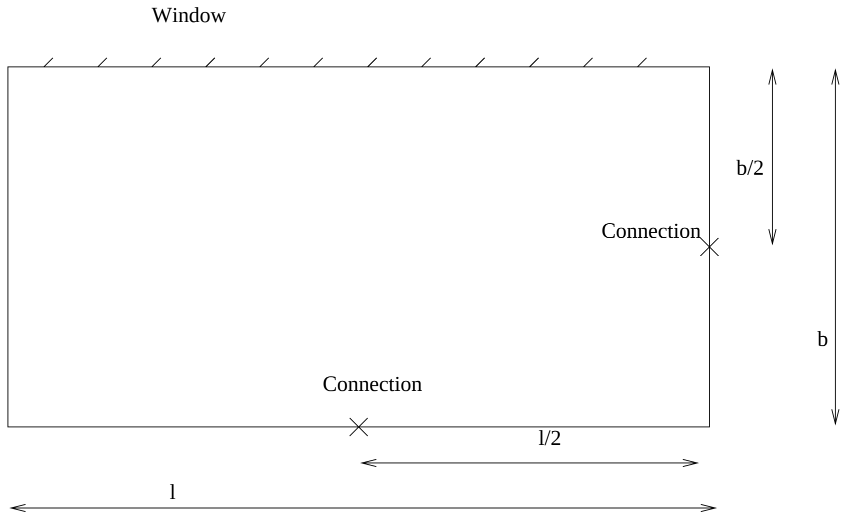


Figure 1.3: Image of the office; Exercise 1.3

store j is d_{ij} km, $i = 1, \ldots, 10$, $j = 1, \ldots, 30$, and a certain warehouse can only serve a department store if the distance is at most D km.

One wishes to minimize the cost of investing in the necessary distribution centers.

(a) Formulate a *linear integer optimization model* describing the optimization problem.

(b) Suppose each department store must be served from *one* of the warehouses. What must be changed in the model? ■

Part II

Fundamentals

Analysis and algebra—A summary

II

The analysis of optimization problems and related optimization algorithms requires the basic understanding of formal logic, linear algebra, and multidimensional analysis. This chapter is not intended as a substitution for the basic courses on these subjects but rather to give a brief review of the notation, definitions, and basic facts which will be used in the subsequent parts without any further notice. If you feel inconvenient with the limited summaries presented in this chapter, contact any of the abundant number of basic text books on the subject.

2.1 Reductio ad absurdum

Together with the absolute majority of contemporary mathematicians we accept proofs by contradiction. The proofs in this group essentially appeal to Aristotles law of the excluded middle, which states that any proposition is either true or false. Thus, if some statement can be shown to lead to a contradiction, we conclude that the original statement is false.

Formally, proofs by contradictions amount to the following:

$$(A \implies B) \iff (\neg A \vee B) \iff (\neg\neg B \vee \neg A) \iff (\neg B \implies \neg A).$$

In the same spirit, when proving $A \iff B$, that is, $(A \implies B) \wedge (B \implies A)$, we often argue that $(A \implies B) \wedge (\neg A \implies \neg B)$ (see, for example, the proof of Farkas' Lemma 3.30).

2.2 Linear algebra

We will always work with finite dimensional Eucledian vector spaces $\mathbb{R}^n$, the natural number n denoting the dimension of the space. Elements $\boldsymbol{v} \in \mathbb{R}^n$ will be referred to as *vectors*, and we will always think of them as of n real numbers stacked on top of each other, i.e., $\boldsymbol{v} = (v_1, \ldots, v_n)^{\mathrm{T}}$, v_i being real numbers, and $^{\mathrm{T}}$ denoting the "transpose" sign. The basic operations defined for two vectors $\boldsymbol{a} = (a_1, \ldots, a_n)^{\mathrm{T}} \in \mathbb{R}^n$ and $\boldsymbol{b} = (b_1, \ldots, b_n)^{\mathrm{T}} \in \mathbb{R}^n$, and an arbitrary scalar $\alpha \in \mathbb{R}$ are as follows:

- addition: $\boldsymbol{a} + \boldsymbol{b} = (a_1 + b_1, \ldots, a_n + b_n)^{\mathrm{T}} \in \mathbb{R}^n$;
- multiplication by a scalar: $\alpha\boldsymbol{a} = (\alpha a_1, \ldots, \alpha a_n)^{\mathrm{T}} \in \mathbb{R}^n$;
- *scalar product* between two vectors: $(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{n} a_i b_i \in \mathbb{R}$. Scalar product will most often be denoted as $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}$ in the subsequent chapters.

A *linear subspace* $L \subset \mathbb{R}^n$ is a set enjoying the following two properties:

- for every $\boldsymbol{a}, \boldsymbol{b} \in L$ it holds that $\boldsymbol{a} + \boldsymbol{b} \in L$, and
- for every $\alpha \in \mathbb{R}, \boldsymbol{a} \in L$ it holds that $\alpha\boldsymbol{a} \in L$.

An *affine subspace* $A \subset \mathbb{R}^n$ is any set that can be represented as $\boldsymbol{v}+L := \{\, \boldsymbol{v} + \boldsymbol{x} \mid \boldsymbol{x} \in L \,\}$ for some vector $\boldsymbol{v} \in \mathbb{R}^n$ and some linear subspace $L \subset \mathbb{R}^n$.

We associate a *norm*, or length, of a vector $\boldsymbol{v} \in \mathbb{R}^n$ with a scalar product as:

$$\|\boldsymbol{v}\| = \sqrt{(\boldsymbol{v}, \boldsymbol{v})}.$$

We will often write $|\boldsymbol{v}|$ in place of $\|\boldsymbol{v}\|$. The Cauchy–Bunyakowski–Schwarz inequality says that $(\boldsymbol{a}, \boldsymbol{b}) \leq \|\boldsymbol{a}\|\|\boldsymbol{b}\|$ for $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$; thus we may define an angle θ between two vectors via $\cos\theta = (\boldsymbol{a}, \boldsymbol{b})/(\|\boldsymbol{a}\|\|\boldsymbol{b}\|)$. Thus, we say that $\boldsymbol{a} \in \mathbb{R}^n$ is *orthogonal* to $\boldsymbol{b} \in \mathbb{R}^n$ iff $(\boldsymbol{a}, \boldsymbol{b}) = 0$ (i.e, when $\cos\theta = 0$). The only vector orthogonal to itself is the zero vector $\mathbf{0}^n = (0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^n$; moreover, this is the only vector with zero norm.

The scalar product is *symmetric* and *bilinear*, i.e., for every $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d} \in \mathbb{R}^n$, $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ it holds that $(\boldsymbol{a}, \boldsymbol{b}) = (\boldsymbol{b}, \boldsymbol{a})$, and $(\alpha\boldsymbol{a} + \beta\boldsymbol{b}, \gamma\boldsymbol{c} + \delta\boldsymbol{d}) = \alpha\gamma(\boldsymbol{a}, \boldsymbol{c}) + \beta\gamma(\boldsymbol{b}, \boldsymbol{c}) + \alpha\delta(\boldsymbol{a}, \boldsymbol{d}) + \beta\delta(\boldsymbol{b}, \boldsymbol{d})$.

A collection of vectors $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ is said to be *linearly independent* iff the equality $\sum_{i=1}^{k} \alpha_i \boldsymbol{v}_i = \mathbf{0}^n$, where $\alpha_1, \ldots, \alpha_k$ are arbitrary real numbers, implies that $\alpha_1 = \cdots = \alpha_k = 0$. Similarly, a collection of vectors $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ is said to be *affinely independent* iff the collection $(\boldsymbol{v}_2 - \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k - \boldsymbol{v}_1)$ is linearly independent.

The largest number of linearly independent vectors in $\mathbb{R}^n$ is n; any collection of n linearly independent vectors from $\mathbb{R}^n$ is referred to as *basis*. The basis $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ is said to be *orthogonal* if $(\boldsymbol{v}_i, \boldsymbol{v}_j) = 0$ for all $i, j = 1, \dots, n$, $i \neq j$. If, in addition, it holds that $\|\boldsymbol{v}_i\| = 1$ for all $i = 1, \dots, n$, the basis is called *orthonormal*.

Given the basis $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ in $\mathbb{R}^n$, every vector $\boldsymbol{v} \in \mathbb{R}^n$ can be written in a unique way as $\boldsymbol{v} = \sum_{i=1}^n \alpha_i \boldsymbol{v}_i$, and the n-tuple $(\alpha_1, \dots, \alpha_n)^{\mathrm{T}}$ will be referred to as *coordinates* of $\boldsymbol{v}$ in this basis. If the basis $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ is orthonormal, the coordinates α_i are computed as $\alpha_i = (\boldsymbol{v}, \boldsymbol{v}_i)$, $i = 1, \dots, n$.

The space $\mathbb{R}^n$ will typically be equipped with the *standard* basis $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n)$, where

$$\boldsymbol{e}_i = (\underbrace{0, \dots, 0}_{i-1 \text{ zeros}}, 1, \underbrace{0, \dots, 0}_{n-i \text{ zeros}})^{\mathrm{T}} \in \mathbb{R}^n.$$

This basis is orthogonal, and for every vector $\boldsymbol{v} = (v_1, \dots, v_n)^{\mathrm{T}} \in \mathbb{R}^n$ we have $(\boldsymbol{v}, \boldsymbol{e}_i) = v_i$, $i = 1, \dots, n$, which allows us to identify vectors and their coordinates.

Now, consider two spaces $\mathbb{R}^n$ and $\mathbb{R}^k$. All linear functions from $\mathbb{R}^n$ to $\mathbb{R}^k$ may be described using a linear space of *real matrices* $\mathbb{R}^{k \times n}$ (i.e., with k rows and n columns). Given a matrix $\boldsymbol{A} \in \mathbb{R}^{k \times n}$ it will often be convenient to view it as a row of its columns, which are thus vectors in $\mathbb{R}^k$. Namely, let $\boldsymbol{A} \in \mathbb{R}^{k \times n}$ have elements a_{ij}, $i = 1, \dots, k$, $j = 1, \dots, n$, then we write $\boldsymbol{A} = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_n)$, where $\boldsymbol{a}_i = (a_{1i}, \dots, a_{ki})^{\mathrm{T}} \in \mathbb{R}^k$, $i = 1, \dots, k$, $j = 1, \dots, n$. The addition of two matrices and scalar-matrix multiplication are defined in a straightforward way. For $\boldsymbol{v} = (v_1, \dots, v_n) \in \mathbb{R}^n$ we define $\boldsymbol{A}\boldsymbol{v} = \sum_{i=1}^n v_i \boldsymbol{a}_i \in \mathbb{R}^k$, where $\boldsymbol{a}_i \in \mathbb{R}^k$ are the columns of $\boldsymbol{A}$. We also define *norm* of the matrix by

$$\|\boldsymbol{A}\| = \max_{\boldsymbol{v} \in \mathbb{R}^n, \|\boldsymbol{v}\| = 1} \|\boldsymbol{A}\boldsymbol{v}\|.$$

Well, this is an example of an optimization problem already!

For a given matrix $\boldsymbol{A} \in \mathbb{R}^{k \times n}$ with elements a_{ij} we define $\boldsymbol{A}^{\mathrm{T}} \in \mathbb{R}^{n \times k}$ as the matrix with elements a_{ji} $i = 1, \dots, k$, $j = 1, \dots, n$. We can give a more elegant, but less straightforward definition: $\boldsymbol{A}^{\mathrm{T}}$ is the unique matrix, satisfying the equality $(\boldsymbol{A}\boldsymbol{v}, \boldsymbol{u}) = (\boldsymbol{v}, \boldsymbol{A}^{\mathrm{T}}\boldsymbol{u})$ for all $\boldsymbol{v} \in \mathbb{R}^n$, $\boldsymbol{u} \in \mathbb{R}^k$. From this definition it should be clear that $\|\boldsymbol{A}\| = \|\boldsymbol{A}^{\mathrm{T}}\|$, and that $(\boldsymbol{A}^{\mathrm{T}})^{\mathrm{T}} = \boldsymbol{A}$.

Given two matrices $\boldsymbol{A} \in \mathbb{R}^{k \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times m}$, we define the *product* $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B} \in \mathbb{R}^{k \times m}$ elementwise by $c_{ij} = \sum_{\ell=1}^n a_{i\ell} b_{\ell j}$, $i = 1, \dots, k$, $j = 1, \dots, m$. In other words, $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B}$ if and only if for all $\boldsymbol{v} \in \mathbb{R}^n$, $\boldsymbol{C}\boldsymbol{v} = \boldsymbol{A}(\boldsymbol{B}\boldsymbol{v})$. By definition, the matrix product is associative (that is,

$\boldsymbol{A}(\boldsymbol{BC}) = (\boldsymbol{AB})\boldsymbol{C}$) for matrices of compatible sizes, but *not* commutative (that is, $\boldsymbol{AB} \neq \boldsymbol{BA}$) in general. It is easy (and instructive) to check that $\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\|$, and that $(\boldsymbol{AB})^{\mathrm{T}} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}$. Vectors $\boldsymbol{v} \in \mathbb{R}^n$ can be (and sometimes will be) viewed as matrices $\boldsymbol{v} \in \mathbb{R}^{n\times 1}$. Check that this embedding is norm-preserving, i.e., the norm of $\boldsymbol{v}$ viewed as a vector equals the norm of $\boldsymbol{v}$ viewed as a matrix with one column.

Of course, no discussion about norms could escape mentioning the *triangle inequality*: for all $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : \|\boldsymbol{a}+\boldsymbol{b}\| \leq \|\boldsymbol{a}\| + \|\boldsymbol{b}\|$, as well as its consequence (check this!) for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{k\times n} : \|\boldsymbol{A}+\boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$. It will often be used in a little bit different form: for all $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$: $\|\boldsymbol{b}\| - \|\boldsymbol{a}\| \leq \|\boldsymbol{b} - \boldsymbol{a}\|$.

For quadratic matrices $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ we can discuss the existence of the unique matrix $\boldsymbol{A}^{-1}$, called the *inverse* of $\boldsymbol{A}$, verifying the equality that for all $\boldsymbol{v} \in \mathbb{R}^n : \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{v} = \boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{v} = \boldsymbol{v}$. If the inverse of a given matrix exists, we call the latter *nonsingular*. The inverse matrix exists iff the columns of $\boldsymbol{A}$ are linearly independent; iff the columns of $\boldsymbol{A}^{\mathrm{T}}$ are linearly independent; iff the system of linear equations $\boldsymbol{Ax} = \boldsymbol{v}$ has a unique solution for every $\boldsymbol{v} \in \mathbb{R}^n$; iff the homogeneous system of equations $\boldsymbol{Ax} = \boldsymbol{0}^n$ has $\boldsymbol{x} = \boldsymbol{0}^n$ as its unique solution. From this definition it follows that $\boldsymbol{A}$ is nonsingular iff $\boldsymbol{A}^{\mathrm{T}}$ is nonsingular, and, furthermore, $(\boldsymbol{A}^{-1})^{\mathrm{T}} = (\boldsymbol{A}^{\mathrm{T}})^{-1}$ and therefore will be denoted simply as $\boldsymbol{A}^{-T}$. At last, if $\boldsymbol{A}$ and $\boldsymbol{B}$ are two nonsingular matrices of the same size, then $\boldsymbol{AB}$ is nonsingular (check!) and $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.

If for some vector $\boldsymbol{v} \in \mathbb{R}^n$, and some scalar $\alpha \in \mathbb{R}$ it holds that $\boldsymbol{Av} = \alpha\boldsymbol{v}$, we call α an *eigenvalue* of $\boldsymbol{A}$ and $\boldsymbol{v}$ an *eigenvector*, corresponding to eigenvalue α. Eigenvectors, corresponding to a given eigenvalue, form a linear *subspace* of $\mathbb{R}^n$; two nonzero eigenvectors, corresponding to two distinct eigenvalues are linearly independent. In general, every matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ has n eigenvalues (counted with multiplicity), maybe complex, which are furthermore roots of the *characteristic equation* $\det(\boldsymbol{A} - \lambda\boldsymbol{I}^n) = 0$, where $\boldsymbol{I}^n \in \mathbb{R}^{n\times n}$ is the *identity matrix*, characterized by the fact that for all $\boldsymbol{v} \in \mathbb{R}^n : \boldsymbol{I}^n\boldsymbol{v} = \boldsymbol{v}$. The norm of the matrix is in fact equal to the largest absolute value of its eigenvalues. The matrix $\boldsymbol{A}$ is nonsingular iff none of its eigenvalues are equal to zero, and in this case the eigenvalues of $\boldsymbol{A}^{-1}$ are equal to the inverted eigenvalues of $\boldsymbol{A}$. The eigenvalues of $\boldsymbol{A}^{\mathrm{T}}$ are equal to the eigenvalues of $\boldsymbol{A}$.

We call $\boldsymbol{A}$ *symmetric* iff $\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{A}$. All eigenvalues of symmetric matrices are real, and eigenvectors corresponding to distinct eigenvalues are orthogonal.

Even if $\boldsymbol{A}$ is not quadratic, then $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ as well as $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ are quadratic and symmetric. If the columns of $\boldsymbol{A}$ are linearly independent, then $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is nonsingular. (Similarly, if the columns of $\boldsymbol{A}^{\mathrm{T}}$ are linearly independent, then $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ is nonsingular.)

Sometimes, we will use the following simple fact: for every $\boldsymbol{A} \in \mathbb{R}^{k\times n}$ with elements a_{ij}, $i = 1,\dots,k$, $j = 1,\dots,n$, it holds that $a_{ij} = (\widetilde{\boldsymbol{e}}_i, \boldsymbol{A}\boldsymbol{e}_j)$, where $(\widetilde{\boldsymbol{e}}_1,\dots,\widetilde{\boldsymbol{e}}_k)$ is the standard basis in $\mathbb{R}^k$, and $(\boldsymbol{e}_1,\dots,\boldsymbol{e}_n)$ is the standard basis in $\mathbb{R}^n$, $i = 1,\dots,k$, $j = 1,\dots,n$.

We will say that $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is *positive semidefinite* (resp., *positive definite*), and denote this by $\boldsymbol{A} \succeq \boldsymbol{0}$ (resp., $\boldsymbol{A} \succ \boldsymbol{0}$) iff for all $\boldsymbol{v} \in \mathbb{R}^n$: $(\boldsymbol{v}, \boldsymbol{A}\boldsymbol{v}) \geq 0$ (resp., for all $\boldsymbol{v} \in \mathbb{R}^n, \boldsymbol{v} \neq \boldsymbol{0}^n : (\boldsymbol{v}, \boldsymbol{A}\boldsymbol{v}) > 0$). The matrix $\boldsymbol{A}$ is positive semidefinite (resp. positive definite) iff its eigenvalues are nonnegative (resp., positive).

For two symmetric matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ we will write $\boldsymbol{A} \succeq \boldsymbol{B}$ (resp., $\boldsymbol{A} \succ \boldsymbol{B}$) iff $\boldsymbol{A} - \boldsymbol{B} \succeq \boldsymbol{0}$ (resp., $\boldsymbol{A} - \boldsymbol{B} \succ \boldsymbol{0}$).

2.3 Analysis

Consider a sequence $\{\boldsymbol{x}_k\} \subset \mathbb{R}^n$. We will write $\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}$, for some $\boldsymbol{x} \in \mathbb{R}^n$, or just $\boldsymbol{x}_k \to \boldsymbol{x}$, iff $\lim_{k\to\infty} \|\boldsymbol{x}_k - \boldsymbol{x}\| = 0$. We will say in this case that $\{\boldsymbol{x}_k\}$ *converges* to $\boldsymbol{x}$, or, equivalently, that $\boldsymbol{x}$ is the *limit* of $\{\boldsymbol{x}_k\}$. Owing to the triangle inequality, every sequence might have at most one limit (check this!). At the same time, there are sequences that do not converge. Moreover, an arbitrary non-converging sequence might contain converging subsequence (or even subsequences). We will refer to the limits of such converging subsequences as *limit points* of a given sequence $\{\boldsymbol{x}_k\}$.

A subset $S \subset \mathbb{R}^n$ is called *bounded* if there exist a constant $C > 0$ such that for all $\boldsymbol{x} \in S : \|\boldsymbol{x}\| \leq C$; otherwise, the set will be called unbounded. Now, let $S \subset \mathbb{R}^n$ be bounded. An interesting and very important fact about the bounded subsets of $\mathbb{R}^n$ is that every sequence $\{\boldsymbol{x}_k\} \subset S$ contains a convergent subsequence.

The set $B_\varepsilon(\boldsymbol{x}) = \{\, \boldsymbol{y} \in \mathbb{R}^n \mid \|\boldsymbol{x} - \boldsymbol{y}\| < \varepsilon \,\}$ is called an *open ball* of radius $\varepsilon > 0$ with center $\boldsymbol{x} \in \mathbb{R}^n$. A set $S \subset \mathbb{R}^n$ is called *open* iff for all $\boldsymbol{x} \in S\ \exists \varepsilon > 0 : B_\varepsilon(\boldsymbol{x}) \subset S$. A set S is *closed* iff its complement $\mathbb{R}^n \setminus S$ is open. An equivalent definition of closedness in terms of sequences is: a set $S \in \mathbb{R}^n$ is closed iff all the limit points of any sequence $\{\boldsymbol{x}_k\} \subset S$ belong to S. There exist sets which are neither closed nor open. The set $\mathbb{R}^n$ is both open and closed (why?).

The *closure* of a set $S \subset \mathbb{R}^n$ (notation: $\operatorname{cl} S$) is the smallest closed set containing S; equivalently, it can be defined as the intersection of all closed sets in $\mathbb{R}^n$ containing S. More constructively, the closure $\operatorname{cl} S$ can be obtained by considering all limit points of all sequences in S. The closure is a closed set, and, quite naturally, the closure of a closed set equals the set itself.

The *interior* of a set $S \subset \mathbb{R}^n$ (notation: $\operatorname{int} S$) is the largest open set contained in S. The interior of an open set equals the set itself.

Finally, the *boundary* of a set $S \subset \mathbb{R}^n$ (notation: $\operatorname{bd} S$, or ∂S) is the set difference $\operatorname{cl} S \setminus \operatorname{int} S$.

A *neighbourhood* of a point $\boldsymbol{x} \in \mathbb{R}^n$ is an arbitrary open set containing $\boldsymbol{x}$.

Consider a function $f : S \to \mathbb{R}$, where $S \subset \mathbb{R}^n$. We say that f is *continuous* at $\boldsymbol{x}_0 \in S$ iff for every sequence $\{\boldsymbol{x}_k\} \subset S$ such that $\boldsymbol{x}_k \to \boldsymbol{x}_0$ it holds that $\lim_{k\to\infty} f(\boldsymbol{x}_k) = f(\boldsymbol{x}_0)$. We say that f is continuous on S iff f is continuous at every point of S.

Now, let $f : S \to \mathbb{R}$ be a continuous function defined on some *open* set S. We say that $f'(\boldsymbol{x}_0; \boldsymbol{d}) \in \mathbb{R}$ is a *directional derivative* of f at $\boldsymbol{x}_0 \in S$ in the direction $\boldsymbol{d} \in \mathbb{R}^n$ if the following limit exists:

$$f'(\boldsymbol{x}_0, \boldsymbol{d}) = \lim_{t \downarrow 0} \frac{f(\boldsymbol{x}_0 + t\boldsymbol{d}) - f(\boldsymbol{x}_0)}{t},$$

and then f will be called directionally differentiable at $\boldsymbol{x}_0 \in S$ in the direction $\boldsymbol{d}$. Clearly, if we fix $\boldsymbol{x}_0 \in S$ and assume that $f'(\boldsymbol{x}_0; \boldsymbol{d})$ exists for some $\boldsymbol{d}$, then for every $\alpha \geq 0$ we have that $f'(\boldsymbol{x}_0; \alpha\boldsymbol{d}) = \alpha f'(\boldsymbol{x}_0; \boldsymbol{d})$. If further $f'(\boldsymbol{x}_0; \boldsymbol{d})$ is linear in $\boldsymbol{d}$, then there exists a vector called the *gradient* of f at $\boldsymbol{x}_0 \in S$, denoted as $\nabla f(\boldsymbol{x}_0) \in \mathbb{R}^n$ such that $f'(\boldsymbol{x}_0; \boldsymbol{d}) = (\nabla f(\boldsymbol{x}_0), \boldsymbol{d})$ and f is called *differentiable* at $\boldsymbol{x}_0 \in S$. Naturally, we say that f is differentiable on S if it is differentiable at every point in S.

Equivalently, the gradient $\nabla f(\boldsymbol{x}_0)$ can be defined as follows: $\nabla f(\boldsymbol{x}_0) \in \mathbb{R}^n$ is the gradient of f at $\boldsymbol{x}_0$ iff there exists a function $o : \mathbb{R} \to \mathbb{R}$ such that

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + (\nabla f(\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0) + o(\|\boldsymbol{x} - \boldsymbol{x}_0\|), \tag{2.1}$$

and moreover,

$$\lim_{t \downarrow 0} \frac{o(t)}{t} = 0. \tag{2.2}$$

For a differentiable function $f : S \to \mathbb{R}$ we can go one step further and define second derivatives of f. Namely, a differentiable function f will be called *twice differentiable* at $\boldsymbol{x}_0 \in S$ iff there exists a symmetric matrix denoted by $\nabla^2 f(\boldsymbol{x}_0)$, referred to as the *Hessian matrix*, and a function $o : \mathbb{R} \to \mathbb{R}$ verifying (2.2), such that

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + (\nabla f(\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0, \nabla^2 f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0)) + o(\|\boldsymbol{x} - \boldsymbol{x}_0\|^2). \tag{2.3}$$

Some times it will be convenient to discuss *vector-valued* functions $\boldsymbol{f} : S \to \mathbb{R}^k$. We say that $\boldsymbol{f} = (f_1, \ldots, f_k)^{\mathrm{T}}$ is continuous if every f_i,

$i = 1, \dots, k$ is; similarly we define differentiability. In the latter case, by $\nabla \boldsymbol{f} \in \mathbb{R}^{n \times k}$ we denote a matrix with columns $(\nabla f_1, \dots, \nabla f_k)$.

We call a continuous function $f : S \to \mathbb{R}$ *continuously differentiable* [notation: $f \in C^1(S)$] if it is differentiable on S and the gradient $\nabla f : S \to \mathbb{R}^n$ is continuous on S. We call $f : S \to \mathbb{R}$ *twice continuously differentiable* [notation: $f \in C^2(S)$], if it is continuously differentiable and in addition every component of $\nabla f : S \to \mathbb{R}^n$ is continuously differentiable.

The following alternative forms of (2.1) and (2.3) will be useful some times. If $f : S \to \mathbb{R}$ is once continuously differentiable on S, and $x_0 \in S$, then for every $\boldsymbol{x}$ in some neighborhood of $\boldsymbol{x}_0$ we have

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + (\nabla f(\boldsymbol{\xi}), \boldsymbol{x} - \boldsymbol{x}_0), \tag{2.4}$$

where $\boldsymbol{\xi} = \lambda \boldsymbol{x}_0 + (1 - \lambda)\boldsymbol{x}$, for some $0 \leq \lambda \leq 1$, is a point between $\boldsymbol{x}$ and $\boldsymbol{x}_0$. (This result is also known as the *mean-value theorem*.) Similarly, for twice differentiable functions we have

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + (\nabla f(\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0, \nabla^2 f(\boldsymbol{\xi})(\boldsymbol{x} - \boldsymbol{x}_0)), \tag{2.5}$$

with the same notation.

If $f, g : \mathbb{R}^n \to \mathbb{R}$ are both differentiable, then $f + g$ and fg are, and $\nabla(f + g) = \nabla f + \nabla g$, $\nabla(fg) = g\nabla f + f\nabla g$. Moreover, if g is never zero, then f/g is differentiable and $\nabla(f/g) = (g\nabla f - f\nabla g)/g^2$.

If both $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^k$ and $h : \mathbb{R}^k \to \mathbb{R}$ are differentiable, then $h(\boldsymbol{F})$ is, and $(\nabla h(\boldsymbol{F}))(\boldsymbol{x}) = (\nabla \boldsymbol{F})(\boldsymbol{x}) \cdot (\nabla h)(\boldsymbol{F}(\boldsymbol{x}))$.

Finally, consider a vector-valued function $\boldsymbol{F} : \mathbb{R}^{k+n} \to \mathbb{R}^k$. Assume that $\boldsymbol{F}$ is continuously differentiable in some neighbourhood $\mathcal{N}_{\boldsymbol{u}} \times \mathcal{N}_{\boldsymbol{x}}$ of the point $(\boldsymbol{u}_0, \boldsymbol{x}_0) \in \mathbb{R}^k \times \mathbb{R}^n$, and that $\boldsymbol{F}(\boldsymbol{u}_0, \boldsymbol{x}_0) = \boldsymbol{0}^k$. If the square matrix $\nabla_{\boldsymbol{u}} \boldsymbol{F}(\boldsymbol{u}_0, \boldsymbol{x}_0)$ is nonsingular, then there exists a unique function $\boldsymbol{\varphi} : \mathcal{N}'_{\boldsymbol{x}} \to \mathcal{N}'_{\boldsymbol{u}}$ such that $\boldsymbol{F}(\boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{x}) \equiv \boldsymbol{0}^k$ in $\mathcal{N}'_{\boldsymbol{x}}$, where $\mathcal{N}'_{\boldsymbol{u}} \times \mathcal{N}'_{\boldsymbol{x}} \subset \mathcal{N}_{\boldsymbol{u}} \times \mathcal{N}_{\boldsymbol{x}}$ is another neighbourhood of $(\boldsymbol{u}_0, \boldsymbol{x}_0)$. Furthermore, $\boldsymbol{\varphi}$ is differentiable at $\boldsymbol{x}_0$, and

$$\nabla \boldsymbol{\varphi}(\boldsymbol{x}_0) = -(\nabla_{\boldsymbol{u}} \boldsymbol{F}(\boldsymbol{u}_0, \boldsymbol{x}_0))^{-1} \nabla_{\boldsymbol{x}} \boldsymbol{F}(\boldsymbol{u}_0, \boldsymbol{x}_0).$$

The function $\boldsymbol{\varphi}$ is known as the *implicit function* defined by the system of equations $\boldsymbol{F}(\boldsymbol{u}, \boldsymbol{x}) = \boldsymbol{0}^k$.

Now we consider two special but very important cases.

Let for some $\boldsymbol{a} \in \mathbb{R}^n$ define a *linear function* $f : \mathbb{R}^n \to \mathbb{R}$ via $f(\boldsymbol{x}) = (\boldsymbol{a}, \boldsymbol{x})$. By the Cauchy–Bunyakowski–Schwarz inequality this function is continuous, and writing $f(\boldsymbol{x}) - f(\boldsymbol{x}_0) = (\boldsymbol{a}, \boldsymbol{x} - \boldsymbol{x}_0)$ for every $\boldsymbol{x}_0 \in \mathbb{R}^n$ we immediately identify from the definitions of the gradient and the Hessian that $\nabla f = \boldsymbol{a}$, $\nabla^2 f = \boldsymbol{0}^{n \times n}$.

Similarly, for some $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ define a *quadratic function* $f(\boldsymbol{x}) = (\boldsymbol{x}, \boldsymbol{A}\boldsymbol{x})$. This function is also continuous, and since $f(\boldsymbol{x}) - f(\boldsymbol{x}_0) = (\boldsymbol{A}\boldsymbol{x}_0, \boldsymbol{x}-\boldsymbol{x}_0)+(\boldsymbol{x}_0, \boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}_0))+(\boldsymbol{x}-\boldsymbol{x}_0, \boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}_0)) = ((\boldsymbol{A}+\boldsymbol{A}^{\mathrm{T}})\boldsymbol{x}_0, \boldsymbol{x}-\boldsymbol{x}_0)+0.5(\boldsymbol{x}-\boldsymbol{x}_0, (\boldsymbol{A}+\boldsymbol{A}^{\mathrm{T}})(\boldsymbol{x}-\boldsymbol{x}_0))$, we identify $\nabla f(\boldsymbol{x}_0) = (\boldsymbol{A}+\boldsymbol{A}^{\mathrm{T}})\boldsymbol{x}_0$, $\nabla^2 f(\boldsymbol{x}_0) = \boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}}$. If the matrix $\boldsymbol{A}$ is symmetric, these expressions reduce to $\nabla f(\boldsymbol{x}_0) = 2\boldsymbol{A}\boldsymbol{x}_0$, $\nabla^2 f(\boldsymbol{x}_0) = 2\boldsymbol{A}$.

Convex analysis

III

3.1 Convexity of sets

Definition 3.1 (convex set) *Let $S \subseteq \mathbb{R}^n$. The set S is* convex *if*

$$\left.\begin{array}{l} \boldsymbol{x}^1, \boldsymbol{x}^2 \in S \\ \lambda \in (0,1) \end{array}\right\} \quad \Longrightarrow \quad \lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2 \in S$$

holds. ∎

A set S is convex if, from everywhere in S, all other points of S are "visible."

Figure 3.1 illustrates a convex set.

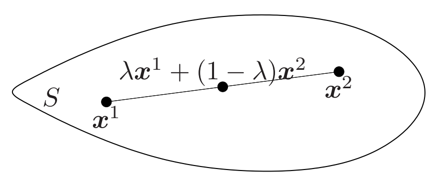


Figure 3.1: A convex set. (For the intermediate vector shown, the value of λ is $\approx 1/2$.)

Two non-convex sets are shown in Figure 3.2.

Example 3.2 (convex and non-convex sets) By using the definition of a convex set, the following can be established:

(a) The empty set is a convex set.

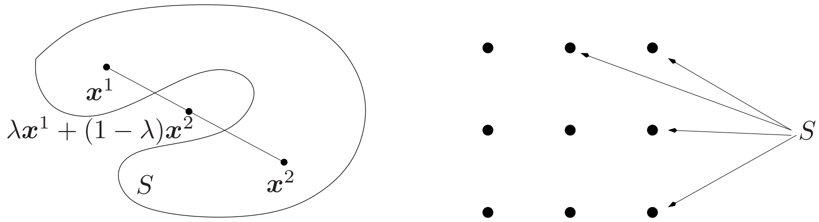


Figure 3.2: Two non-convex sets.

(b) The set $\{ \boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| \leq a \}$ is convex for every value of $a \in \mathbb{R}$. (Note: $\|\cdot\|$ here denotes any vector norm, but we will almost always use the 2-norm,

$$\|\boldsymbol{x}\|_2 := \sqrt{\sum_{j=1}^n x_j^2}.$$

We will most often not write the index $_2$, but instead use the 2-norm implicitly whenever writing $\|\cdot\|$.)

(c) The set $\{ \boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| = a \}$ is non-convex for every $a > 0$.

(d) The set $\{0, 1, 2\}$ is non-convex. (The second illustration in Figure 3.2 is such a case of a set of integral points in $\mathbb{R}^2$.) ■

Proposition 3.3 (convex intersection) *Suppose that S_k, $k \in \mathcal{K}$, is any collection of convex sets. Then, the intersection*

$$S := \cap_{k \in \mathcal{K}} S_k$$

is a convex set.

Proof. Let both $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$ belong to S. (If two such points cannot be found, then the result holds vacuously.) Then, $\boldsymbol{x}^1 \in S_k$ and $\boldsymbol{x}^2 \in S_k$ for all $k \in \mathcal{K}$. Take $\lambda \in (0, 1)$. Then, $\lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2 \in S_k, k \in \mathcal{K}$, by the convexity of the sets S_k. So, $\lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2 \in \cap_{k \in \mathcal{K}} S_k = S$. ■

3.2 Polyhedral theory

3.2.1 Convex hulls

Consider the set $V = \{\boldsymbol{v}^1, \boldsymbol{v}^2\}$, where $\boldsymbol{v}^1, \boldsymbol{v}^2 \in \mathbb{R}^n$ and $\boldsymbol{v}^1 \neq \boldsymbol{v}^2$. A set naturally related to V is the line in $\mathbb{R}^n$ through $\boldsymbol{v}^1$ and $\boldsymbol{v}^2$ [see Figure 3.3(b)], that is, $\{ \lambda \boldsymbol{v}^1 + (1-\lambda)\boldsymbol{v}^2 \mid \lambda \in \mathbb{R} \} = \{ \lambda_1 \boldsymbol{v}^1 + \lambda_2 \boldsymbol{v}^2 \mid \lambda_1, \lambda_2 \in$

$\mathbb{R};\ \lambda_1+\lambda_2=1\,\}$. Another set naturally related to V is the line segment between $\boldsymbol{v}^1$ and $\boldsymbol{v}^2$ [see Figure 3.3(c)], that is, $\{\,\lambda\boldsymbol{v}^1+(1-\lambda)\boldsymbol{v}^2 \mid \lambda\in[0,1]\,\}=\{\,\lambda_1\boldsymbol{v}^1+\lambda_2\boldsymbol{v}^2 \mid \lambda_1,\lambda_2\geq 0;\ \lambda_1+\lambda_2=1\,\}$. Motivated by this we define the *affine hull* and the *convex hull* of a set in $\mathbb{R}^n$.

Definition 3.4 (affine hull) *The* affine hull *of a finite set $V=\{\boldsymbol{v}^1,\dots,\boldsymbol{v}^k\}\subset\mathbb{R}^n$ is the set*

$$\operatorname{aff} V := \left\{ \lambda_1\boldsymbol{v}^1+\dots+\lambda_k\boldsymbol{v}^k \;\middle|\; \lambda_1,\dots,\lambda_k\in\mathbb{R};\ \sum_{i=1}^k \lambda_i = 1 \right\}.$$

The affine hull *of an infinite set $V\subseteq\mathbb{R}^n$ is the smallest affine subspace that includes V.*

A point $\lambda_1\boldsymbol{v}^1+\dots+\lambda_k\boldsymbol{v}^k$, where $\boldsymbol{v}^1,\dots,\boldsymbol{v}^k\in V$ and $\lambda_1,\dots,\lambda_k\in\mathbb{R}$ such that $\sum_{i=1}^k\lambda_i=1$, is called an affine combination *of the points $\boldsymbol{v}^1,\dots,\boldsymbol{v}^k$ (the number k of points in the sum must be finite).* ■

Definition 3.5 (convex hull) *The* convex hull *of a finite set $V=\{\boldsymbol{v}^1,\dots,\boldsymbol{v}^k\}\subset\mathbb{R}^n$ is the set*

$$\operatorname{conv} V := \left\{ \lambda_1\boldsymbol{v}^1+\dots+\lambda_k\boldsymbol{v}^k \;\middle|\; \lambda_1,\dots,\lambda_k\geq 0;\ \sum_{i=1}^k \lambda_i = 1 \right\}.$$

The convex hull *of an infinite set $V\subseteq\mathbb{R}^n$ is the smallest convex set that includes V.*

A point $\lambda_1\boldsymbol{v}^1+\dots+\lambda_k\boldsymbol{v}^k$, where $\boldsymbol{v}^1,\dots,\boldsymbol{v}^k\in V$ and $\lambda_1,\dots,\lambda_k\geq 0$ such that $\sum_{i=1}^k\lambda_i=1$, is called a convex combination *of the points $\boldsymbol{v}^1,\dots,\boldsymbol{v}^k$ (the number k of points in the sum must be finite).* ■

Example 3.6 (affine hull, convex hull) (a) The affine hull of three or more points in $\mathbb{R}^2$ not all lying on the same line is $\mathbb{R}^2$ itself. The convex hull of five points in $\mathbb{R}^2$ is shown in Figure 3.4 (observe that the "corners" of the convex hull of the points are some of the points themselves).

(b) The affine hull of three points not all lying on the same line in $\mathbb{R}^3$ is the plane through the points.

(c) The affine hull of an affine space is the space itself and the convex hull of a convex set is the set itself. ■

From the definition of convex hull of a finite set it follows that the convex hull equals the set of all convex combinations of points in the set. It turns out that this also holds for infinite sets.

Proposition 3.7 *Let $V\subseteq\mathbb{R}^n$. Then,* $\operatorname{conv} V$ *is the set of all convex combinations of points of V.*

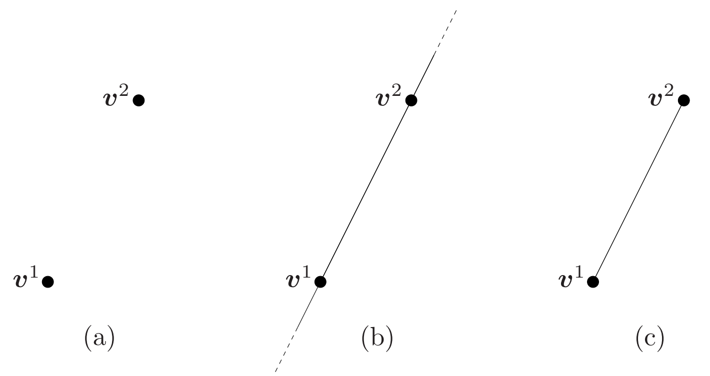


Figure 3.3: (a) The set V. (b) The set aff V. (c) The set conv V.

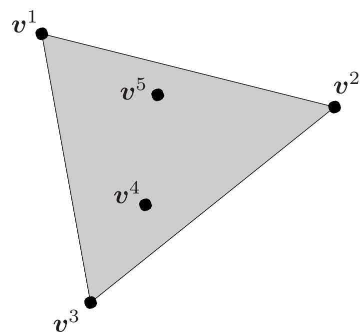


Figure 3.4: The convex hull of five points in $\mathbb{R}^2$.

Proof. Let Q be the set of all convex combinations of points of V. The inclusion $Q \subseteq \operatorname{conv} V$ follows from the definition of a convex set (since $\operatorname{conv} V$ is a convex set). We next show that Q is a convex set. If $\boldsymbol{x}^1, \boldsymbol{x}^2 \in Q$, then $\boldsymbol{x}^1 = \alpha_1 \boldsymbol{a}^1 + \cdots + \alpha_k \boldsymbol{a}^k$ and $\boldsymbol{x}^2 = \beta_1 \boldsymbol{b}^1 + \cdots + \beta_m \boldsymbol{b}^m$ for some $\boldsymbol{a}^1, \ldots, \boldsymbol{a}^k, \boldsymbol{b}^1, \ldots, \boldsymbol{b}^m \in V$ and $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_m \geq 0$ such that $\sum_{i=1}^k \alpha_i = \sum_{i=1}^m \beta_i = 1$. Let $\lambda \in (0, 1)$. Then

$$\begin{aligned}\lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2 = {} & \lambda\alpha_1 \boldsymbol{a}^1 + \cdots + \lambda\alpha_k \boldsymbol{a}^k \\ & + (1-\lambda)\beta_1 \boldsymbol{b}^1 + \cdots + (1-\lambda)\beta_m \boldsymbol{b}^m,\end{aligned}$$

and since $\lambda\alpha_1 + \cdots + \lambda\alpha_k + (1-\lambda)\beta_1 + \cdots + (1-\lambda)\beta_m = 1$, we have that $\lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2 \in Q$, so Q is convex. Since Q is convex and $V \subseteq Q$ it follows that $\operatorname{conv} V \subseteq Q$ (from the definition of convex hull of an infinite set in $\mathbb{R}^n$ it follows that $\operatorname{conv} V$ is the smallest convex set that contains V). Therefore $Q = \operatorname{conv} V$. ■

Proposition 3.7 shows that every point of the convex hull of a set can be written as a convex combination of points from the set. It tells, however, nothing about how many points that are required. This is the content of Carathéodory's Theorem.

Theorem 3.8 (Carathéodory's Theorem) *Let $\boldsymbol{x} \in \operatorname{conv} V$, where $V \subseteq \mathbb{R}^n$. Then, $\boldsymbol{x}$ can be expressed as a convex combination of $n+1$ or fewer points of V.*

Proof. From Proposition 3.7 it follows that $\boldsymbol{x} = \lambda_1 \boldsymbol{a}^1 + \cdots + \lambda_m \boldsymbol{a}^m$ for some $\boldsymbol{a}^1, \dots, \boldsymbol{a}^m \in V$ and $\lambda_1, \dots, \lambda_m \geq 0$ such that $\sum_{i=1}^m \lambda_i = 1$. We assume that this representation of $\boldsymbol{x}$ is chosen so that $\boldsymbol{x}$ cannot be expressed as a convex combination of fewer than m points of V. It follows that no two of the points $\boldsymbol{a}^1, \dots, \boldsymbol{a}^m$ are equal and that $\lambda_1, \dots, \lambda_m > 0$. We prove the theorem by showing that $m \leq n+1$. Assume that $m > n+1$. Then the set $\{\boldsymbol{a}^1, \dots, \boldsymbol{a}^m\}$ must be affinely dependent, so there exist $\alpha_1, \dots, \alpha_m \in \mathbb{R}$, not all zero, such that $\sum_{i=1}^m \alpha_i \boldsymbol{a}^i = \mathbf{0}^n$ and $\sum_{i=1}^m \alpha_i = 0$. Let $\varepsilon > 0$ be such that $\lambda_1 + \varepsilon\alpha_1, \dots, \lambda_m + \varepsilon\alpha_m$ are non-negative with at least one of them zero (such an ε exists since the λ's are all positive and at least one of the α's must be negative). Now we have that $\boldsymbol{x} = \sum_{i=1}^m (\lambda_i + \varepsilon\alpha_i)\boldsymbol{a}^i$ and if terms with zero coefficients are omitted this is a representation of $\boldsymbol{x}$ with fewer than m points; this is a contradiction. ■

3.2.2 Polytopes

We are now ready to define the geometrical object *polytope*.

Definition 3.9 (polytope) *A subset P of $\mathbb{R}^n$ is a* polytope *if it is the convex hull of finitely many points in $\mathbb{R}^n$.* ■

Example 3.10 (polytopes) (a) The set shown in Figure 3.4 is a polytope.

(b) A cube and a tetrahedron are polytopes in $\mathbb{R}^3$. ■

We next show how to characterize a polytope as the convex hull of its *extreme points*.

Definition 3.11 (extreme point) *A point $\boldsymbol{v}$ of a convex set P is called an* extreme point *if whenever $\boldsymbol{v} = \lambda\boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2$, where $\boldsymbol{x}^1, \boldsymbol{x}^2 \in P$ and $\lambda \in (0,1)$, then $\boldsymbol{v} = \boldsymbol{x}^1 = \boldsymbol{x}^2$.* ■

Example 3.12 (extreme points) The set shown in Figure 3.3(c) has the extreme points $\boldsymbol{v}^1$ and $\boldsymbol{v}^2$. The set shown in Figure 3.4 has the extreme points $\boldsymbol{v}^1$, $\boldsymbol{v}^2$, and $\boldsymbol{v}^3$. The set shown in Figure 3.3(b) do not have any extreme points. ∎

Lemma 3.13 *Let P be the polytope* $\operatorname{conv} V$*, where* $V = \{\boldsymbol{v}^1, \dots, \boldsymbol{v}^k\} \subset \mathbb{R}^n$*. Then, each extreme point of P lies in V.*

Proof. Assume that $\boldsymbol{w} \notin V$ is an extreme point of P. We have that $\boldsymbol{w} = \sum_{i=1}^k \lambda_i \boldsymbol{v}^i$, for some $\lambda_i \geq 0$ such that $\sum_{i=1}^k \lambda_i = 1$. At least one of the λ_i's must be nonzero, say λ_1. If $\lambda_1 = 1$ then $\boldsymbol{w} = \boldsymbol{v}^1$, a contradiction, so $\lambda_1 \in (0, 1)$. We have that

$$\boldsymbol{w} = \lambda_1 \boldsymbol{v}^1 + (1 - \lambda_1) \sum_{i=2}^k \frac{\lambda_i}{1 - \lambda_1} \boldsymbol{v}^i.$$

Since $\sum_{i=2}^k \lambda_i/(1 - \lambda_1) = 1$ we have that $\sum_{i=2}^k \lambda_i/(1 - \lambda_1)\boldsymbol{v}^i \in P$, but $\boldsymbol{w}$ is an extreme point of P so $\boldsymbol{w} = \boldsymbol{v}^1$, a contradiction. ∎

Proposition 3.14 *Let P be the polytope* $\operatorname{conv} V$*, where* $V = \{\boldsymbol{v}^1, \dots, \boldsymbol{v}^k\} \subset \mathbb{R}^n$*. Then P is equal to the convex hull of its extreme points.*

Proof. Let Q be the set of extreme points of P. If $\boldsymbol{v}^i \in Q$ for all $i = 1, \dots, k$ we are done, so assume that $\boldsymbol{v}^1 \notin Q$. Then $\boldsymbol{v}^1 = \lambda \boldsymbol{u} + (1 - \lambda)\boldsymbol{w}$ for some $\lambda \in (0, 1)$ and $\boldsymbol{u}, \boldsymbol{w} \in P$, $\boldsymbol{u} \neq \boldsymbol{w}$. Further, $\boldsymbol{u} = \sum_{i=1}^k \alpha_i \boldsymbol{v}^i$ and $\boldsymbol{w} = \sum_{i=1}^k \beta_i \boldsymbol{v}^i$, for some $\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k \geq 0$ such that $\sum_{i=1}^k \alpha_i = \sum_{i=1}^k \beta_i = 1$. Hence

$$\boldsymbol{v}^1 = \lambda \sum_{i=1}^k \alpha_i \boldsymbol{v}^i + (1 - \lambda) \sum_{i=1}^k \beta_i \boldsymbol{v}^i = \sum_{i=1}^k (\lambda \alpha_i + (1 - \lambda)\beta_i) \boldsymbol{v}^i.$$

It must hold that $\alpha_1, \beta_1 \neq 1$, since otherwise $\boldsymbol{u} = \boldsymbol{w} = \boldsymbol{v}^1$, a contradiction. Therefore

$$\boldsymbol{v}^1 = \sum_{i=2}^k \frac{\lambda \alpha_i + (1 - \lambda)\beta_i}{1 - (\lambda \alpha_1 + (1 - \lambda)\beta_1)} \boldsymbol{v}^i,$$

and since $\sum_{i=2}^k (\lambda \alpha_i + (1 - \lambda)\beta_i)/(1 - \lambda \alpha_1 - (1 - \lambda)\beta_1) = 1$ it follows that $\operatorname{conv} V = \operatorname{conv}(V \setminus \{\boldsymbol{v}^1\})$. Similarly every $\boldsymbol{v}^i \notin Q$ can be removed and we end up with a set $T \subseteq V$ such that $\operatorname{conv} T = \operatorname{conv} V$ and $T \subseteq Q$. But from Lemma 3.13 we have that every extreme point of the set $\operatorname{conv} T$ lies in T and since $\operatorname{conv} T = \operatorname{conv} V$ it follows that Q is the set of extreme points of $\operatorname{conv} T$, so $Q \subseteq T$. Hence $T = Q$ and we are done. ∎

3.2.3 Polyhedra

Closely related to the polytope is the polyhedron. We will show that every polyhedron is the sum of a polytope and a convex cone. In the next subsection we show that a set is a polytope if and only if it is a bounded polyhedron.

Definition 3.15 (polyhedron) *A subset P of $\mathbb{R}^n$ is a* polyhedron *if there exist a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and a vector $\boldsymbol{b} \in \mathbb{R}^m$ such that*

$$P = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \,\}.$$

■

The importance of polyhedra is obvious, since the set of feasible solutions of every linear programming problem is a polyhedron.

Example 3.16 (polyhedra) (a) Figure 3.5 shows the bounded polyhedron $P = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 \geq 2;\ x_1 + x_2 \leq 6;\ 2x_1 - x_2 \leq 4 \,\}$.

(b) The unbounded polyhedron $P = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1+x_2 \geq 2;\ x_1-x_2 \leq 2;\ 3x_1 - x_2 \geq 0 \,\}$ is shown in Figure 3.6. ■

Often it is hard to decide whether a point in a convex set is an extreme point or not. This is not the case for the polyhedron since there is an algebraic characterization of the extreme points of a polyhedron. Given an $\tilde{\boldsymbol{x}} \in \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \,\}$ we call to the rows of $\boldsymbol{A}\tilde{\boldsymbol{x}} \leq \boldsymbol{b}$ that are fulfilled with equality as the *equality subsystem* of $\boldsymbol{A}\tilde{\boldsymbol{x}} \leq \boldsymbol{b}$, and denote it by $\tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{b}}$. The number of rows in $\tilde{\boldsymbol{A}}$ is denoted by $\tilde{m}$.

Theorem 3.17 (algebraic characterization of extreme points) *Let $\tilde{\boldsymbol{x}} \in P = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \,\}$, where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ has $\operatorname{rank} \boldsymbol{A} = n$ and $\boldsymbol{b} \in \mathbb{R}^m$. Further, let $\tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{b}}$ be the equality subsystem of $\boldsymbol{A}\tilde{\boldsymbol{x}} \leq \boldsymbol{b}$. Then $\tilde{\boldsymbol{x}}$ is an extreme point of P if and only if $\operatorname{rank} \tilde{\boldsymbol{A}} = n$.*

Proof. [$\Longrightarrow$] Suppose that $\tilde{\boldsymbol{x}}$ is an extreme point of P. If $\boldsymbol{A}\tilde{\boldsymbol{x}} < \boldsymbol{b}$ then $\tilde{\boldsymbol{x}} + \varepsilon\boldsymbol{1}^n, \tilde{\boldsymbol{x}} - \varepsilon\boldsymbol{1}^n \in P$ if $\varepsilon > 0$ is sufficiently small. But $\tilde{\boldsymbol{x}} = 1/2(\tilde{\boldsymbol{x}}+\varepsilon\boldsymbol{1}^n)+1/2(\tilde{\boldsymbol{x}}-\varepsilon\boldsymbol{1}^n)$ which contradicts that $\tilde{\boldsymbol{x}}$ is an extreme point, so assume that at least one of the rows in $\boldsymbol{A}\tilde{\boldsymbol{x}} \leq \boldsymbol{b}$ is fulfilled with equality. If $\tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{b}}$ is the equality subsystem of $\boldsymbol{A}\tilde{\boldsymbol{x}} \leq \boldsymbol{b}$ and $\operatorname{rank} \tilde{\boldsymbol{A}} \leq n - 1$, then there exists a $\boldsymbol{w} \neq \boldsymbol{0}^n$ such that $\tilde{\boldsymbol{A}}\boldsymbol{w} = \boldsymbol{0}^{\tilde{m}}$, so $\tilde{\boldsymbol{x}} + \varepsilon\boldsymbol{w}, \tilde{\boldsymbol{x}} - \varepsilon\boldsymbol{w} \in P$ if $\varepsilon > 0$ is sufficiently small. But $\tilde{\boldsymbol{x}} = 1/2(\tilde{\boldsymbol{x}}+\varepsilon\boldsymbol{w}) + 1/2(\tilde{\boldsymbol{x}} - \varepsilon\boldsymbol{w})$, which contradicts that $\tilde{\boldsymbol{x}}$ is an extreme point. Hence $\operatorname{rank} \tilde{\boldsymbol{A}} = n$.

[$\Longleftarrow$] Assume that $\operatorname{rank} \tilde{\boldsymbol{A}} = n$. Then $\tilde{\boldsymbol{x}}$ is the unique solution to $\tilde{\boldsymbol{A}}\boldsymbol{x} = \tilde{\boldsymbol{b}}$. If $\tilde{\boldsymbol{x}}$ is not an extreme point of P it follows that $\tilde{\boldsymbol{x}} = \lambda\boldsymbol{u} + (1 - \lambda)\boldsymbol{v}$ for some $\lambda \in (0,1)$ and $\boldsymbol{u}, \boldsymbol{v} \in P$, $\boldsymbol{u} \neq \boldsymbol{v}$. This yields

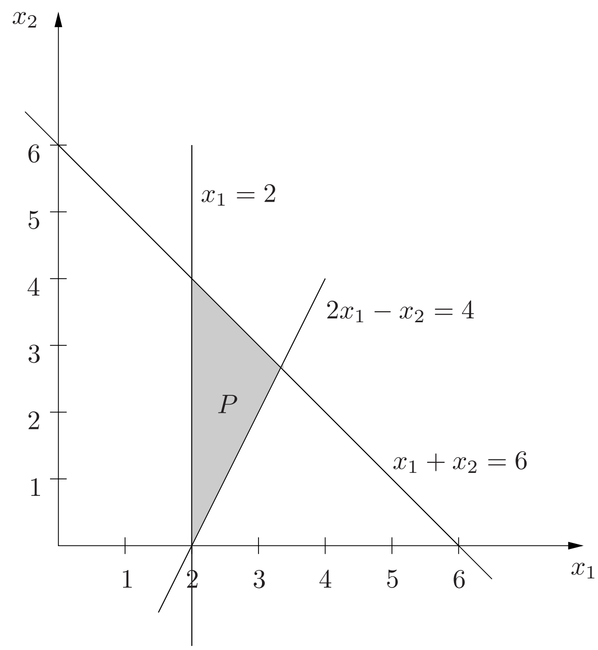


Figure 3.5: Illustration of the bounded polyhedron $P = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 \geq 2;\; x_1 + x_2 \leq 6;\; 2x_1 - x_2 \leq 4 \,\}$.

that $\lambda \tilde{\boldsymbol{A}}\boldsymbol{u} + (1-\lambda)\tilde{\boldsymbol{A}}\boldsymbol{v} = \tilde{\boldsymbol{b}}$, and since $\boldsymbol{A}\boldsymbol{u} \leq \boldsymbol{b}$ and $\boldsymbol{A}\boldsymbol{v} \leq \boldsymbol{b}$ it follows that $\tilde{\boldsymbol{A}}\boldsymbol{u} = \tilde{\boldsymbol{A}}\boldsymbol{v} = \tilde{\boldsymbol{b}}$, which contradicts that $\tilde{\boldsymbol{x}}$ is the unique solution to $\tilde{\boldsymbol{A}}\boldsymbol{x} = \tilde{\boldsymbol{b}}$. Therefore $\tilde{\boldsymbol{x}}$ must be an extreme point. ■

Corollary 3.18 *The number of extreme points of the polyhedron* $P = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \,\}$, *where* $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ *and* $\boldsymbol{b} \in \mathbb{R}^m$, *is finite.*

Proof. The theorem implies that the number of extreme points of P never exceeds the number of ways in which n objects can be chosen from a set of m objects, that is, the number of extreme points is less than or equal to

$$\binom{m}{n} = \frac{m!}{n!(m-n)!}.$$

■

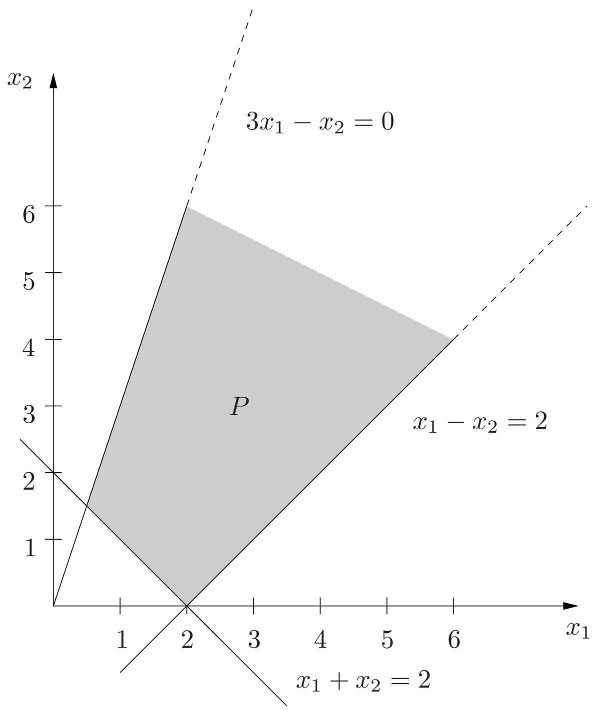


Figure 3.6: Illustration of the unbounded polyhedron $P = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 + x_2 \geq 2;\ x_1 - x_2 \leq 2;\ 3x_1 - x_2 \geq 0 \,\}$.

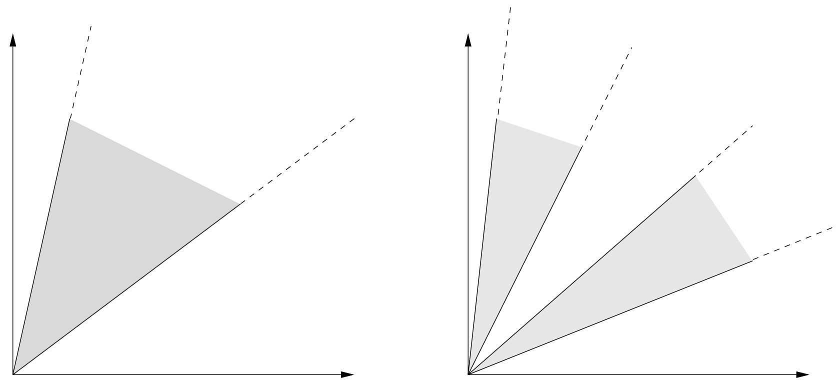
Remark 3.19 Since the number of extreme points is finite, the convex hull of the extreme points of a polyhedron is a polytope. ■

Definition 3.20 (cone) *A subset C of $\mathbb{R}^n$ is a* cone *if $\lambda \boldsymbol{x} \in C$ whenever $\boldsymbol{x} \in C$ and $\lambda > 0$.* ■

Example 3.21 (cone) (a) The set $\{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{0}^m \,\}$, where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, is a cone. Since this set is a polyhedron this type of cone is usually called a *polyhedral cone*.

(b) Figure 3.7(a) illustrates a convex cone and Figure 3.7(b) illustrates a non-convex cone in $\mathbb{R}^2$. ■

We have arrived at the most important theorem of this section, namely the Representation Theorem, which tells that every polyhedron is the sum of a polytope and a polyhedral cone. The Representation



(a) (b)

Figure 3.7: (a) A convex cone in $\mathbb{R}^2$. (b) A non-convex cone in $\mathbb{R}^2$.

Theorem will have great importance in the linear programming theory in Chapter 8.

Theorem 3.22 (Representation Theorem) *Let $Q = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{Ax} \leq \boldsymbol{b} \,\}$, where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$, P is the convex hull of the extreme points of Q, and $C = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{Ax} \leq \boldsymbol{0}^m \,\}$. If $\operatorname{rank} \boldsymbol{A} = n$ then $Q = P + C = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x} = \boldsymbol{u} + \boldsymbol{v}$ for some $\boldsymbol{u} \in P$ and $\boldsymbol{v} \in C \,\}$. In other words, every polyhedron (that has at least one extreme point) is the sum of a polytope and a polyhedral cone.*

Proof. Let $\tilde{\boldsymbol{x}} \in Q$ and $\tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{b}}$ be the corresponding equality subsystem of $\boldsymbol{A}\tilde{\boldsymbol{x}} \leq \boldsymbol{b}$. We prove the theorem by induction on the rank of $\tilde{\boldsymbol{A}}$.

If $\operatorname{rank} \tilde{\boldsymbol{A}} = n$ it follows from Theorem 3.17 that $\tilde{\boldsymbol{x}}$ is an extreme point of Q, so $\tilde{\boldsymbol{x}} \in P + C$, since $\boldsymbol{0}^n \in C$. Now assume that $\tilde{\boldsymbol{x}} \in P + C$ for all $\tilde{\boldsymbol{x}} \in Q$ with $k \leq \operatorname{rank} \tilde{\boldsymbol{A}} \leq n$, and choose $\tilde{\boldsymbol{x}} \in Q$ with $\operatorname{rank} \tilde{\boldsymbol{A}} = k - 1$. Then there exists $\boldsymbol{w} \neq \boldsymbol{0}^n$ such that $\tilde{\boldsymbol{A}}\boldsymbol{w} = \boldsymbol{0}^{\tilde{n}}$. If $|\lambda|$ is sufficiently small it follows that $\tilde{\boldsymbol{x}} + \lambda\boldsymbol{w} \in Q$. (Why?) If $\tilde{\boldsymbol{x}} + \lambda\boldsymbol{w} \in Q$ for all $\lambda \in \mathbb{R}$ we must have $\boldsymbol{Aw} = \boldsymbol{0}^n$ which implies $\operatorname{rank} \boldsymbol{A} \leq n - 1$, a contradiction. Suppose that there exists a largest λ^+ such that $\tilde{\boldsymbol{x}} + \lambda^+\boldsymbol{w} \in Q$. Then if $\bar{\boldsymbol{A}}(\tilde{\boldsymbol{x}} + \lambda^+\boldsymbol{w}) = \bar{\boldsymbol{b}}$ is the equality subsystem of $\boldsymbol{A}(\tilde{\boldsymbol{x}} + \lambda^+\boldsymbol{w}) \leq \boldsymbol{b}$ we must have $\operatorname{rank} \bar{\boldsymbol{A}} \geq k$. (Why?) By the induction hypothesis it then follows that $\tilde{\boldsymbol{x}} + \lambda^+\boldsymbol{w} \in P + C$. On the other hand, if $\tilde{\boldsymbol{x}} + \lambda\boldsymbol{w} \in Q$ for all $\lambda \geq 0$ then $\boldsymbol{Aw} \leq \boldsymbol{0}^m$, so $\boldsymbol{w} \in C$. Similarly, if $\tilde{\boldsymbol{x}} + \lambda(-\boldsymbol{w}) \in Q$ for all $\lambda \geq 0$ then $-\boldsymbol{w} \in C$, and if there exists a largest λ^- such that $\tilde{\boldsymbol{x}} + \lambda^-(-\boldsymbol{w}) \in Q$ then $\tilde{\boldsymbol{x}} + \lambda^-(-\boldsymbol{w}) \in P + C$.

Above we got a contradiction if none of λ^+ or λ^- existed. If only one of them exists, say λ^+, then $\tilde{\boldsymbol{x}} + \lambda^+ \boldsymbol{w} \in P + C$ and $-\boldsymbol{w} \in C$, and it follows that $\tilde{\boldsymbol{x}} \in P + C$. Otherwise, if both λ^+ and λ^- exist then $\tilde{\boldsymbol{x}} + \lambda^+ \boldsymbol{w} \in P + C$ and $\tilde{\boldsymbol{x}} + \lambda^-(-\boldsymbol{w}) \in P + C$, and $\tilde{\boldsymbol{x}}$ can be written as a convex combination of these points, which gives $\tilde{\boldsymbol{x}} \in P + C$. We have shown that $\tilde{\boldsymbol{x}} \in P + C$ for all $\tilde{\boldsymbol{x}} \in Q$ with $k - 1 \leq \operatorname{rank} \tilde{A} \leq n$ and the theorem follows by induction. ∎

Example 3.23 (illustration of the Representation Theorem) In Figure 3.8(a) we have a bounded polyhedron. The interior point $\tilde{\boldsymbol{x}}$ can be written as a convex combination of the extreme point $\boldsymbol{x}^5$ and the point $\boldsymbol{v}$ on the boundary, that is, there is a $\lambda \in (0, 1)$ such that

$$\tilde{\boldsymbol{x}} = \lambda \boldsymbol{x}^5 + (1 - \lambda)\boldsymbol{v}.$$

Further, the point $\boldsymbol{v}$ can be written as a convex combination of the extreme points $\boldsymbol{x}^2$ and $\boldsymbol{x}^3$, that is, there exists a $\mu \in (0, 1)$ such that

$$\boldsymbol{v} = \mu \boldsymbol{x}^2 + (1 - \mu)\boldsymbol{x}^3.$$

This gives that

$$\tilde{\boldsymbol{x}} = \lambda \boldsymbol{x}^5 + (1 - \lambda)\mu \boldsymbol{x}^2 + (1 - \lambda)(1 - \mu)\boldsymbol{x}^3,$$

and since $\lambda, (1 - \lambda)\mu, (1 - \lambda)(1 - \mu) \geq 0$ and

$$\lambda + (1 - \lambda)\mu + (1 - \lambda)(1 - \mu) = 1$$

we have that $\tilde{\boldsymbol{x}}$ lies in the convex hull of the extreme points $\boldsymbol{x}^2$, $\boldsymbol{x}^3$, and $\boldsymbol{x}^5$.

In Figure 3.8(b) we have an unbounded polyhedron. The interior point $\tilde{\boldsymbol{x}}$ can be written as a convex combination of the extreme point $\boldsymbol{x}^3$ and the point $\boldsymbol{v}$ on the boundary, that is, there exists a $\lambda \in (0, 1)$ such that

$$\tilde{\boldsymbol{x}} = \lambda \boldsymbol{x}^3 + (1 - \lambda)\boldsymbol{v}.$$

The point $\boldsymbol{v}$ lies on the halfline $\{\, \boldsymbol{x} \in \mathbb{R}^2 \mid \boldsymbol{x} = \boldsymbol{x}^2 + \mu(\boldsymbol{x}^1 - \boldsymbol{x}^2),\ \mu \geq 0 \,\}$. All the points on this halfline are feasible, which gives that if the polyhedron is given by $\{\, \boldsymbol{x} \in \mathbb{R}^2 \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \,\}$ then

$$\boldsymbol{A}(\boldsymbol{x}^2 + \mu(\boldsymbol{x}^1 - \boldsymbol{x}^2)) = \boldsymbol{A}\boldsymbol{x}^2 + \mu \boldsymbol{A}(\boldsymbol{x}^1 - \boldsymbol{x}^2) \leq \boldsymbol{b}, \qquad \forall \mu \geq 0.$$

But then we must have that $\boldsymbol{A}(\boldsymbol{x}^1 - \boldsymbol{x}^2) \leq \boldsymbol{0}^2$ since otherwise some component of $\mu \boldsymbol{A}(\boldsymbol{x}^1 - \boldsymbol{x}^2)$ tends to infinity as μ tends to infinity. Therefore

$\boldsymbol{x}^1 - \boldsymbol{x}^2$ lies in the cone $C = \{ \boldsymbol{x} \in \mathbb{R}^2 \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{0}^2 \}$. Now there exists a $\mu \in (0,1)$ such that

$$\boldsymbol{v} = \boldsymbol{x}^2 + \mu(\boldsymbol{x}^1 - \boldsymbol{x}^2),$$

and it follows that

$$\tilde{\boldsymbol{x}} = \lambda \boldsymbol{x}^3 + (1-\lambda)\boldsymbol{x}^2 + (1-\lambda)\mu(\boldsymbol{x}^1 - \boldsymbol{x}^2),$$

so since $(1-\lambda)\mu \geq 0$ and $\boldsymbol{x}^1 - \boldsymbol{x}^2 \in C$, $\tilde{\boldsymbol{x}}$ is the sum of a point in the convex hull of the extreme points and a point in the polyhedral cone C. ∎

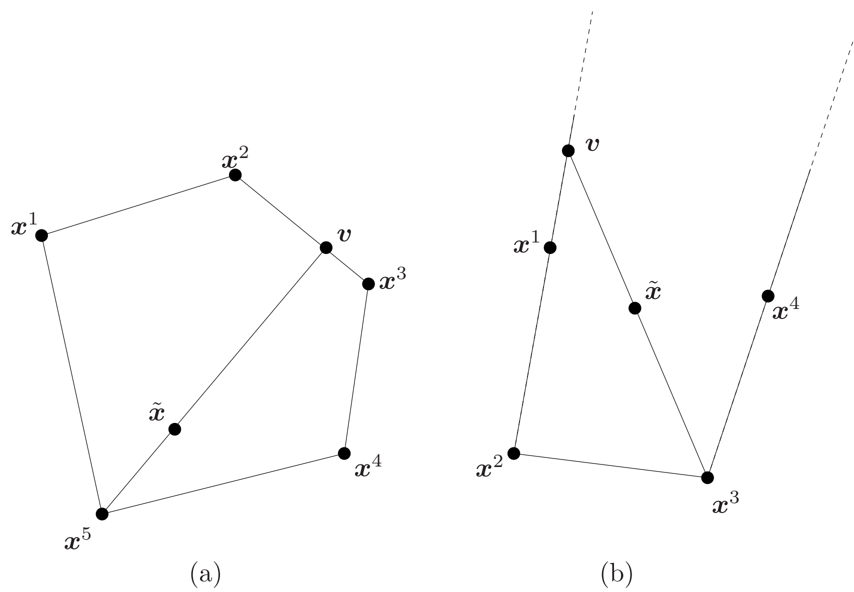


Figure 3.8: Illustration of the Representation Theorem (a) in the bounded case, and (b) in the unbounded case.

3.2.4 The Separation Theorem and Farkas' Lemma

We introduce the important concept of separation and use it to show that every polytope is a polyhedron.

Theorem 3.24 (Separation Theorem) *Suppose that the set $C \subseteq \mathbb{R}^n$ is closed and convex, and that the point $\boldsymbol{y}$ does not lie in C. Then there* *exist a vector* $\boldsymbol{\pi} \neq \mathbf{0}^n$ *and* $\alpha \in \mathbb{R}$ *such that* $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{y} > \alpha$ *and* $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{x} \leq \alpha$ *for all* $\boldsymbol{x} \in C$. ■

We postpone the proof of this theorem since it requires the Weierstrass Theorem 4.6 and the first order necessary optimality condition given in Proposition 4.22.b. Instead the proof is presented in Section 4.4.

The Separation Theorem is easy to describe geometrically: If a point $\boldsymbol{y}$ does not lie in a closed and convex set C, then there exists a hyperplane that separates $\boldsymbol{y}$ from C.

Example 3.25 (illustration of the Separation Theorem) Consider the closed and convex set $C = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid \|\boldsymbol{x}\| \leq 1 \,\}$ (i.e., C is the unit disc in $\mathbb{R}^2$), and the point $\boldsymbol{y} = (1.5, 1.5)$. Since $\boldsymbol{y} \notin C$ the Separation Theorem tells that there exists a line in $\mathbb{R}^2$ that separates $\boldsymbol{y}$ from C. This line is however not unique. In Figure 3.9 we see that the line given by $\boldsymbol{\pi} = (1, 1)^{\mathrm{T}}$ and $\alpha = 2$ is a candidate. (The proof of Theorem 3.24 actually constructs a tangent plane to C.) ■

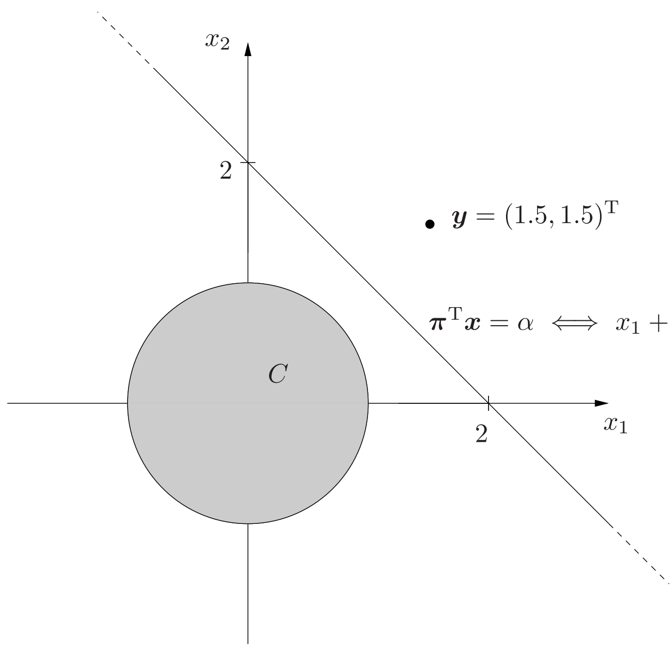


Figure 3.9: Illustration of the Separation Theorem: the unit disk is separated from $\boldsymbol{y}$ by the line $\{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 + x_2 = 2 \,\}$.

Theorem 3.26 *A set P is a polytope if and only if it is a bounded polyhedron.*

Proof. [$\Longleftarrow$] From the Representation Theorem 3.22 we get that a bounded polyhedron is the convex hull of its extreme points and hence by Remark 3.19 a polytope.

[$\Longrightarrow$] Let $V = \{\boldsymbol{v}^1, \dots, \boldsymbol{v}^k\} \subset \mathbb{R}^n$ and let P be the polytope $\operatorname{conv} V$. In order to prove that P is a polyhedron we must show that P is the solution set of some system of linear inequalities. The idea of the proof is to define a bounded polyhedron consisting of the coefficients and right-hand sides of all valid inequalities for P and then apply the Representation Theorem to select a finite subset of those valid inequalities.

To carry this out, consider the set $Q \subset \mathbb{R}^{n+1}$ defined as

$$\left\{ \begin{pmatrix} \boldsymbol{a} \\ b \end{pmatrix} \;\middle|\; \boldsymbol{a} \in \mathbb{R}^n;\, b \in \mathbb{R};\, -\mathbf{1}^n \leq \boldsymbol{a} \leq \mathbf{1}^n;\, -1 \leq b \leq 1;\, \boldsymbol{a}^{\mathrm{T}}\boldsymbol{v} \leq b,\, \boldsymbol{v} \in V \right\}.$$

Since V is a finite set, Q is a polyhedron. Further Q is bounded, so by the Representation Theorem we know that Q is the convex hull of its extreme points

$$\begin{pmatrix} \boldsymbol{a}^1 \\ b_1 \end{pmatrix}, \dots, \begin{pmatrix} \boldsymbol{a}^m \\ b_m \end{pmatrix}.$$

We will prove that the linear system

$$(\boldsymbol{a}^1)^{\mathrm{T}}\boldsymbol{x} \leq b_1, \dots, (\boldsymbol{a}^m)^{\mathrm{T}}\boldsymbol{x} \leq b_m, \tag{3.1}$$

defines P. We first show that P is contained in the solution set of (3.1). So, suppose that $\tilde{\boldsymbol{x}} \in P$. Then $\tilde{\boldsymbol{x}} = \lambda_1 \boldsymbol{v}^1 + \cdots + \lambda_k \boldsymbol{v}^k$ for some $\lambda_1, \dots, \lambda_k \geq 0$ such that $\sum_{i=1}^k \lambda_i = 1$. Thus, for each $i = 1, \dots, m$, we have

$$\begin{aligned}(\boldsymbol{a}^i)^{\mathrm{T}}\tilde{\boldsymbol{x}} = (\boldsymbol{a}^i)^{\mathrm{T}}(\lambda_1 \boldsymbol{v}^1 + \cdots + \lambda_k \boldsymbol{v}^k) &= \lambda_1 (\boldsymbol{a}^i)^{\mathrm{T}}\boldsymbol{v}^1 + \cdots + \lambda_k (\boldsymbol{a}^i)^{\mathrm{T}}\boldsymbol{v}^k \\ &\leq \lambda_1 b_i + \cdots + \lambda_k b_i = b_i,\end{aligned}$$

so $\tilde{\boldsymbol{x}}$ satisfies all inequalities in (3.1).

In order to show that the solution set of (3.1) is contained in P, let $\tilde{\boldsymbol{x}}$ be a solution to (3.1) and suppose that $\tilde{\boldsymbol{x}} \notin P$. Then, by the Separation Theorem 3.24 there exist a vector $\boldsymbol{\pi} \neq \mathbf{0}^n$ and $\alpha \in \mathbb{R}$ such that $\boldsymbol{\pi}^{\mathrm{T}}\tilde{\boldsymbol{x}} > \alpha$ and $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{x} \leq \alpha$ for all $\boldsymbol{x} \in P$. By scaling $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{x} \leq \alpha$ by a positive constant if necessary, we may assume that $-\mathbf{1}^n \leq \boldsymbol{\pi} \leq \mathbf{1}^n$ and $-1 \leq \alpha \leq 1$. That is, we may assume that $\begin{pmatrix} \boldsymbol{\pi} \\ \alpha \end{pmatrix} \in Q$. So we may write

$$\begin{pmatrix} \boldsymbol{\pi} \\ \alpha \end{pmatrix} = \lambda_1 \begin{pmatrix} \boldsymbol{a}^1 \\ b_1 \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} \boldsymbol{a}^m \\ b_m \end{pmatrix},$$

for some $\lambda_1, \ldots, \lambda_m \geq 0$ such that $\sum_{i=1}^m \lambda_i = 1$. Therefore,

$$\boldsymbol{\pi}^{\mathrm{T}}\tilde{\boldsymbol{x}} = \lambda_1(\boldsymbol{a}^1)^{\mathrm{T}}\tilde{\boldsymbol{x}} + \cdots + \lambda_m(\boldsymbol{a}^m)^{\mathrm{T}}\tilde{\boldsymbol{x}} \leq \lambda_1 b_1 + \cdots + \lambda_m b_m = \alpha.$$

But this is a contradiction, since $\boldsymbol{\pi}^{\mathrm{T}}\tilde{\boldsymbol{x}} > \alpha$. So $\tilde{\boldsymbol{x}} \in P$, which completes the proof. ■

We introduce the concept of finitely generated cones. In the proof of Farkas' Lemma below we will use that finitely generated cones are convex and closed and in order to show that we prove that finitely generated cones are polyhedral sets.

Definition 3.27 (finitely generated cone) *A* finitely generated cone *is one that is generated by a finite set, that is, a cone of the form*

$$\operatorname{cone}\{\boldsymbol{v}^1, \ldots, \boldsymbol{v}^m\} := \{\, \lambda_1\boldsymbol{v}^1 + \cdots + \lambda_m\boldsymbol{v}^m \mid \lambda_1, \ldots, \lambda_m \geq 0 \,\},$$

where $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^m \in \mathbb{R}^n$. Note that if $\boldsymbol{A}$ is an $m \times n$ matrix, then the set $\{\, \boldsymbol{y} \in \mathbb{R}^m \mid \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x},\ \boldsymbol{x} \geq \boldsymbol{0}^n \,\}$ is a finitely generated cone. ■

Recall that a cone that is a polyhedron is called a *polyhedral cone*. We show that a finitely generated cone is always a polyhedral cone and vice versa.

Theorem 3.28 *A convex cone in $\mathbb{R}^n$ is finitely generated if and only if it is polyhedral.*

Proof. [$\Longrightarrow$] Assume that C is the finitely generated cone

$$\operatorname{cone}\{\boldsymbol{v}^1, \ldots, \boldsymbol{v}^m\},$$

where $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^m \in \mathbb{R}^n$. From Theorem 3.26 we know that polytopes are polyhedral sets, so $\operatorname{conv}\{\boldsymbol{0}^n, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^m\}$ is the solution set of some linear inequalities

$$(\boldsymbol{a}^1)^{\mathrm{T}}\boldsymbol{x} \leq b_1, \ldots, (\boldsymbol{a}^k)^{\mathrm{T}}\boldsymbol{x} \leq b_k. \tag{3.2}$$

Since the solution set of these inequalities contains $\boldsymbol{0}^n$ we must have $b_1, \ldots, b_k \geq 0$. We show that C is the polyhedral cone A that equals the solution set of the inequalities of (3.2) for which $b_i = 0$. Since $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^m \in A$ we have $C \subseteq A$. In order to show that $A \subseteq C$, assume that $\boldsymbol{w} \in A$. Then $\lambda\boldsymbol{w}$ is in the solution set of (3.2) if $\lambda > 0$ is sufficiently small. Hence there exists a $\lambda > 0$ such that

$$\begin{aligned}\lambda\boldsymbol{w} &\in \{\, \boldsymbol{x} \in \mathbb{R}^n \mid (\boldsymbol{a}^1)^{\mathrm{T}}\boldsymbol{x} \leq b_1, \ldots, (\boldsymbol{a}^k)^{\mathrm{T}}\boldsymbol{x} \leq b_k \,\}\\ &= \operatorname{conv}\{\boldsymbol{0}^n, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^m\} \subseteq C,\end{aligned}$$

so $\boldsymbol{w} \in (1/\lambda)C = C$. Hence $A \subseteq C$, and $C = A$.

[$\Longleftarrow$] Suppose that C is a polyhedral cone in $\mathbb{R}^n$. Let P be a polytope in $\mathbb{R}^n$ such that $\mathbf{0}^n \in \operatorname{int} P$ (that is, $\mathbf{0}^n$ lies in the interior of P). Then $C \cap P$ is a bounded polyhedron and hence the representation theorem gives that $C \cap P = \operatorname{conv}\{\boldsymbol{v}^1, \dots, \boldsymbol{v}^m\}$, where $\boldsymbol{v}^1, \dots, \boldsymbol{v}^m$ is the extreme points of $C \cap P$. We show that C is the finitely generated cone $\operatorname{cone}\{\boldsymbol{v}^1, \dots, \boldsymbol{v}^m\}$. Since $\boldsymbol{v}^1, \dots, \boldsymbol{v}^m \in C$ and C is a polyhedral cone we get that $\operatorname{cone}\{\boldsymbol{v}^1, \dots, \boldsymbol{v}^m\} \subseteq C$. If $\boldsymbol{c} \in C$, then, since $\mathbf{0}^n \in \operatorname{int} P$, there exists a $\lambda > 0$ such that $\lambda\boldsymbol{c} \in P$. Thus

$$\lambda\boldsymbol{c} \in C \cap P = \operatorname{conv}\{\boldsymbol{v}^1, \dots, \boldsymbol{v}^m\} \subseteq \operatorname{cone}\{\boldsymbol{v}^1, \dots, \boldsymbol{v}^m\},$$

and so $\boldsymbol{c} \in (1/\lambda)\operatorname{cone}\{\boldsymbol{v}^1, \dots, \boldsymbol{v}^m\} = \operatorname{cone}\{\boldsymbol{v}^1, \dots, \boldsymbol{v}^m\}$. Hence it follows that $C \subseteq \operatorname{cone}\{\boldsymbol{v}^1, \dots, \boldsymbol{v}^m\}$, and $C = \operatorname{cone}\{\boldsymbol{v}^1, \dots, \boldsymbol{v}^m\}$. ■

Corollary 3.29 *Finitely generated cones in $\mathbb{R}^n$ are convex and closed.*

Proof. Halfspaces, that is, sets of the form $\{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} \leq b \,\}$ for some vector $\boldsymbol{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$, are convex and closed. (Why?) By the theorem a finitely generated cone is the intersection of finitely many halfspaces and thus the corollary follows from Proposition 3.3 and the fact that intersections of closed sets are closed. ■

We close this section by proving the famous Farkas' Lemma by using the Separation Theorem 3.24 and the fact that finitely generated cones are convex and closed.

Theorem 3.30 (Farkas' Lemma) *Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. Then, exactly one of the systems*

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} &= \boldsymbol{b}, \\ \boldsymbol{x} &\geq \mathbf{0}^n, \end{aligned} \tag{I}$$

and

$$\begin{aligned} \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\pi} &\leq \mathbf{0}^n, \\ \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\pi} &> 0, \end{aligned} \tag{II}$$

has a feasible solution, and the other system is inconsistent.

Proof. Let $C = \{\, \boldsymbol{y} \in \mathbb{R}^m \mid \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x},\ \boldsymbol{x} \geq \mathbf{0}^n \,\}$. If (I) is infeasible then $\boldsymbol{b} \notin C$. The set C is a finitely generated cone. Hence, by Corollary 3.29, it follows that C is convex and closed so by the Separation Theorem 3.24

there exist a vector $\boldsymbol{\pi} \in \mathbb{R}^m$ and $\alpha \in \mathbb{R}$ such that $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\pi} > \alpha$ and $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\pi} \leq \alpha$ for all $\boldsymbol{y} \in C$, that is,

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\pi} \leq \alpha, \qquad \forall \boldsymbol{x} \geq \boldsymbol{0}^n. \tag{3.3}$$

Since $\boldsymbol{0}^m \in C$ it follows that $\alpha \geq 0$, so $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\pi} > 0$, and if there exists an $\tilde{\boldsymbol{x}} \geq \boldsymbol{0}^n$ such that $\tilde{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\pi} > 0$ then (3.3) cannot hold for any α (if $\lambda \geq 0$ then $\lambda\tilde{\boldsymbol{x}} \geq \boldsymbol{0}^n$ and $(\lambda\tilde{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\pi} = \lambda\tilde{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\pi}$ tends to infinity as λ tends to infinity). Therefore we must have that $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\pi} \leq 0$ for all $\boldsymbol{x} \geq \boldsymbol{0}^n$, and this holds if and only if $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\pi} \leq \boldsymbol{0}^n$, which means that (II) is feasible.

On the other hand, if (I) has a feasible solution, say $\tilde{\boldsymbol{x}} \geq \boldsymbol{0}^n$, then $\boldsymbol{A}\tilde{\boldsymbol{x}} = \boldsymbol{b}$, so if there is a solution to (II), say $\tilde{\boldsymbol{\pi}}$, then $\tilde{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\tilde{\boldsymbol{\pi}} = \boldsymbol{b}^{\mathrm{T}}\tilde{\boldsymbol{\pi}} > 0$. But then $\boldsymbol{A}^{\mathrm{T}}\tilde{\boldsymbol{\pi}} > \boldsymbol{0}^n$ (since $\tilde{\boldsymbol{x}} \geq \boldsymbol{0}^n$), a contradiction. Hence (II) is infeasible. ∎

3.3 Convex functions

Definition 3.31 (convex function) *Suppose that $S \subseteq \mathbb{R}^n$. A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is* convex *at $\bar{\boldsymbol{x}} \in S$ if*

$$\left.\begin{array}{l} \boldsymbol{x} \in S \\ \lambda \in (0,1) \\ \lambda\bar{\boldsymbol{x}} + (1-\lambda)\boldsymbol{x} \in S \end{array}\right\} \Longrightarrow f(\lambda\bar{\boldsymbol{x}} + (1-\lambda)\boldsymbol{x}) \leq \lambda f(\bar{\boldsymbol{x}}) + (1-\lambda)f(\boldsymbol{x}).$$

The function f is convex *on S if it is convex at every $\bar{\boldsymbol{x}} \in S$.* ∎

In other words, a convex function is such that a linear interpolation never is lower than the function itself.[1]

From the definition follows that a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex on a convex set $S \subseteq \mathbb{R}^n$ if and only if

$$\left.\begin{array}{l} \boldsymbol{x}^1, \boldsymbol{x}^2 \in S \\ \lambda \in (0,1) \end{array}\right\} \Longrightarrow f(\lambda\boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x})^2 \leq \lambda f(\boldsymbol{x}^1) + (1-\lambda)f(\boldsymbol{x}^2).$$

Definition 3.32 (concave function) *Suppose that $S \subseteq \mathbb{R}^n$. A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is* concave *at $\bar{\boldsymbol{x}} \in S$ if $-f$ is convex at $\bar{\boldsymbol{x}}$.*

The function f is concave *on S if it is concave at every $\bar{\boldsymbol{x}} \in S$.* ∎

[1] Words like "lower" and "above" should be understood in the sense of the comparison between the y-coordinates of the respective function at the same coordinates in x.

Definition 3.33 (strictly convex (concave) function) *A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is* strictly convex at *$\bar{\boldsymbol{x}} \in S$ if*

$$\left.\begin{array}{l} \boldsymbol{x} \in S,\ \boldsymbol{x} \neq \bar{\boldsymbol{x}} \\ \lambda \in (0,1) \\ \lambda\bar{\boldsymbol{x}} + (1-\lambda)\boldsymbol{x} \in S \end{array}\right\} \Longrightarrow f(\lambda\bar{\boldsymbol{x}} + (1-\lambda)\boldsymbol{x}) < \lambda f(\bar{\boldsymbol{x}}) + (1-\lambda) f(\boldsymbol{x}).$$

The function f strictly convex (concave) on *S if it is strictly convex (concave) at every $\bar{\boldsymbol{x}} \in S$.* ∎

In other words, a strictly convex function is such that a linear interpolation is strictly above the function itself.

Figure 3.10 illustrates a strictly convex function.

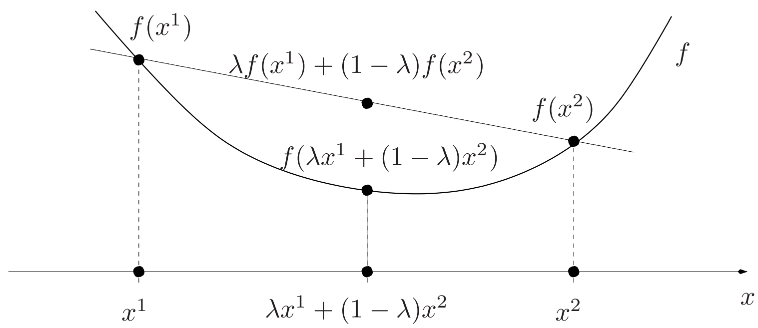


Figure 3.10: A convex function.

Example 3.34 (convex functions) *By using the definition of a convex function, the following can be established:*

(a) *The function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(\boldsymbol{x}) := \|\boldsymbol{x}\|$ is convex on $\mathbb{R}^n$.*

(b) *Let $\boldsymbol{c} \in \mathbb{R}^n$, $a \in \mathbb{R}$. The affine function $\boldsymbol{x} \mapsto f(\boldsymbol{x}) := \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + a = \sum_{j=1}^n c_j x_j + a$ is both* convex *and* concave *on $\mathbb{R}^n$. These are also the only finite functions that are both convex and concave.* ∎

Figure 3.11 illustrates a non-convex function.

Proposition 3.35 (sums of convex functions) *Suppose that $S \subseteq \mathbb{R}^n$. Let f_k, $k \in \mathcal{K}$, with $\mathcal{K}$ finite, be a collection of functions $f_k : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. Let $\alpha_k \geq 0$, $k \in \mathcal{K}$. If each function f_k, $k \in \mathcal{K}$, is convex at $\bar{\boldsymbol{x}} \in S$, then so is the function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ defined by $f(\boldsymbol{x}) := \sum_{k \in \mathcal{K}} \alpha_k f_k(\boldsymbol{x})$.*

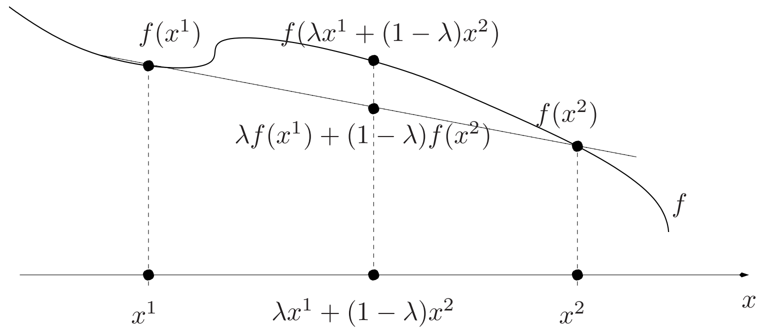


Figure 3.11: A non-convex function.

Proof. The proof is left as an exercise. ∎

Proposition 3.36 (convexity of composite functions) *Suppose that $S \subseteq \mathbb{R}^n$ and $P \subseteq \mathbb{R}$. Let further $g : S \to \mathbb{R}$ be a function which is convex on S, and $f : P \to \mathbb{R}$ be convex and non-decreasing $[y \geq x \Longrightarrow f(y) \geq f(x)]$ on P. Then, the composite function $f(g)$ is convex on the set $\{\, \boldsymbol{x} \in \mathbb{R}^n \mid g(\boldsymbol{x}) \in P \,\}$.*

Proof. Let $\boldsymbol{x}^1, \boldsymbol{x}^2 \in S \cap \{\, \boldsymbol{x} \in \mathbb{R}^n \mid g(\boldsymbol{x}) \in P \,\}$, and $\lambda \in (0, 1)$. Then,

$$\begin{aligned} f(g(\lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2)) &\leq f(\lambda g(\boldsymbol{x}^1) + (1-\lambda) g(\boldsymbol{x}^2)) \\ &\leq \lambda f(g(\boldsymbol{x}^1)) + (1-\lambda) f(g(\boldsymbol{x}^2)), \end{aligned}$$

where the first inequality follows from the convexity of g and the property of f being increasing, and the second inequality from the convexity of f. ∎

The following example functions are important in the development of penalty methods in linear and nonlinear optimization; their convexity is crucial is developing a convergence theory for such algorithms.

Example 3.37 (convex composite functions) *Suppose that the function $g : \mathbb{R}^n \to \mathbb{R}$ is convex.*

(a) *The function $\boldsymbol{x} \mapsto -\log(-g(\boldsymbol{x}))$ is convex on the set $\{\, \boldsymbol{x} \in \mathbb{R}^n \mid g(\boldsymbol{x}) < 0 \,\}$. (This function will be of interest in the analysis of interior point methods; see Section 13.1.)*

(b) *The function $\boldsymbol{x} \mapsto -1/g(\boldsymbol{x})$ is convex on the set $\{\, \boldsymbol{x} \in \mathbb{R}^n \mid g(\boldsymbol{x}) < 0 \,\}$.*

[*Note:* This function is convex, but the above rule for composite functions cannot be used. Utilize the definition of a convex function instead.]

(b) *The function* $\boldsymbol{x} \mapsto 1/\log(-g(\boldsymbol{x}))$ *is convex on the set* $\{\, \boldsymbol{x} \in \mathbb{R}^n \mid g(\boldsymbol{x}) < -1 \,\}$.

[*Note:* This function is convex, but the above rule for composite functions cannot be used. Utilize the definition of a convex function instead. The domain of the function must here be limited, because $x \mapsto 1/x$ is convex only for positive x.] ■

We next characterize the convexity of a function on $\mathbb{R}^n$ by the convexity of its *epigraph* in $\mathbb{R}^{n+1}$.

[*Note:* the *graph* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the boundary of epi f, which still resides in $\mathbb{R}^{n+1}$. See Figure 3.12 for an example, corresponding to the convex function in Figure 3.10.]

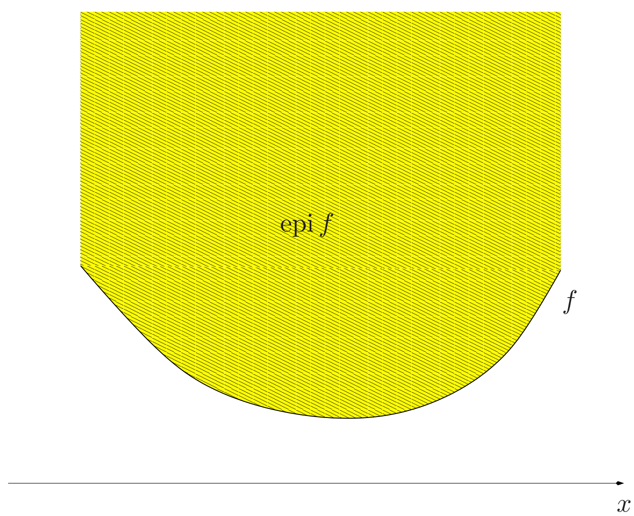


Figure 3.12: A convex function and its epigraph.

Definition 3.38 (epigraph) *The epigraph of a function* $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *is the set*

$$\operatorname{epi} f := \{\, (\boldsymbol{x}, \alpha) \in \mathbb{R}^{n+1} \mid f(\boldsymbol{x}) \le \alpha \,\}. \tag{3.4}$$

The epigraph of the function f *restricted to the set* $S \subseteq \mathbb{R}^n$ *is*

$$\operatorname{epi}_S f := \{\, (\boldsymbol{x}, \alpha) \in S \times \mathbb{R} \mid f(\boldsymbol{x}) \le \alpha \,\}. \tag{3.5}$$

■

Theorem 3.39 *Suppose that $S \subseteq \mathbb{R}^n$ is a convex set. Then, the function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex on S if, and only if, its epigraph restricted to S is a convex set in $\mathbb{R}^{n+1}$.*

Proof. [$\Longrightarrow$] Suppose that f is convex on S. Let $(\boldsymbol{x}^1, \alpha_1), (\boldsymbol{x}^2, \alpha_2) \in \operatorname{epi}_S f$. Let $\lambda \in (0,1)$. By the convexity of f on S,

$$\begin{aligned} f(\lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2) &\leq \lambda f(\boldsymbol{x}^1) + (1-\lambda) f(\boldsymbol{x}^2) \\ &\leq \lambda \alpha_1 + (1-\lambda)\alpha_2. \end{aligned}$$

Hence, $[\lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2, \lambda\alpha_1 + (1-\lambda)\alpha_2] \in \operatorname{epi}_S f$, so $\operatorname{epi}_S f$ is a convex set in $\mathbb{R}^{n+1}$.

[$\Longleftarrow$] Suppose that $\operatorname{epi}_S f$ is convex. Let $\boldsymbol{x}^1, \boldsymbol{x}^2 \in S$, whence

$$(\boldsymbol{x}^1, f(\boldsymbol{x}^1)), (\boldsymbol{x}^2, f(\boldsymbol{x}^2)) \in \operatorname{epi}_S f.$$

Let $\lambda \in (0,1)$. By the convexity of $\operatorname{epi}_S f$, it follows that

$$[\lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2, \lambda f(\boldsymbol{x}^1) + (1-\lambda) f(\boldsymbol{x}^2)] \in \operatorname{epi}_S f,$$

that is, $f(\lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2) \leq \lambda f(\boldsymbol{x}^1) + (1-\lambda) f(\boldsymbol{x}^2)$. Hence, f is convex on S. ■

When f is in C^1 (once differentiable, with continuous partial derivatives) or C^2 (twice differentiable, with continuous second partial derivatives), then convexity can be characterized also in terms of these derivatives. The results show how with stronger differentiability properties the characterizations become more and more useful in practice.

Theorem 3.40 (convexity characterizations in C^1) *Let $f \in C^1$ on an open convex set S.*

(a) *f is convex on $S \iff f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x})$, for all $\boldsymbol{x}, \boldsymbol{y} \in S$.*

(b) *f is convex on $S \iff [\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})]^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}) \geq 0$, for all $\boldsymbol{x}, \boldsymbol{y} \in S$.*

The result in (a) states, in words, that "every tangent plane to the function surface in $\mathbb{R}^{n+1}$ lies on, or below, the epigraph of f", or, that "a first-order approximation is below f."

The result in (b) states that ∇f is "monotone on S."

[*Note:* when $n = 1$, the result in (b) states that f is convex if and only if its derivative f' in non-decreasing, that is, that it is monotonically increasing.]

Proof. (a) [$\Longrightarrow$] Take $\boldsymbol{x}^1, \boldsymbol{x}^2 \in S$ and $\lambda \in (0,1)$. Then,

$$\begin{aligned}\lambda f(\boldsymbol{x}^1) + (1-\lambda) f(\boldsymbol{x}^2) &\geq f(\lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2) \\ &\Longleftrightarrow [\lambda > 0] \\ f(\boldsymbol{x}^1) - f(\boldsymbol{x}^2) &\geq (1/\lambda)[f(\lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2) - f(\boldsymbol{x}^2)].\end{aligned}$$

Let $\lambda \downarrow 0$. Then, the right-hand side of the above inequality tends to the directional derivative of f at $\boldsymbol{x}^2$ in the direction of $(\boldsymbol{x}^1 - \boldsymbol{x}^2)$, so that in the limit it becomes

$$f(\boldsymbol{x}^1) - f(\boldsymbol{x}^2) \geq \nabla f(\boldsymbol{x}^2)^{\mathrm{T}}(\boldsymbol{x}^1 - \boldsymbol{x}^2).$$

The result follows.

[$\Longleftarrow$] We have that

$$\begin{aligned}f(\boldsymbol{x}^1) &\geq f(\lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2) + (1-\lambda)\nabla f(\lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2)^{\mathrm{T}}(\boldsymbol{x}^1 - \boldsymbol{x}^2), \\ f(\boldsymbol{x}^2) &\geq f(\lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2) + \lambda \nabla f(\lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2)^{\mathrm{T}}(\boldsymbol{x}^2 - \boldsymbol{x}^1).\end{aligned}$$

Multiply the inequalities by λ and $(1-\lambda)$, respectively, and add them together to get the result sought.

(b) [$\Longrightarrow$] Using (a), and the two inequalities

$$\begin{aligned}f(\boldsymbol{y}) &\geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}), \qquad \boldsymbol{x}, \boldsymbol{y} \in S, \\ f(\boldsymbol{x}) &\geq f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}), \qquad \boldsymbol{x}, \boldsymbol{y} \in S,\end{aligned}$$

added together, yields that $[\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})]^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}) \geq 0$, for all $\boldsymbol{x}, \boldsymbol{y} \in S$.

[$\Longleftarrow$] The mean-value theorem states that

$$f(\boldsymbol{x}^2) - f(\boldsymbol{x}^1) = \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{x}^2 - \boldsymbol{x}^1),$$

where $\boldsymbol{x} = \lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2$ for some $\lambda \in (0,1)$. By assumption, $[\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^1)]^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^1) \geq 0$, so $(1-\lambda)[\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^1)]^{\mathrm{T}}(\boldsymbol{x}^2 - \boldsymbol{x}^1) \geq 0$. From this follows that $\nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{x}^2 - \boldsymbol{x}^1) \geq \nabla f(\boldsymbol{x}^1)^{\mathrm{T}}(\boldsymbol{x}^2 - x^1)$. Used above, we get that $f(\boldsymbol{x}^2) \geq f(\boldsymbol{x}^1) + \nabla f(\boldsymbol{x}^1)^{\mathrm{T}}(\boldsymbol{x}^2 - \boldsymbol{x}^1)$. We are done. ∎

Figure 3.13 illustrates part (a) of Theorem 3.40.

By replacing the inequalities in (a) and (b) in the theorem by *strict* inequalities, and adding the requirement that $\boldsymbol{x} \neq \boldsymbol{y}$ holds in the statements, we can establish a characterization also of *strictly* convex functions. The statement in (a) then says that the tangential hyperplane lies strictly below the function except at the tangent point, and (b) states that the gradient mapping is *strictly* monotone.

Still more can be said in C^2:

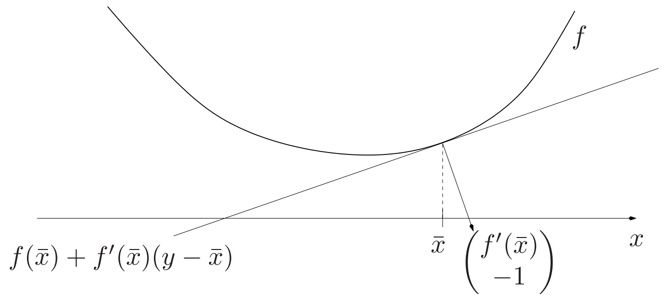


Figure 3.13: A tangent plane to the graph of a convex function.

Theorem 3.41 (convexity characterizations in C^2) *Let f be in C^2 on an open, convex set $S \subseteq \mathbb{R}^n$.*

(a) *f is convex on $S \iff \nabla^2 f(\boldsymbol{x})$ is positive semidefinite for all $\boldsymbol{x} \in S$.*

(b) *$\nabla^2 f(\boldsymbol{x})$ is positive definite for all $x \in S \Longrightarrow f$ is strictly convex on S.*

[*Note:* When $n = 1$ and S is an interval, the above reduce to the following familiar results: (a) f is convex on S if and only if $f''(x) \geq 0$ for every $x \in S$; (b) f is strictly convex on S if $f''(x) > 0$ for every $x \in S$.]

Proof.

(a) [$\Longrightarrow$] Suppose that f is convex and let $\bar{\boldsymbol{x}} \in S$. We must show that $\boldsymbol{p}^{\mathrm{T}} \nabla^2 f(\bar{\boldsymbol{x}}) \boldsymbol{p} \geq 0$ for all $\boldsymbol{p} \in \mathbb{R}^n$ holds.

Since S open, for any given $\boldsymbol{p} \in \mathbb{R}^n$, $\bar{\boldsymbol{x}} + \alpha \boldsymbol{p} \in S$ whenever $|\alpha| \neq 0$ is small enough. We utilize Theorem 3.40(a) as follows: by the twice differentiability of f,

$$f(\bar{\boldsymbol{x}} + \alpha \boldsymbol{p}) \geq f(\bar{\boldsymbol{x}}) + \alpha \nabla f(\bar{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{p}, \tag{3.6}$$

$$f(\bar{\boldsymbol{x}} + \alpha \boldsymbol{p}) = f(\bar{\boldsymbol{x}}) + \alpha \nabla f(\bar{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{p} + \frac{1}{2} \alpha^2 \boldsymbol{p}^{\mathrm{T}} \nabla^2 f(\bar{\boldsymbol{x}}) \boldsymbol{p} + o(\alpha^2). \tag{3.7}$$

Subtracting (3.7) from (3.6), we get

$$\frac{1}{2} \alpha^2 \boldsymbol{p}^{\mathrm{T}} \nabla^2 f(\bar{\boldsymbol{x}}) \boldsymbol{p} + o(\alpha^2) \geq 0.$$

Dividing by α^2 and letting $\alpha \downarrow 0$ it follows that $\boldsymbol{p}^{\mathrm{T}} \nabla^2 f(\bar{\boldsymbol{x}}) \boldsymbol{p} \geq 0$.

[$\Longleftarrow$] Suppose that the Hessian matrix is positive semi-definite at each point in S. The proof depends on the following second-order mean-value theorem: for every $\boldsymbol{x}, \boldsymbol{y} \in S$, there exists $\ell \in [0, 1]$ such that

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2} (\boldsymbol{y} - \boldsymbol{x})^{\mathrm{T}} \nabla^2 f[\boldsymbol{x} + \ell(\boldsymbol{y} - \boldsymbol{x})](\boldsymbol{y} - \boldsymbol{x}). \tag{3.8}$$

By assumption, the last term in (3.8) is non-negative, whence we obtain the convexity characterization in Theorem 3.40(a).

(b) [$\Longrightarrow$] By the assumptions, the last term in (3.8) is always positive when $\boldsymbol{y} \neq \boldsymbol{x}$, whence we obtain the strict convexity characterization in C^1. ∎

It is important to note that the opposite direction in the result (b) is false. A simple example that establishes this fact is the function defined by $f(x) = x^4$, $S = \mathbb{R}$; f is strictly convex on $\mathbb{R}$ (why?), but its second derivative at zero is $f''(0) = 0$!

The case of quadratic functions f is interesting to mention in particular. For quadratic functions, that is, functions of the form

$$f(\boldsymbol{x}) = (1/2)\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{q}^{\mathrm{T}}\boldsymbol{x} + a,$$

for some symmetric matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$, vector $\boldsymbol{q} \in \mathbb{R}^n$ and constant $a \in \mathbb{R}$, it holds that $\nabla^2 f(\boldsymbol{x}) \equiv \boldsymbol{Q}$ for every $\boldsymbol{x}$ where f is defined, so the value $\nabla^2 f(\boldsymbol{x})$ does not depend on $\boldsymbol{x}$. In this case, we can state a stronger result than in Theorem 3.41: *the quadratic function f is convex on the open, convex set $S \subseteq \mathbb{R}^n$ if and only if $\boldsymbol{Q}$ is positive semi-definite*; *f is strictly convex on S if and only if $\boldsymbol{Q}$ is positive definite*. To prove this result is simple from the above result for general C^2 functions, and is left as an exercise.

What happens when S is not full-dimensional (which is often the case)? Take, for example, $f(\boldsymbol{x}) := x_1^2 - x_2^2$ and $S := \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 \in \mathbb{R};\ x_2 = 0 \,\}$. Then, f is convex on S but $\nabla^2 f(\boldsymbol{x})$ is not positive semi-definite anywhere on S. The below result covers this type of case. Its proof is left as an exercise.

Theorem 3.42 (convexity characterizations in C^2, part II) *Let $S \subseteq \mathbb{R}^n$ be a nonempty convex set and $f : \mathbb{R}^n \to \mathbb{R}$ be in C^2 on $\mathbb{R}^n$. Let C be the subspace parallel to the affine hull of S. Then,*

$$f \text{ is convex on } S \quad \Longleftrightarrow \quad \boldsymbol{p}^{\mathrm{T}}\nabla^2 f(\boldsymbol{x})\boldsymbol{p} \geq 0 \text{ for every } \boldsymbol{x} \in S \text{ and } \boldsymbol{p} \in C.$$

In particular, when S has a nonempty interior, f is convex if and only if $\nabla^2 f(\boldsymbol{x})$ is positive semi-definite for every $\boldsymbol{x} \in S$. ∎

We have already seen that the convexity of a function is intimately connected to the convexity of a certain set, namely the epigraph of the function. The following result shows that a particular type of set, defined by those vectors that bound a convex function from above, is a convex set. Later, we will utilize this result to establish the convexity of feasible sets in some optimization problems.

Definition 3.43 (level set) *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a function. The* level set *of g with respect to the value $b \in \mathbb{R}$ is the set*

$$\operatorname{lev}_g(b) := \{\, x \in \mathbb{R}^n \mid g(x) \leq b \,\}. \tag{3.9}$$

■

Figure 3.14 illustrates a level set of a convex function.

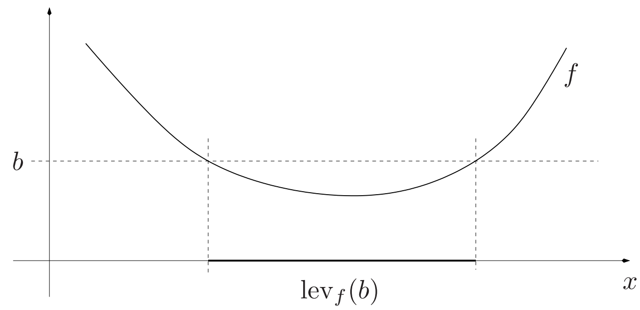


Figure 3.14: A level set of a convex function.

Proposition 3.44 (convex level sets from convex functions) *Suppose that the function $g : \mathbb{R}^n \to \mathbb{R}$ is convex. Then, for every value of $b \in \mathbb{R}$, the level set $\operatorname{lev}_g(b)$ is a convex set. It is moreover closed.*

Proof. The result follows immediately from the definitions of a convex set and a convex function. Let $\boldsymbol{x}^1, \boldsymbol{x}^2$ both satisfy the constraint that $g(\boldsymbol{x}) \leq b$ holds, and let $\lambda \in (0, 1)$. (If not two such points $\boldsymbol{x}^1, \boldsymbol{x}^2$ can be found, then the result holds vacuously.) Then, by the convexity of g, $g(\lambda \boldsymbol{x}^1 + (1 - \lambda)\boldsymbol{x}^2) \leq \lambda b + (1 - \lambda)b = b$, so the set $\operatorname{lev}_g(b)$ is convex.

The fact that a convex function which is defined on $\mathbb{R}^n$ is continuous establishes that the set $\operatorname{lev}_g(b)$ is always closed.[2] (Why?) ■

Definition 3.45 (convex problem) *Suppose that the set $X \subseteq \mathbb{R}^n$ is closed and convex. Suppose further that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and that the functions $g_i : \mathbb{R}^n \to \mathbb{R}$, $i \in \mathcal{I}$, all are concave. Suppose, finally, that the functions $g_i : \mathbb{R}^n \to \mathbb{R}$, $i \in \mathcal{E}$, all are affine. Then, the problem* (1.1) *is called a* convex problem. ■

[2] That convex functions are continuous will be established in Theorem 4.26.

The name is natural, because the objective function is a convex one, and the feasible set is closed and convex as well. In order to establish the latter, we refer first to Proposition 3.44 together with the concavity Definition 3.32 to establish that the inequality constraints define convex sets [note that in the problem (1.1) the inequalities are given as $\geq$-constraints], and ask the reader to prove that a constraint of the form $\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} = b_i$ defines a convex set as well. Finally, we refer to Proposition 3.3 to establish that the intersection of all the convex sets defined by S, $\mathcal{I}$, and $\mathcal{E}$ is convex.

3.4 Application: the projection of a vector onto a convex set

In Figure 3.15 we illustrate the Euclidean projection of some vectors onto a convex set.

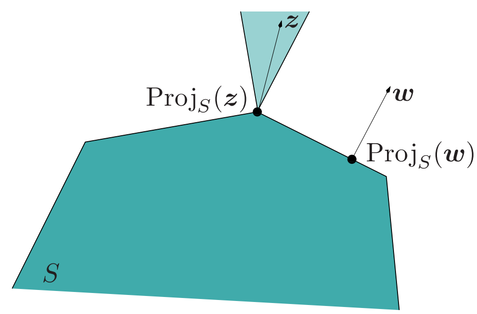


Figure 3.15: The projection of two vectors onto a convex set.

Starting with the vector $\boldsymbol{w}$, we see that its Euclidean projection corresponds to the vector in S which is nearest (in Euclidean norm) to $\boldsymbol{w}$; the vector $\boldsymbol{w} - \mathrm{Proj}_S(\boldsymbol{w})$ clearly is *normal* to the set S. The point $\boldsymbol{z}$ has the Euclidean projection $\mathrm{Proj}_S(\boldsymbol{z})$, but there are also several other vectors with the same projection; the figure shows in a special shading the set of vectors $\boldsymbol{z}$ which all have that same projection onto S. This set is a cone, which we refer to as the *normal cone* to S at $\boldsymbol{x} = \mathrm{Proj}_S(\boldsymbol{z})$. In the case of the point $\mathrm{Proj}_S(\boldsymbol{w})$ the normal cone reduces to a ray—which of course is also a cone. (The difference between these two sets is largely the consequence of the fact that there is only one constraint active at $\boldsymbol{w}$, while there are two constraints active at $\boldsymbol{z}$; when developing the KKT conditions in Chapter 5 we shall see how strongly the number of active constraints influence the appearance of the optimality conditions.)

We will also return to this image already in Section 4.6.3, because it contains the building blocks of the optimality conditions for an optimization problem with an objective function in C^1 over a convex set. For now, we will establish only one property of the projection operation Proj_S, namely that the *distance function*, dist_S, defined by

$$\mathrm{dist}_S(\boldsymbol{x}) := \|\boldsymbol{x} - \mathrm{Proj}_S(\boldsymbol{x})\|, \qquad \boldsymbol{x} \in \mathbb{R}^n, \tag{3.10}$$

is a convex function on $\mathbb{R}^n$. In particular, then, this function is continuous. (Later, we will establish also that the projection operation Proj_S is a well-defined operation whenever S is nonempty, closed and convex, and that the operation has particularly nice continuity properties. Before we can do so, however, we need to establish some results on the existence of optimal solutions.)

Let $\boldsymbol{x}^1, \boldsymbol{x}^2 \in \mathbb{R}^n$, and $\lambda \in (0,1)$. Then,

$$\begin{aligned}
\mathrm{dist}_S(\lambda\boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2) &= \|(\lambda\boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2) \\
&\quad - \mathrm{Proj}_S(\lambda\boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2)\| \\
&\leq \|(\lambda\boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2) \\
&\quad - (\lambda\mathrm{Proj}_S(\boldsymbol{x}^1) + (1-\lambda)\mathrm{Proj}_S(\boldsymbol{x}^2))\| \\
&\leq \lambda\|\boldsymbol{x}^1 - \mathrm{Proj}_S(\boldsymbol{x}^1)\| \\
&\quad + (1-\lambda)\|\boldsymbol{x}^2 - \mathrm{Proj}_S(\boldsymbol{x}^2)\| \\
&= \lambda\mathrm{dist}_S(\boldsymbol{x}^1) + (1-\lambda)\mathrm{dist}_S(\boldsymbol{x}^2),
\end{aligned}$$

where the first inequality comes from the fact that $\lambda\mathrm{Proj}_S(\boldsymbol{x}^1) + (1-\lambda)\mathrm{Proj}_S(\boldsymbol{x}^2) \in S$, but it does not necessarily define $\mathrm{Proj}_S(\lambda\boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2)$ (it may have a longer distance), and the second is the triangle inequality.

The proof is illustrated in Figure 3.16.

3.5 Notes and further reading

The subject of this chapter—convex analysis—has a long history, going back about a century. Much of the early work on convex sets and functions, for example, the theory of separation of convex sets, go back to the work of Minkowski [Min10, Min11]. More modern expositions are found in [Fen51, Roc70, StW70], which all are classical in the field. More easily accessible are the modern books [BoL00, BNO03]. Lighter introductions

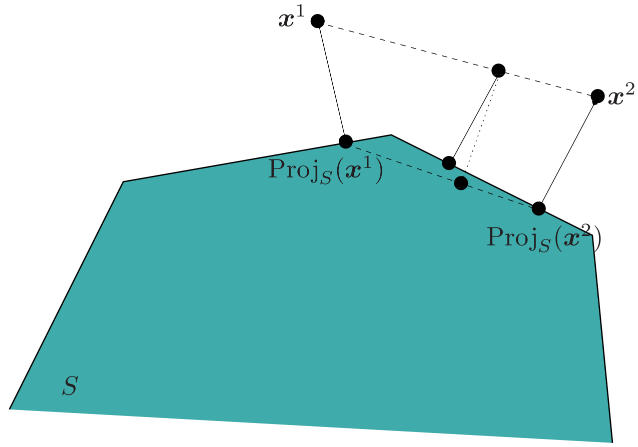


Figure 3.16: The distance function is convex. From the intermediate vector $\lambda\boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2$ shown the distance to the vector $\lambda\mathrm{Proj}_S(\boldsymbol{x}^1) + (1-\lambda)\mathrm{Proj}_S(\boldsymbol{x}^2))$ [the dotted line segment] clearly is longer than to its projection on S [shown as a solid line].

are also found in [BSS93, HiL93]. The most influencial of all of these books is *Convex Analysis* by R. T. Rockafellar [Roc70].

Carathéodory's Theorem 3.8 is found in [Car07, Car11]. Farkas' Lemma in Theorem 3.30 is due to Farkas [Far1902]. Theorem 3.42 is given as Exercise 1.8 in [BNO03].

The early history of polyhedral convexity is found in [Mot36].

3.6 Exercises

Exercise 3.1 (convexity of polyhedra) Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. Show that the polyhedron

$$P = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \,\},$$

is a convex set. ∎

Exercise 3.2 (polyhedra) Which of the following sets are polyhedra?

a) $S = \{y_1\boldsymbol{a} + y_2\boldsymbol{b} \mid -1 \leq y_1 \leq 1, \ -1 \leq y_2 \leq 1\}$, where $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ are fixed.

b) $S = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x} \geq \boldsymbol{0}^n, \ \boldsymbol{x}^{\mathrm{T}}\boldsymbol{1}^n = 1, \ \sum_{i=1}^n x_i a_i = b_1, \ \sum_{i=1}^n x_i a_i^2 = b_2\}$, where $a_i \in \mathbb{R}$ for $i = 1, \dots, n$, and $b_1, b_2 \in \mathbb{R}$ are fixed.

c) $S = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x} \geq \boldsymbol{0}^n,\ \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} \leq 1 \text{ for all } \boldsymbol{y} \text{ such that } \|\boldsymbol{y}\|_2 = 1\}$.

d) $S = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x} \geq \boldsymbol{0}^n,\ \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} \leq 1 \text{ for all } \boldsymbol{y} \text{ such that } \sum_{i=1}^n |y_i| = 1\}$.

e) $S = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x} - \boldsymbol{x}^0\|_2 \leq \|\boldsymbol{x} - \boldsymbol{x}^1\|_2\}$, where $\boldsymbol{x}^0, \boldsymbol{x}^1 \in \mathbb{R}^n$ are fixed.

f) $S = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x} - \boldsymbol{x}^0\|_2 \leq \|\boldsymbol{x} - \boldsymbol{x}^i\|_2,\ i = 1, \dots, k\}$, where $\boldsymbol{x}^0, \dots, \boldsymbol{x}^k \in \mathbb{R}^n$ are fixed.

■

Exercise 3.3 (extreme points) Consider the polyhedron P defined by

$$\begin{aligned} x_1 + x_2 &\leq 2, \\ x_2 &\leq 1, \\ x_3 &\leq 2, \\ x_2 + x_3 &\leq 2. \end{aligned}$$

a) Is $\boldsymbol{x}^1 = (1, 1, 0)^{\mathrm{T}}$ an extreme point to P?

b) Is $\boldsymbol{x}^2 = (1, 1, 1)^{\mathrm{T}}$ an extreme point to P?

■

Exercise 3.4 (existence of extreme points in LPs) Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ be such that rank $\boldsymbol{A} = m$, and let $\boldsymbol{b} \in \mathbb{R}^m$. Show that if the polyhedron

$$P = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}; \quad \boldsymbol{x} \geq \boldsymbol{0}^n \,\}$$

has a feasible solution, then it has an extreme point. ■

Exercise 3.5 (illustration of the Representation Theorem) Let

$$\begin{aligned} Q &= \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid -2x_1 + x_2 \leq 1; \quad x_1 - x_2 \leq 1; \quad -x_1 - x_2 \leq -1 \,\}, \\ C &= \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid -2x_1 + x_2 \leq 0; \quad x_1 - x_2 \leq 0; \quad -x_1 - x_2 \leq 0 \,\}, \end{aligned}$$

and P be the convex hull of the extreme points of Q. Show that the feasible point $\tilde{\boldsymbol{x}} = (1, 1)^{\mathrm{T}}$ can be written as

$$\tilde{\boldsymbol{x}} = \boldsymbol{p} + \boldsymbol{c},$$

where $\boldsymbol{p} \in P$ and $\boldsymbol{c} \in C$. ■

Exercise 3.6 (separation) Show that there is only one hyperplane in $\mathbb{R}^3$ which separates the disjoint closed convex sets A and B defined by the equations

$$A = \{ (0, x_2, 1)^{\mathrm{T}} \mid x_2 \in \mathbb{R} \}, \quad B = \{ \boldsymbol{x} \in \mathbb{R}^3 \mid \boldsymbol{x} \geq \mathbf{0}^3,\ x_1 x_2 \geq x_3^2 \}$$

and that this hyperplane meets both A and B. ■

Exercise 3.7 (separation) Show that each closed convex set A in $\mathbb{R}^n$ is the intersection of all the closed halfspaces in $\mathbb{R}^n$ containing A. ■

Exercise 3.8 (application of Farkas' Lemma) In a paper submitted for publication in an operations research journal, the author considered the set

$$P = \left\{ \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} \in \mathbb{R}^{n+m} \;\middle|\; \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} \geq \boldsymbol{c}; \quad \boldsymbol{x} \geq \mathbf{0}^n; \quad \boldsymbol{y} \geq \mathbf{0}^m \right\},$$

where $\boldsymbol{A}$ is an $m \times n$ matrix, $\boldsymbol{B}$ a positive semi-definite $m \times m$ matrix and $\boldsymbol{c} \in \mathbb{R}^m$. The author explicitly assumed that the set P is compact in $\mathbb{R}^{n+m}$. A reviewer of the paper pointed out that the only compact set of the above form is the empty set. Prove the reviewer's assertion. ■

Exercise 3.9 (convex sets) Let $S_1 := \{ \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 + x_2 \leq 1;\ x_1 \geq 0 \}$, $S_2 := \{ \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 - x_2 \geq 0;\ x_1 \leq 1 \}$, and $S := S_1 \cup S_2$. Prove that S_1 and S_2 are convex sets and that S is not convex. Hence, the union of convex sets is not necessarily a convex set. ■

Exercise 3.10 (convex functions) Determine if the function f defined by $f(\boldsymbol{x}) := 2x_1^2 - 3x_1x_2 + 5x_2^2 - 2x_1 + 6x_2$ is convex, concave, or neither, on $\mathbb{R}^2$. ■

Exercise 3.11 (convex functions) Let $a > 0$. Consider the following functions in one variable:

a) $f(x) := \ln x$, for $x > 0$;
b) $f(x) := -\ln x$, for $x > 0$;
c) $f(x) := -\ln(1 - e^{-ax})$, for $x > 0$;
d) $f(x) := \ln(1 + e^{ax})$;
e) $f(x) := e^{ax}$;
f) $f(x) := x \ln x$, for $x > 0$.

Which of these functions are convex (respectively, strictly convex)? ■

Exercise 3.12 (convex functions) Consider the following functions: a) $f(\boldsymbol{x}) := \ln(e^{x_1} + e^{x_2})$;

b) $f(\boldsymbol{x}) := \ln \sum_{j=1}^{n} e^{a_j x_j}$, where a_j, $j = 1, \ldots, n$, are constants;

c) $f(\boldsymbol{x}) := \sqrt{\displaystyle\sum_{j=1}^{n} x_j^2}$;

d) $f(\boldsymbol{x}) := x_1^2/x_2$, for $x_2 > 0$;

e) $f(\boldsymbol{x}) := -\sqrt{x_1 x_2}$, for $x_1, x_2 > 0$;

f) $f(\boldsymbol{x}) := -\left(\displaystyle\prod_{j=1}^{n} x_j\right)^{1/n}$, for $x_j > 0$, $j = 1, \ldots, n$.

Which of these functions are convex (respectively, strictly convex)? ■

Exercise 3.13 (convex functions) Consider the following function:

$$f(x, y) := 2x^2 - 2xy + \frac{1}{2}y^2 + 3x - y.$$

a) Express the function in matrix–vector form.

b) Is the Hessian singular?

c) Is f a convex function? ■

Exercise 3.14 (convex sets) Consider the following sets:

a) $\{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1^2 + x_2^2 \leq 1;\ x_1^2 + x_2^2 \geq 1/4 \,\}$;

b) $\{\, \boldsymbol{x} \in \mathbb{R}^n \mid x_j \geq 0,\ j = 1, \ldots, n \,\}$;

c) $\{\, \boldsymbol{x} \in \mathbb{R}^n \mid x_1^2 + x_2^2 + \cdots + x_n^2 = 1 \,\}$;

d) $\{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 + x_2^2 \leq 5;\ x_1^2 - x_2 \leq 10;\ x_1 \geq 0;\ x_2 \geq 0 \,\}$;

e) $\{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 - x_2^2 \geq 1;\ x_1^3 + x_2^2 \leq 10;\ 2x_1 + x_2 \leq 8;\ x_1 \geq 1;\ x_2 \geq 0 \,\}$.

Investigate whether each of them is convex or not. In the latter case, provide a counter-example. ■

Exercise 3.15 (convex sets) Is the set defined by

$$S := \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid 2e^{-x_1 + x_2^2} \leq 4,\ -x_1^2 + 3x_1 x_2 - 3x_2^2 \geq -1 \,\}$$

a convex set? ■

Exercise 3.16 (convex sets) Is the set defined by

$$S := \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 - x_2^2 \geq 1,\ x_1^3 + x_2^2 \leq 10,\ 2x_1 + x_2 \leq 8,\ x_1 \geq 1,\ x_2 \geq 0 \,\}$$

a convex set? ■

Exercise 3.17 (convex problem) Suppose that the function $g : \mathbb{R}^n \to \mathbb{R}$ is convex on $\mathbb{R}^n$ and that $\boldsymbol{d} \in \mathbb{R}^n$. Is the problem to

$$
\begin{aligned}
\text{maximize } & -\sum_{j=1}^{n} x_j^2, \\
\text{subject to } & -\frac{1}{\ln(-g(\boldsymbol{x}))} \geq 0, \\
& \boldsymbol{d}^{\mathrm{T}}\boldsymbol{x} = 2, \\
& g(\boldsymbol{x}) \leq -2, \\
& \boldsymbol{x} \geq \boldsymbol{0}^n
\end{aligned}
$$

a convex problem? ■

Exercise 3.18 (convex problem) Is the problem to

$$
\begin{aligned}
\text{maximize } & x_1 \ln x_1, \\
\text{subject to } & x_1^2 + x_2^2 \geq 0, \\
& \boldsymbol{x} \geq \boldsymbol{0}^2
\end{aligned}
$$

a convex problem? ■

Part III

Optimality Conditions

An introduction to optimality conditions

IV

4.1 Local and global optimality

Consider the problem to

$$\text{minimize } f(\boldsymbol{x}), \tag{4.1a}$$
$$\text{subject to } \boldsymbol{x} \in S, \tag{4.1b}$$

where $S \subseteq \mathbb{R}^n$ is a nonempty set and $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a given function.

Consider the function given in Figure 4.1.

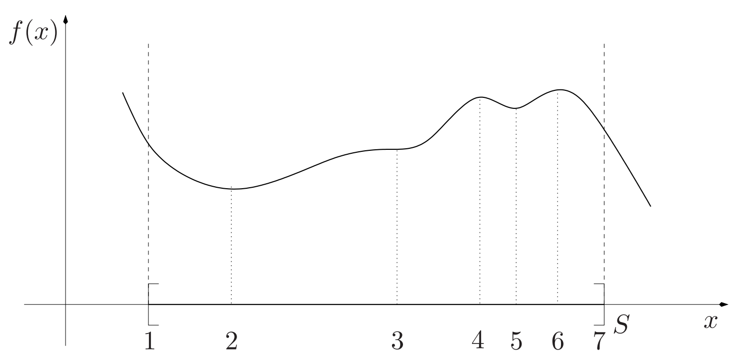


Figure 4.1: A one-dimensional function and its possible optimal points.

For a minimization problem over f in one variable over an interval S, the interesting points are:

(i) boundary points of S;

(ii) stationary points, that is, where $f'(x) = 0$;
(iii) discontinuities in f or f'.

In the case of the function in Figure 4.1 we have:

(i) 1, 7;
(ii) 2, 3, 4, 5, 6;
(iii) none.

Definition 4.1 (global minimum) *Consider the problem* (4.1). *Let* $\boldsymbol{x}^* \in S$. *Then, we say that* $\boldsymbol{x}^*$ *is a* global minimum *of* f *over* S *if it attains the lowest value of* f *over* S.

In other words, $\boldsymbol{x}^* \in S$ *is a global minimum of* f *over* S *if*

$$f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}), \qquad \boldsymbol{x} \in S, \tag{4.2}$$

holds. ■

Let $B_\varepsilon(\boldsymbol{x}^*) := \{ \boldsymbol{y} \in \mathbb{R}^n \mid \|\boldsymbol{y} - \boldsymbol{x}^*\| < \varepsilon \}$ be the Euclidean ball with radius ε centered at $\boldsymbol{x}^*$.

Definition 4.2 (local minimum) *Consider the problem* (4.1). *Let* $\boldsymbol{x}^* \in S$.

(a) *We say that* $\boldsymbol{x}^*$ *is a* local minimum *of* f *over* S *if there exists a small enough ball intersected with* S *around* $\boldsymbol{x}^*$ *such that it is a globally optimal solution in that smaller set.*

In other words, $\boldsymbol{x}^* \in S$ *is a local minimum of* f *over* S *if*

$$\exists \varepsilon > 0 \text{ such that } f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}), \qquad \boldsymbol{x} \in S \cap B_\varepsilon(\boldsymbol{x}^*). \tag{4.3}$$

(b) *We say that* $\boldsymbol{x}^* \in S$ *is a* strict local minimum *of* f *over* S *if, in* (4.3), *the inequality holds strictly for* $\boldsymbol{x} \neq \boldsymbol{x}^*$. ■

Note that a global minimum in particular is a local minimum. When is a local minimum a global one? This question is resolved in the case of convex problems, as the following fundamental theorem shows.

Theorem 4.3 (Fundamental Theorem of global optimality) *Consider the problem* (4.1), *where* S *is a convex set and* f *is convex on* S. *Then, every local minimum of* f *over* S *is also a global minimum.*

Proof. Suppose that $\boldsymbol{x}^*$ is a local minimum but not a global one, while $\bar{\boldsymbol{x}}$ is a global minimum. Then, $f(\bar{\boldsymbol{x}}) < f(\boldsymbol{x}^*)$. Let $\lambda \in (0, 1)$. By the convexity of S and f, $\lambda\bar{\boldsymbol{x}} + (1 - \lambda)\boldsymbol{x}^* \in S$, and $f(\lambda\bar{\boldsymbol{x}} + (1 - \lambda)\boldsymbol{x}^*) \leq \lambda f(\bar{\boldsymbol{x}}) + (1 - \lambda)f(\boldsymbol{x}^*) < f(\boldsymbol{x}^*)$. Choosing $\lambda > 0$ small enough then leads to a contradiction to the local optimality of $\boldsymbol{x}^*$. ■

There is an intuitive image that can be seen from the proof design: If $\boldsymbol{x}^*$ is a local minimum, then f cannot "go down-hill" from $\boldsymbol{x}^*$ in any direction, but if $\bar{\boldsymbol{x}}$ has a lower value, then f has to go down-hill sooner or later. This cannot be the shape of any convex function.

The example in Figure 4.2 shows a case where, without convexity, a vector $\boldsymbol{x}^*$ may be a local minimum of a function $f \in C^1$ with respect to every *line segment* that passes through $\boldsymbol{x}^*$, and yet it is not even a local minimum of f over $\mathbb{R}^n$.

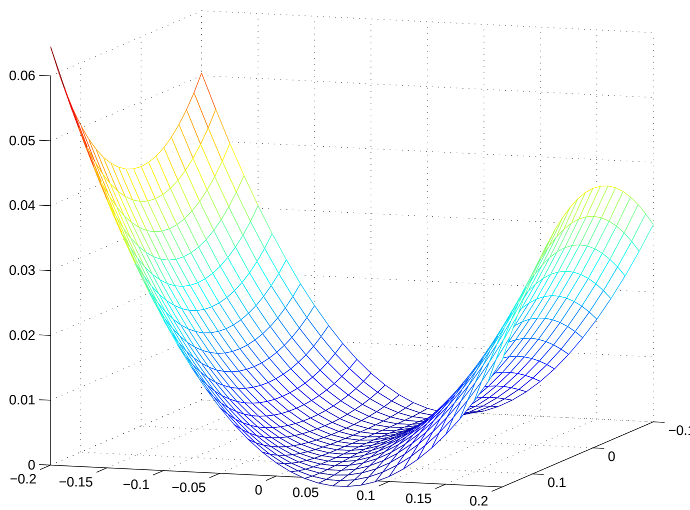


Figure 4.2: A three-dimensional graph of the function $f(x, y) = (y - x^2)(y - 4x^2)$. The origin is a local minimum with respect to every line that passes through it, but it is not a local minimum of f.

In fact, this situation may also occur in the convex case when $f \notin C^1$.

In the simple one-dimensional example in Figure 4.1, finding and checking the different points of the form (i)–(iii) was easy; there are of course examples even in $\mathbb{R}$ which makes this "algorithm" impossible to use, and when considering the multi-dimensional case (that is, $n > 1$) this is a completely absurd "method" for solving a problem.

In the following we will develop necessary and sufficient conditions for $\boldsymbol{x}^*$ to be a local or a global optimal solution to the problem (4.1) for any dimension $n \geq 1$, and which are useful and possible to check. Before we do that, however, we will establish exactly when a globally optimal solution to the problem (4.1) exists.

4.2 Existence of optimal solutions

We first pave the way for a classic result from calculus: Weierstrass' Theorem.

Definition 4.4 (weakly coercive, coercive functions) *Let $S \subseteq \mathbb{R}^n$ be a nonempty and closed set, and $f : S \to \mathbb{R}$ be a given function.*

(a) *We say that f is* weakly coercive *with respect to the set S if S is bounded or the sequence $\{f(\boldsymbol{x}_k)\}$ tends to infinity whenever the sequence $\{\boldsymbol{x}_k\} \subset S$ tends to infinity in norm.*

In other words, f is weakly coercive if either S is bounded or

$$\lim_{\substack{\|\boldsymbol{x}_k\| \to \infty \\ \boldsymbol{x}_k \in S}} f(\boldsymbol{x}_k) = \infty$$

holds.

(b) *We say that f is* coercive *with respect to the set S if S is bounded or the sequence $\{f(\boldsymbol{x}_k)/\|\boldsymbol{x}_k\|\}$ tends to infinity whenever the sequence $\{\boldsymbol{x}_k\} \subset S$ tends to infinity in norm.*

In other words, f is coercive if either S is bounded or

$$\lim_{\substack{\|\boldsymbol{x}_k\| \to \infty \\ \boldsymbol{x}_k \in S}} f(\boldsymbol{x}_k)/\|\boldsymbol{x}_k\| = \infty$$

holds. ■

The weak coercivity of $f : S \to \mathbb{R}$ is (for nonempty sets S) equivalent to the property that f has bounded level sets restricted to S (cf. Definition 3.43). (Why?)

A coercive function clearly grows faster than any linear function. In fact, for convex functions f, f being coercive is equivalent to $\boldsymbol{x} \mapsto f(\boldsymbol{x}) - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ being weakly coercive for every vector $\boldsymbol{a} \in \mathbb{R}^n$. This property is a very useful one for certain analyses in the context of Lagrangian duality.[1]

We next introduce two extended notions of continuity.

[1] For example, in Section 6.4.2 we suppose that the ground set X is compact in order for the Lagrangian dual function q to be finite. It is possible to replace the boundedness condition on X with a coercivity condition on f.

Definition 4.5 (semi-continuity) *Consider a function $f : S \to \mathbb{R}$, where $S \subseteq \mathbb{R}^n$ is nonempty.*

(a) *The function f is said to be* lower semi-continuous *at $\bar{\boldsymbol{x}} \in S$ if the value $f(\bar{\boldsymbol{x}})$ is less than or equal to every limit of f as $\{\boldsymbol{x}_k\} \to \bar{\boldsymbol{x}}$.*

In other words, f is lower semi-continuous *at $\bar{\boldsymbol{x}} \in S$ if*

$$\{\boldsymbol{x}_k\} \to \bar{\boldsymbol{x}} \qquad \Longrightarrow \qquad f(\bar{\boldsymbol{x}}) \leq \liminf_{k\to\infty} f(\boldsymbol{x}_k).$$

(b) *The function f is said to be* upper semi-continuous *at $\bar{\boldsymbol{x}} \in S$ if the value $f(\bar{\boldsymbol{x}})$ is greater than or equal to every limit of f as $\{\boldsymbol{x}_k\} \to \bar{\boldsymbol{x}}$.*

In other words, f is upper semi-continuous *at $\bar{\boldsymbol{x}} \in S$ if*

$$\{\boldsymbol{x}_k\} \to \bar{\boldsymbol{x}} \qquad \Longrightarrow \qquad f(\bar{\boldsymbol{x}}) \geq \limsup_{k\to\infty} f(\boldsymbol{x}_k).$$

We say that f is lower semi-continuous on S *(respectively,* upper semi-continuous on S*) if it is lower semi-continuous (respectively, upper semi-continuous) at every $\bar{\boldsymbol{x}} \in S$.* ■

Lower semi-continuous functions in one variable have the appearance shown in Figure 4.3.

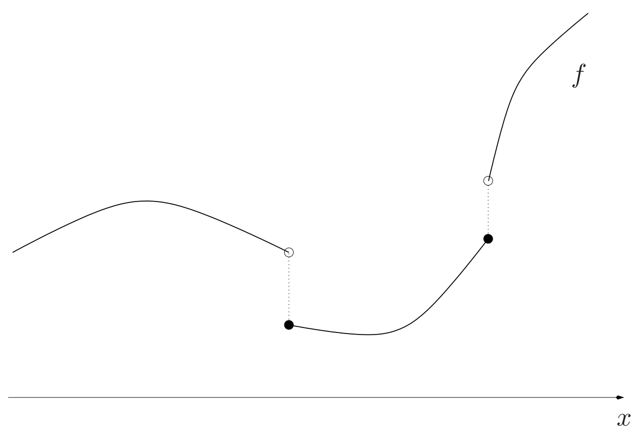


Figure 4.3: A lower semi-continuous function in one variable.

Establish the following important relations:

(a) The inequalities displayed in Definition 4.5 can be replaced by equalities, since the respective opposite inequlities are trivially satisfied.

(b) The function f mentioned in Definition 4.5 is *continuous* at $\bar{\boldsymbol{x}} \in S$ if and only if it is *both* lower and upper semi-continuous at $\bar{\boldsymbol{x}}$.

(c) Lower semi-continuity of f is equivalent to the closedness of all its level sets $\mathrm{lev}_f(b)$, $b \in \mathbb{R}$ (cf. Definition 3.43), as well as the closedness of its epigraph (cf. Definition 3.38).

Next follows the famous existence theorem credited to Karl Weierstrass (see, however, Section 4.7).

Theorem 4.6 (Weierstrass' Theorem) *Let $S \subseteq \mathbb{R}^n$ be a nonempty and closed set, and $f : S \to \mathbb{R}$ be a lower semi-continuous function on S. If f is weakly coercive with respect to S, then there exists a nonempty, closed and bounded (thus compact) set of globally optimal solutions to the problem* (4.1).

Proof. We first assume that S is bounded, and proceed by choosing a sequence $\{\boldsymbol{x}_k\}$ in S such that

$$\lim_{k\to\infty} f(\boldsymbol{x}_k) = \inf_{\boldsymbol{x}\in S} f(\boldsymbol{x}).$$

(The infimum of f over S is the lowest limit of all sequences of the form $\{f(\boldsymbol{x}_k)\}$ with $\{\boldsymbol{x}_k\} \subset S$, so such a sequence of $\boldsymbol{x}_k$ is what we here are choosing.)

Due to the boundedness of S, the sequence $\{\boldsymbol{x}_k\}$ must have limit points, all of which lie in S because of the closedness of S. Let $\bar{\boldsymbol{x}}$ be an arbitrary limit point of $\{\boldsymbol{x}_k\}$, corresponding to the subsequence $\mathcal{K} \subseteq \mathbb{Z}_+$. Then, by the lower semi-continuity of f,

$$f(\bar{\boldsymbol{x}}) \leq \lim_{k\in\mathcal{K}} f(\boldsymbol{x}_k) = \inf_{\boldsymbol{x}\in S} f(\boldsymbol{x}).$$

Since $\bar{\boldsymbol{x}}$ attains the infimum of f over S, $\bar{\boldsymbol{x}}$ is a global minimum of f over S. This limit point of $\{\boldsymbol{x}_k\}$ was arbitrarily chosen; any other choice (provided more than one exists) has the same (optimal) objective value.

Suppose next that f is weakly coercive, and consider the same sequence $\{\boldsymbol{x}_k\}$ in S. Then, by the weak coercivity assumption, either $\{\boldsymbol{x}_k\}$ is bounded or the sequence $\{f(\boldsymbol{x}_k)\}$ tends to infinity. The non-emptiness of S implies that $\inf_{\boldsymbol{x}\in S} f(\boldsymbol{x}) < \infty$ holds, and hence we conclude that $\{\boldsymbol{x}_k\}$ is bounded. We can then utilize the same arguments as in the previous paragraph and conclude that also in this case there exists a globally optimal solution. We are done. ■

Before moving on we take a closer look at the proof of this result, because it is important in order to understand the importance of some of the assumptions that we make about the optimization models that we pose. We notice that the closedness of S is really crucial; if it is not then a sequence generated in S may converge to a point outside of S, which means that we would converge to an infeasible and of course also non-optimal solution. This is the reason why the generic optimization model (1.1) stated in Chapter 1 does not contain any constraints of the form

$$g_i(\boldsymbol{x}) < 0, \qquad i \in \mathcal{SI},$$

where $\mathcal{SI}$ denotes *strict inequality*. The reason is that such constraints in general may describe non-closed sets.

Weierstrass' Theorem 4.6 is next improved for special classes of the problem (4.1). The main purpose of presenting these results is to show the role of convexity and to illustrate the special properties of convex quadratic programs and linear programs. The proofs are rather complex and are therefore left out.

Theorem 4.7 (existence of solutions, convex polynomials) *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a convex polynomial function. Suppose further that the set S can be described by inequality constraints of the form $g_i(\boldsymbol{x}) \leq 0$, $i = 1, \dots, m$, where each function g_i is convex and polynomial. The problem* (4.1) *then is convex. Moreover, it has a nonempty (as well as closed and convex) set of globally optimal solutions if and only if f is lower bounded on S.* ∎

In the following result, we let S be a nonempty polyhedron, and suppose that it is possible to describe it as the following finite (cf. Definition 3.15) set of linear constraints:

$$S = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}; \quad \boldsymbol{E}\boldsymbol{x} = \boldsymbol{d} \,\}, \tag{4.4}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{E} \in \mathbb{R}^{\ell \times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, and $\boldsymbol{d} \in \mathbb{R}^\ell$. The *recession cone* to S then is the following set, defining the set of directions that are feasible at every point in S:[2]

$$\operatorname{rec}_S := \{\, \boldsymbol{p} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{p} \leq \boldsymbol{0}^m; \quad \boldsymbol{E}\boldsymbol{p} = \boldsymbol{0}^\ell \,\}. \tag{4.5}$$

(For the definition of the set of feasible directions at *a given vector* $\boldsymbol{x}$, see Definition 4.19.)

We also suppose that

$$f(\boldsymbol{x}) := \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{q}^{\mathrm{T}}\boldsymbol{x}, \qquad \boldsymbol{x} \in \mathbb{R}^n, \tag{4.6}$$

where $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ is a symmetric and positive semi-definite matrix and $\boldsymbol{q} \in \mathbb{R}^n$. We define the recession cone to any convex function $f : \mathbb{R}^n \to \mathbb{R}$ as follows: the *recession cone* to f is the recession cone to the level

[2] Recall the cone C in the Representation Theorem 3.22.

set of f (cf. Definition 3.43), defined for any value of b for which the corresponding level set of f is nonempty. (Check that this cone actually is independent of the value of b under this only requirement. Also confirm that if the level set $\operatorname{lev}_f(b)$ is (nonempty and) bounded for some $b \in \mathbb{R}$ then it is bounded for every $b \in \mathbb{R}$, thanks to the convexity of f.) In the special case of the convex quadratic function given in (4.6),

$$\operatorname{rec}_f = \{\, \boldsymbol{p} \in \mathbb{R}^n \mid \boldsymbol{Q}\boldsymbol{p} = \boldsymbol{0}^n; \quad \boldsymbol{q}^{\mathrm{T}}\boldsymbol{p} \leq 0 \,\}.$$

This is the set of directions that nowhere are ascent directions to f.

Corollary 4.8 (the Frank–Wolfe Theorem) *Suppose that S is the polyhedron described by* (4.4) *and f is the convex quadratic function given by* (4.6)*, so that the problem* (4.1) *is a convex quadratic programming problem. Then, the following three statements are equivalent.*

(a) *The problem* (4.1) *has a nonempty (as well as a closed and convex) set of globally optimal solutions.*

(b) *f is lower bounded on S.*

(c) *For every vector $\boldsymbol{p}$ in the intersection of the recession cone* rec_S *to S and the null space $N(\boldsymbol{Q})$ of the matrix $\boldsymbol{Q}$, it holds that $\boldsymbol{q}^{\mathrm{T}}\boldsymbol{p} \geq 0$. In other words,*

$$\boldsymbol{p} \in \operatorname{rec}_S \cap N(\boldsymbol{Q}) \qquad \Longrightarrow \qquad \boldsymbol{q}^{\mathrm{T}}\boldsymbol{p} \geq 0$$

holds. ■

The statement in (c) shows that the conditions for the existence of an optimal solution in the case of convex quadratic programs are milder than in the general convex case. In the latter case, we can state a slight improvement over the Weierstrass Theorem 4.6 that if, in the problem (4.1), f is convex on S where the latter is nonempty, closed and convex, then the problem has a nonempty, convex and compact set of globally optimal solutions if and only if $\operatorname{rec}_S \cap \operatorname{rec}_f = \{\boldsymbol{0}^n\}$. The improvements in the above results for polyhedral, in particular quadratic, programs stems from the fact that convex polynomial functions cannot be lower bounded and yet not have a global minimum.

[*Note:* Consider the special case of the problem (4.1) where $f(x) := 1/x$ and $S := [1, +\infty)$. It is clear that f is bounded from below on S, in fact by the value zero which is the infimum of f over S, but it never attains the value zero on S, and therefore this problem has no optimal solution. Of course, f is not a polynomial function.]

Corollary 4.9 (a fundamental theorem in linear programming) *Suppose, in the Frank–Wolfe Theorem, that f is linear, that is, that $\boldsymbol{Q} = \boldsymbol{0}^{n \times n}$.*

Then, the problem (4.1) *is identical to a linear programming (LP) problem. Then, the following three statements are equivalent.*

(a) *The problem* (4.1) *has a nonempty (as well as a closed and convex polyhedral) set of globally optimal solutions.*

(b) f *is lower bounded on* S.

(c) *For every vector* $\boldsymbol{p}$ *in the recession cone* rec_S *to* S, *it holds that* $\boldsymbol{q}^{\mathrm{T}}\boldsymbol{p} \geq 0$. *In other words,*

$$\boldsymbol{p} \in \mathrm{rec}_S \qquad \Longrightarrow \qquad \boldsymbol{q}^{\mathrm{T}}\boldsymbol{p} \geq 0$$

holds. ∎

Corollary 4.9 will in fact be established later on in Theorem 8.10, by the use of polyhedral convexity, when we specialize our treatment of nonlinear optimization to that of linear optimization. Since we have already established the Representation Theorem 3.22, proving Corollary 4.9 for the case of LP will be easy: since the objective function is linear, every feasible direction $\boldsymbol{p} \in \mathrm{rec}_S$ with $\boldsymbol{q}^{\mathrm{T}}\boldsymbol{p} < 0$ leads to an unbounded solution from any vector $\boldsymbol{x} \in S$.

Under strict convexity, we can finally establish the following result.

Proposition 4.10 (unique solution under strict convexity) *Suppose that in the problem* (4.1) f *is strictly convex on* S *and the set* S *is convex. Then, there can be at most one globally optimal solution.*

Proof. Suppose, by means of contradiction, that $\boldsymbol{x}^*$ and $\boldsymbol{x}^{**}$ are two different globally optimal solutions. Then, for every $\lambda \in (0,1)$, we have that

$$f(\lambda\boldsymbol{x}^* + (1-\lambda)\boldsymbol{x}^{**}) < \lambda f(\boldsymbol{x}^*) + (1-\lambda)f(\boldsymbol{x}^{**}) = f(\boldsymbol{x}^*)\,[= f(\boldsymbol{x}^{**})].$$

Since $\lambda\boldsymbol{x}^* + (1-\lambda)\boldsymbol{x}^{**} \in S$, we have found an entire interval of points which are strictly better than $\boldsymbol{x}^*$ or $\boldsymbol{x}^{**}$. This is impossible, whence we are done. ∎

We finally characterize a class of optimization problems over polytopes whose optimal solution set, if nonempty, includes an extreme point.

Consider the optimization problem to

$$\begin{aligned} &\text{maximize} && f(\boldsymbol{x}), \\ &\text{subject to} && \boldsymbol{x} \in P, \end{aligned} \tag{4.7}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and $P \subset \mathbb{R}^n$ is a nonempty, bounded polyhedron (that is, a polytope). Then, from the Representation Theorem 3.22 it follows below that an optimal solution can be found among the extreme points of P. Theorem 8.10 establishes a corresponding result for linear programs that does not rely on Weierstrass' Theorem.

Theorem 4.11 (optimal extreme point) *An optimal solution to* (4.7) *can be found among the extreme points of* P.

Proof. The function f is continuous (since it is convex, cf. Theorem 4.26 below); further, P is a nonempty and compact set. Hence, there exists an optimal solution $\tilde{\boldsymbol{x}}$ to (4.7) by Weierstrass' Theorem 4.6. The Representation Theorem 3.22 implies that $\tilde{\boldsymbol{x}} = \lambda_1 \boldsymbol{v}^1 + \cdots + \lambda_k \boldsymbol{v}^k$ for some extreme points $\boldsymbol{v}^1, \dots, \boldsymbol{v}^k$ of P and $\lambda_1, \dots, \lambda_k \geq 0$ such that $\sum_{i=1}^k \lambda_i = 1$. But then (from the convexity of f)

$$\begin{aligned} f(\tilde{\boldsymbol{x}}) &= f(\lambda_1 \boldsymbol{v}^1 + \cdots + \lambda_k \boldsymbol{v}^k) \leq \lambda_1 f(\boldsymbol{v}^1) + \cdots + \lambda_k f(\boldsymbol{v}^k) \\ &\leq \lambda_1 f(\tilde{\boldsymbol{x}}) + \cdots + \lambda_k f(\tilde{\boldsymbol{x}}) = f(\tilde{\boldsymbol{x}}), \end{aligned}$$

which gives that $f(\tilde{\boldsymbol{x}}) = f(\boldsymbol{v}^i)$ for some $i = 1, \dots, k$. ∎

Remark 4.12 Every linear function is convex, so Theorem 4.11 implies, in particular, that every linear program over a nonempty and bounded polyhedron has an optimal extreme point. ∎

4.3 Optimality in unconstrained optimization

In Theorem 4.3 we have established that locally optimal solutions also are global in the convex case. What are the necessary and sufficient conditions for a vector $\boldsymbol{x}^*$ to be a local optimum? This is an important question, because the algorithms that we will investigate for solving important classes of optimization problems are always devised based on those conditions that we would like to fulfill. This is a statement that seems to be true universally: *efficient, locally or globally convergent iterative algorithms for an optimization problem are directly based on its necessary and/or sufficient local optimality conditions.*

We begin by establishing these conditions for the case of unconstrained optimization, where the objective function is in C^1. Every proof is based on the Taylor expansion up to order one or two.

Our problem here is the following:

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\operatorname{minimize}} \ f(\boldsymbol{x}), \tag{4.8}$$

where f is in C^1 on $\mathbb{R}^n$ [for short we say: in C^1].

Theorem 4.13 (necessary optimality conditions, C^1 case) *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is in C^1 on $\mathbb{R}^n$. Then,*

$$\boldsymbol{x}^* \text{ is a local minimum of } f \text{ on } \mathbb{R}^n \quad \Longrightarrow \quad \nabla f(\boldsymbol{x}^*) = \mathbf{0}^n.$$

Note that

$$\nabla f(\boldsymbol{x}) = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_j}\right)_{j=1}^{n},$$

so the requirement thus is that $\frac{\partial f(\boldsymbol{x}^*)}{\partial x_j} = 0$, $j = 1, \dots, n$.

Just as for the case $n = 1$, we refer to this condition as $\boldsymbol{x}^*$ being a *stationary point* of f.

[*Note:* For $n = 1$, Theorem 4.13 reduces to: $x^* \in \mathbb{R}$ is a local minimum $\Longrightarrow f'(x^*) = 0$.]

Proof. (By contradiction.) Suppose that $\boldsymbol{x}^*$ is a local minimum, but that $\nabla f(\boldsymbol{x}^*) \neq \mathbf{0}^n$. Let $\boldsymbol{p} := -\nabla f(\boldsymbol{x}^*)$, and study the Taylor expansion around $\boldsymbol{x} = \boldsymbol{x}^*$ in the direction of $\boldsymbol{p}$:

$$f(\boldsymbol{x}^* + \alpha \boldsymbol{p}) = f(\boldsymbol{x}^*) + \alpha \nabla f(\boldsymbol{x}^*)^{\mathrm{T}} \boldsymbol{p} + o(\alpha),$$

where $o : \mathbb{R} \to \mathbb{R}$ is such that $o(s)/s \to 0$ when $s \downarrow 0$. We get that

$$\begin{aligned} f(\boldsymbol{x}^* + \alpha \boldsymbol{p}) &= f(\boldsymbol{x}^*) - \alpha \|\nabla f(\boldsymbol{x}^*)\|^2 + o(\alpha) \\ &< f(\boldsymbol{x}^*) \text{ for all small enough } \alpha > 0, \end{aligned}$$

since $\|\nabla f(\boldsymbol{x}^*)\| \neq 0$. This completes the proof. ∎

The opposite direction is false. Take, for example, $f(x) = x^3$. Then, $\bar{x} = 0$ is stationary, but it is neither a local minimum or a local maximum.

The proof is instrumental in that it provides a sufficient condition for a vector $\boldsymbol{p}$ to define a *descent direction*, that is, a direction such that a small step along it yields a lower objective value. We first define this notion properly.

Definition 4.14 (descent direction) *Let the function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ be given. Let $\boldsymbol{x} \in \mathbb{R}^n$ be a vector such that $f(\boldsymbol{x})$ is finite. Let $\boldsymbol{p} \in \mathbb{R}^n$. We say that the vector $\boldsymbol{p} \in \mathbb{R}^n$ is a* descent direction *with respect to f at $\boldsymbol{x}$ if*

$$\exists \delta > 0 \text{ such that } f(\boldsymbol{x} + \alpha \boldsymbol{p}) < f(\boldsymbol{x}) \text{ for every } \alpha \in (0, \delta]$$

holds. ∎

Proposition 4.15 (sufficient condition for descent) *Suppose that* $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *is in* C^1 *around a point* $\boldsymbol{x}$ *for which* $f(\boldsymbol{x}) < +\infty$, *and that* $\boldsymbol{p} \in \mathbb{R}^n$. *If* $\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{p} < 0$ *then the vector* $\boldsymbol{p}$ *defines a direction of descent with respect to* f *at* $\boldsymbol{x}$.

Proof. Since f is in C^1 around $\boldsymbol{x}$, we can construct a Taylor expansion of f, as above:

$$f(\boldsymbol{x} + \alpha\boldsymbol{p}) = f(\boldsymbol{x}) + \alpha\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{p} + o(\alpha).$$

Since $\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{p} < 0$, we obtain that $f(\boldsymbol{x} + \alpha\boldsymbol{p}) < f(\boldsymbol{x})$ for all sufficiently small values of $\alpha > 0$. ∎

Notice that at a point $\boldsymbol{x} \in \mathbb{R}^n$ there may be other descent directions $\boldsymbol{p} \in \mathbb{R}^n$ beside those satisfying that $\nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{p} < 0$; in Example 11.2(b) we show how directions of *negative curvature* stemming from eigenvectors corresponding to negative eigenvalues of the Hessian matrix $\nabla^2 f(\boldsymbol{x})$ can be utilized.

If f has stronger differentiability properties, then we can say even more what a local optimum must be like.

Theorem 4.16 (necessary optimality conditions, C^2 case) *Suppose that* $f : \mathbb{R}^n \to \mathbb{R}$ *is in* C^2 *on* $\mathbb{R}^n$. *Then,*

$$\boldsymbol{x}^* \text{ is a local minimum of } f \implies \begin{cases} \nabla f(\boldsymbol{x}^*) = \boldsymbol{0}^n \\ \nabla^2 f(\boldsymbol{x}^*) \text{ is positive semi-definite.} \end{cases}$$

[*Note:* For $n = 1$, Theorem 4.16 reduces to: $x^* \in \mathbb{R}$ is a local minimum $\implies f'(x^*) = 0$ and $f''(x^*) \geq 0$.]

Proof. Consider the Taylor expansion of f up to order two around $\boldsymbol{x}^*$ and in the direction of a vector $\boldsymbol{p} \in \mathbb{R}^n$:

$$f(\boldsymbol{x}^* + \alpha\boldsymbol{p}) = f(\boldsymbol{x}^*) + \alpha\nabla f(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{p} + \frac{\alpha^2}{2}\boldsymbol{p}^{\mathrm{T}}\nabla^2 f(\boldsymbol{x}^*)\boldsymbol{p} + o(\alpha^2).$$

Suppose that $\boldsymbol{x}^*$ satisfies $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}^n$, but that there is a vector $\boldsymbol{p} \neq \boldsymbol{0}^n$ with $\boldsymbol{p}^{\mathrm{T}}\nabla^2 f(\boldsymbol{x}^*)\boldsymbol{p} < 0$, that is, $\nabla^2 f(\boldsymbol{x}^*)$ is not positive semidefinite. Then the above yields that $f(\boldsymbol{x}^* + \alpha\boldsymbol{p}) < f(\boldsymbol{x}^*)$ for all small enough $\alpha > 0$, whence $\boldsymbol{x}^*$ cannot be a local minimum. ∎

Also in this case, the opposite direction is false; the same counter-example as that after Theorem 4.13 applies.

In Example 11.2(b) we provide an example descent direction that has the form provided in the above proof; it is based on $\boldsymbol{p}$ being an eigenvector corresponding to a negative eigenvalue of $\nabla^2 f(\boldsymbol{x}^*)$.

The next result shows that under some circumstances, we can establish local optimality for a stationary point.

Theorem 4.17 (sufficient optimality conditions, C^2 case) *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is in C^2 on $\mathbb{R}^n$. Then,*

$$\left.\begin{array}{r}\nabla f(\boldsymbol{x}^*) = \mathbf{0}^n \\ \nabla^2 f(\boldsymbol{x}^*) \text{ is positive definite}\end{array}\right\} \Longrightarrow \boldsymbol{x}^* \text{ is a strict local minimum of } f.$$

[*Note:* For $n = 1$, Theorem 4.17 reduces to: $f'(x^*) = 0$ and $f''(x^*) > 0 \Longrightarrow x^* \in \mathbb{R}$ is a strict local minimum.]

Proof. Suppose that $\nabla f(\boldsymbol{x}^*) = \mathbf{0}^n$ and $\nabla^2 f(\boldsymbol{x}^*)$ is positive definite. Take an arbitrary vector $\boldsymbol{p} \in \mathbb{R}^n, \boldsymbol{p} \neq \mathbf{0}^n$. Then,

$$\begin{aligned} f(\boldsymbol{x}^* + \alpha \boldsymbol{p}) &= f(\boldsymbol{x}^*) + \alpha \underbrace{\nabla f(\boldsymbol{x}^*)^{\mathrm{T}} \boldsymbol{p}}_{=0} + \frac{\alpha^2}{2} \underbrace{\boldsymbol{p}^{\mathrm{T}} \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{p}}_{>0} + o(\alpha^2) \\ &> f(\boldsymbol{x}^*), \quad \text{for all small enough } \alpha > 0. \end{aligned}$$

As $\boldsymbol{p}$ was arbitrary, it implies that $\boldsymbol{x}^*$ is a strict local minimum. ∎

We naturally face the following question: When is a stationary point a global minimum? The answer is given next. (It is instrumental to investigate the connection between this result and the Fundamental Theorem 4.3.)

Theorem 4.18 (necessary and sufficient global optimality conditions) *Let $f \in C^1$, and let f be convex. Then,*

$$\boldsymbol{x}^* \text{ is a global minimum of } f \quad \Longleftrightarrow \quad \nabla f(\boldsymbol{x}^*) = \mathbf{0}^n.$$

Proof. [$\Longrightarrow$] This has already be shown in Theorem 4.13, since a global minimum is a local minimum.

[$\Longleftarrow$] The convexity of f yields that for every $\boldsymbol{y} \in \mathbb{R}^n$,

$$\begin{aligned} f(\boldsymbol{y}) &\geq f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^*)^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{x}^*) \\ &= f(\boldsymbol{x}^*), \end{aligned}$$

where the equality stems from the property that $\nabla f(\boldsymbol{x}^*) = \mathbf{0}^n$, by assumption. ∎

4.4 Optimality for optimization over convex sets

We consider a quite general optimization problem of the form:

$$\text{minimize } f(\boldsymbol{x}), \tag{4.9a}$$

$$\text{subject to } \boldsymbol{x} \in S, \tag{4.9b}$$

where $S \subseteq \mathbb{R}^n$ is nonempty, closed and convex, and $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is in C^1 on S.

A noticeable difference to unconstrained optimization is the fact that whether a vector $\boldsymbol{p} \in \mathbb{R}^n$ can be used as a direction of movement from a point $\boldsymbol{x} \in S$ depends on the constraints defining S; if $\boldsymbol{x}$ is an interior point of S then every $\boldsymbol{p} \in \mathbb{R}^n$ is a *feasible direction*, otherwise only certain directions will be feasible. That is, it all depends on whether there are any *active constraints* of S at $\boldsymbol{x}$ or not. We will define these terms in detail next, and then develop necessary and sufficient optimality conditions based on them. These conditions are natural extensions of those for the case of unconstrained optimization and reduces to them when $S = \mathbb{R}^n$. Further, we will develop a way of measuring the distance to an optimal solution in terms of the value of the objective function f which is valid for convex problems. As a result of this development, we will also be able to finally establish the Separation Theorem 3.24, whose proof has been postponed until now. (See Section 4.6.2 for the proof.)

Definition 4.19 (feasible direction) *Suppose that $\boldsymbol{x} \in S$, where $S \subseteq \mathbb{R}^n$, and that $\boldsymbol{p} \in \mathbb{R}^n$. Then, the vector $\boldsymbol{p}$ defines a* feasible direction *at $\boldsymbol{x}$ if a small enough step in the direction of $\boldsymbol{p}$ does not lead outside of the set S.*

In other words, the vector $\boldsymbol{p}$ defines a feasible direction *at $\boldsymbol{x} \in S$ if*

$$\exists \delta > 0 \textit{ such that } \boldsymbol{x} + \alpha \boldsymbol{p} \in S \textit{ for all } \alpha \in [0, \delta]$$

holds. ■

Recall that in the discussion following Theorem 4.7 we defined the set of feasible directions of a polyhedral set, that is, the set of directions that are feasible at every feasible point. For a general set S it would hence be the set

$$\{\, \boldsymbol{p} \in \mathbb{R}^m \mid \forall \boldsymbol{x} \in S \; \exists \delta > 0 \text{ such that } \boldsymbol{x} + \alpha \boldsymbol{p} \in S \text{ for all } \alpha \in [0, \delta] \,\}.$$

For nonempty, closed and convex sets S, this set is nonempty if and only if the set S also is unbounded. (Why?)

Definition 4.20 (active constraints) *Suppose that the set $S \subset \mathbb{R}^n$ is defined by a finite collection of equality and inequality constraints:*

$$S = \{ \boldsymbol{x} \in \mathbb{R}^n \mid g_i(\boldsymbol{x}) = 0, \quad i \in \mathcal{E}; \quad g_i(\boldsymbol{x}) \leq 0, \quad i \in \mathcal{I} \},$$

where $g_i : \mathbb{R}^n \to \mathbb{R}$ ($i \in \mathcal{E} \cup \mathcal{I}$) are given functions. Suppose that $\boldsymbol{x} \in S$. The set of active constraints *at $\boldsymbol{x}$ is the union of all the equality constraints and the set of inequality constraints that are satisfied with equality at $\boldsymbol{x}$, that is, the set $\mathcal{E} \cup \mathcal{I}(\boldsymbol{x})$, where $\mathcal{I}(\boldsymbol{x}) := \{ i \in \mathcal{I} \mid g_i(\boldsymbol{x}) = 0 \}$.* ■

Example 4.21 (feasible directions for linear constraints) Suppose, as a special case, that the constraints are all linear, that is, that for every $i \in \mathcal{E}$, $g_i(\boldsymbol{x}) := \boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{x} - d_i$ ($\boldsymbol{e}_i \in \mathbb{R}^n$; $d_i \in \mathbb{R}$), and for every $i \in \mathcal{I}$, $g_i(\boldsymbol{x}) := \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} - b_i$ ($\boldsymbol{a}_i \in \mathbb{R}^n$; $b_i \in \mathbb{R}$). In other words, in matrix notation, $S = \{ \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{E}\boldsymbol{x} = \boldsymbol{d}; \quad \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \}$.

Suppose further that $\boldsymbol{x} \in S$. Then, the set of feasible directions at $\boldsymbol{x}$ is the set

$$\{ \boldsymbol{p} \in \mathbb{R}^n \mid \boldsymbol{E}\boldsymbol{p} = \boldsymbol{0}^{\ell}; \quad \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{p} \leq 0, \ i \in \mathcal{I}(\boldsymbol{x}) \}.$$

Just as S, this is a polyhedron. Moreover, it is a polyhedral cone. ■

Clearly, the set of feasible directions of the polyhedral set S (or, the recession cone of S) is

$$\mathrm{rec}_S := \{ \boldsymbol{p} \in \mathbb{R}^n \mid \boldsymbol{E}\boldsymbol{p} = \boldsymbol{0}^{\ell}; \quad \boldsymbol{A}\boldsymbol{p} \leq \boldsymbol{0}^m \},$$

as stated in (4.5). Note moreover that the above set rec_S represents the cone C in the Representation Theorem 3.22.[3]

We can now more or less repeat the arguments for the unconstrained case in order to establish a necessary optimality condition for constrained optimization problems. This condition will immediately be refined for convex feasible sets, then later on in Chapter 5 be given a general statement for the case of explicit constraints in the form of the famous Karush–Kuhn–Tucker conditions in nonlinear programming.

Proposition 4.22 (necessary optimality conditions, C^1 case) *Suppose that $S \subseteq \mathbb{R}^n$ and that $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is in C^1 around a point $\boldsymbol{x} \in S$ for which $f(\boldsymbol{x}) < +\infty$.*

(a) *If $\boldsymbol{x}^* \in S$ is a local minimum of f on S then $\nabla f(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{p} \geq 0$ holds for every feasible direction $\boldsymbol{p}$ at $\boldsymbol{x}^*$.*

[3] While that theorem was stated for sets defined only by linear inequalities, we can always rewrite the equalities $\boldsymbol{E}\boldsymbol{x} = \boldsymbol{d}$ as $\boldsymbol{E}\boldsymbol{x} \leq \boldsymbol{d}$, $-\boldsymbol{E}\boldsymbol{x} \leq -\boldsymbol{d}$; the corresponding feasible directions are then given by $\boldsymbol{E}\boldsymbol{p} \leq \boldsymbol{0}^{\ell}$, $-\boldsymbol{E}\boldsymbol{p} \leq \boldsymbol{0}^{\ell}$, that is, $\boldsymbol{E}\boldsymbol{p} = \boldsymbol{0}^{\ell}$.

(b) *Suppose that S is convex and that f is in C^1 on S. If $\boldsymbol{x}^* \in S$ is a local minimum of f on S then*

$$\nabla f(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0, \qquad \boldsymbol{x} \in S, \tag{4.10}$$

holds.

Notice that we above in (4.10) say that $\nabla f(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0$ holds for every $\boldsymbol{x} \in S$; the quantifier $\forall$ is not explicitly stated.

Proof. (a) We again utilize the Taylor expansion of f around $\boldsymbol{x}^*$:

$$f(\boldsymbol{x}^* + \alpha \boldsymbol{p}) = f(\boldsymbol{x}^*) + \alpha \nabla f(\boldsymbol{x}^*)^{\mathrm{T}} \boldsymbol{p} + o(\alpha).$$

The proof is by contradiction. As was shown in Proposition 4.15, if there is a direction $\boldsymbol{p}$ for which it holds that $\nabla f(\boldsymbol{x}^*)^{\mathrm{T}} \boldsymbol{p} < 0$, then $f(\boldsymbol{x}^* + \alpha \boldsymbol{p}) < f(\boldsymbol{x}^*)$ for all sufficiently small values of $\alpha > 0$. It suffices here to state that $\boldsymbol{p}$ should also be a feasible direction in order to reach a contradiction to the local optimality of $\boldsymbol{x}^*$.

(b) If S is convex then every feasible direction $\boldsymbol{p}$ can be written as a positive scalar times the vector $\boldsymbol{x} - \boldsymbol{x}^*$ for *some* vector $\boldsymbol{x} \in S$. (Why?) The expression (4.10) then follows from the statement in (a). ∎

The inequality (4.10) is sometimes referred to as a *variational inequality*. We will utilize it for several purposes: (i) to derive equivalent optimality conditions involving a linear optimization problem as well as the Euclidean projection operation Proj_S introduced in Section 3.4; (ii) to derive a descent algorithm for the problem (4.9); (iii) to derive a near-optimality condition for convex optimization problems; and (iv) we will extend it to non-convex sets in the form of the Karush–Kuhn–Tucker conditions.

In Theorem 4.13 we established that for unconstrained C^1 optimization the necessary optimality condition is that $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}^n$ holds. Notice that that is exactly what becomes of the variational inequality (4.10) when $S = \mathbb{R}^n$, because the only way in which that inequality can hold for every $\boldsymbol{x} \in \mathbb{R}^n$ is that $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}^n$ holds. Just as we did in the case of unconstrained optimization, we will call a vector $\boldsymbol{x}^* \in S$ satisfying (4.10) a *stationary point*.

We will next provide two statements equivalent to the variational inequality (4.10). First up, though, we will provide the extension to Theorem 4.18 to the convex constrained case. Notice the resemblance of their respective proofs.

Theorem 4.23 (necessary and sufficient global optimality conditions) *Suppose that $S \subseteq \mathbb{R}^n$ is nonempty and convex. Let $f \in C^1$ on S, and let f be* *convex. Then,*

$$\boldsymbol{x}^* \text{ is a global minimum of } f \text{ on } S \quad \Longleftrightarrow \quad (4.10) \text{ holds.}$$

Proof. [$\Longrightarrow$] This has already been shown in Proposition 4.22(b), since a global minimum is a local minimum.

[$\Longleftarrow$] The convexity of f yields [cf. Theorem 3.40(a)] that for every $\boldsymbol{y} \in S$,

$$\begin{aligned} f(\boldsymbol{y}) &\geq f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}^*) \\ &\geq f(\boldsymbol{x}^*), \end{aligned}$$

where the second inequality stems from (4.10), by assumption. ∎

First, we will provide the connection to the projection of a vector onto a convex set, discussed in Section 3.4. We claim that the property (4.10) is equivalent to

$$\boldsymbol{x}^* = \mathrm{Proj}_S[\boldsymbol{x}^* - \nabla f(\boldsymbol{x}^*)], \tag{4.11}$$

or, more generally,

$$\boldsymbol{x}^* = \mathrm{Proj}_S[\boldsymbol{x}^* - \alpha \nabla f(\boldsymbol{x}^*)], \qquad \alpha > 0.$$

In other words, a point is stationary if and only if a step in the direction of the steepest descent direction followed by a Euclidean projection onto S means that we have not moved at all. To prove this, we will utilize Proposition 4.22(b) for the optimization problem corresponding to this projection. We are interested in finding the point $\boldsymbol{x} \in S$ that minimizes the distance to the vector $\boldsymbol{z} := \boldsymbol{x}^* - \nabla f(\boldsymbol{x}^*)$. We can write this as a strictly convex optimization problem as follows:

$$\underset{\boldsymbol{x} \in S}{\text{minimize}}\ h(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2. \tag{4.12}$$

The necessary optimality conditions for this problem, as stated in Proposition 4.22(b), is that

$$\nabla h(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) \geq 0, \qquad \boldsymbol{y} \in S, \tag{4.13}$$

holds. Here, $\nabla h(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{z} = \boldsymbol{x} - [\boldsymbol{x}^* - \nabla f(\boldsymbol{x}^*)]$. Since h is convex, by Theorem 4.23, we know that the variational inequality (4.13) characterizes $\boldsymbol{x}$ as the globally optimal solution to the projection problem. We claimed that $\boldsymbol{x} = \boldsymbol{x}^*$ is the solution to this problem if and only if $\boldsymbol{x}^*$

is stationary in the problem (4.9). But this follows immediately, since the variational inequality (4.13), for the special choice of h and $\boldsymbol{x} = \boldsymbol{x}^*$, becomes

$$\nabla f(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}^*) \geq 0, \qquad \boldsymbol{y} \in S,$$

that is, a statement identical to (4.10). The characterization (4.11) is interesting in that it states that if $\boldsymbol{x}^*$ is *not* stationary, then the projection operation defined therein then must provide a step *away* from $\boldsymbol{x}^*$; this step will in fact yield a reduced value of f under some additional conditions on the step length α, and so it defines a descent algorithm for (4.9); see Exercise 4.5, and the text in Section 12.4.

So far, we have two equivalent characterizations of a stationary point of f at $\boldsymbol{x}^*$: (4.10) and (4.11). The following one is based on a linear optimization problem.

Notice that (4.10) states that $\nabla f(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{x} \geq \nabla f(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{x}^*$ for every $\boldsymbol{x} \in S$. Since we obtain equality by setting $\boldsymbol{x} = \boldsymbol{x}^*$ we see that $\boldsymbol{x}^*$ in fact is a globally optimal solution to the problem to

$$\underset{\boldsymbol{x} \in S}{\operatorname{minimize}} \ \nabla f(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{x}.$$

In other words, (4.10) is equivalent to the statement

$$\underset{\boldsymbol{x} \in S}{\operatorname{minimum}} \ \nabla f(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) = 0. \tag{4.14}$$

It is quite obvious that if at some point $\boldsymbol{x} \in S$,

$$\underset{\boldsymbol{y} \in S}{\operatorname{minimum}} \ \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) < 0,$$

then the direction of $\boldsymbol{p} := \boldsymbol{y} - \boldsymbol{x}$ is a feasible descent direction with respect to f at $\boldsymbol{x}$. Again, we have a building block of a descent algorithm for the problem (4.9). [The algorithms that immediately spring out from this characterization are called the *Frank–Wolfe* and *Simplicial decomposition* algorithms, when S is polyhedral; we notice that in the polyhedral case, the linear minimization problem is an LP problem. Read more about these algorithms in Sections 12.2 and 12.3.] Now having got three equivalent stationarity conditions, (4.10), (4.11), and (4.14), we finally provide a fourth one. This one is intimately associated with the projection operation, and it introduces an important geometric concept into the theory of optimality, namely the *normal cone* to a convex set S.

We studied a particular choice of $\boldsymbol{z}$ above, but let us consider an extension of Figure 3.15 which provided an image of the Euclidean projection.

Notice from the above arguments that if we wish to project the vector $\boldsymbol{z} \in \mathbb{R}^n$ onto S, then the resulting (unique) projection is the vector $\boldsymbol{x}$ for which the following holds:

$$[\boldsymbol{x} - \boldsymbol{z}]^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) \geq 0, \qquad \boldsymbol{y} \in S.$$

Changing sign for clarity, this is the same as

$$[\boldsymbol{z} - \boldsymbol{x}]^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) \leq 0, \qquad \boldsymbol{y} \in S.$$

The interpretation of this inequality is that the angle between the two vectors $\boldsymbol{z} - \boldsymbol{x}$ (the vector that points towards the point being projected) and the vector $\boldsymbol{y} - \boldsymbol{x}$ (the vector that points towards any vector $\boldsymbol{y} \in S$) is $\geq 90°$. So, the projection operation has the characterization

$$[\boldsymbol{z} - \mathrm{Proj}_S(\boldsymbol{z})]^{\mathrm{T}}(\boldsymbol{y} - \mathrm{Proj}_S(\boldsymbol{z})) \leq 0, \qquad \boldsymbol{y} \in S. \tag{4.15}$$

The above is summarized in Figure 4.4 for $\boldsymbol{x} = \boldsymbol{x}^*$ and $\boldsymbol{z} = \boldsymbol{x}^* - \nabla f(\boldsymbol{x}^*)$.

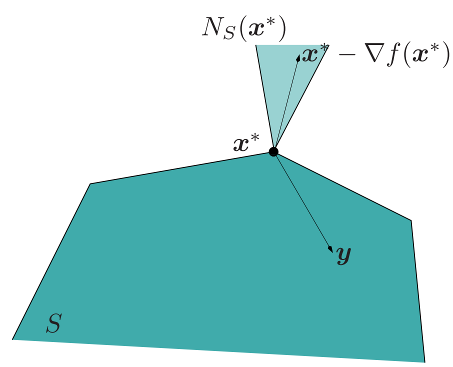


Figure 4.4: Normal cone characterization of a stationary point.

Here, the point being projected is $\boldsymbol{z} = \boldsymbol{x}^* - \nabla f(\boldsymbol{x}^*)$, as used in the characterization of stationarity.

What is left to complete the picture is to define the normal cone, depicted here as $N_S(\boldsymbol{x}^*)$ in the lighter shade.

Definition 4.24 (normal cone) *Suppose that the set $S \subseteq \mathbb{R}^n$ is closed and convex. Let $\boldsymbol{x} \in \mathbb{R}^n$. Then, the* normal cone *to S at $\boldsymbol{x}$ is the set*

$$N_S(\boldsymbol{x}) := \begin{cases} \{\, \boldsymbol{v} \in \mathbb{R}^n \mid \boldsymbol{v}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) \leq 0, \qquad \boldsymbol{y} \in S \,\}, & \text{if } \boldsymbol{x} \in S, \\ \emptyset, & \text{otherwise.} \end{cases} \tag{4.16}$$

■

According to the definition, we can now define our fourth characterization of a stationary point at $\boldsymbol{x}^*$ as follows:

$$-\nabla f(\boldsymbol{x}^*) \in N_S(\boldsymbol{x}^*). \tag{4.17}$$

What this condition states geometrically is that the angle between the negative gradient and any feasible direction is $\geq 90°$, which, of course, is the same as stating that at $\boldsymbol{x}^*$ there exist no feasible descent directions. The four conditions (4.10), (4.11), (4.14), and (4.17) are equivalent, and so according to Theorem 4.23(b) they all are also both necessary and sufficient for the global optimality of $\boldsymbol{x}^*$ as soon as f is convex.

We remark that in the special case when S is an affine subspace (such as the solution set of a number of linear equations, $S := \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{E}\boldsymbol{x} = \boldsymbol{d} \,\}$), the statement (4.17) means that at a stationary point $\boldsymbol{x}^*$, $\nabla f(\boldsymbol{x}^*)$ is parallel to the normal of the subspace.

The normal cone inclusion (4.17) will later be extended to more general sets, where S is described by a finite collection of possibly nonconvex constraints. The extension will lead us to the famous Karush–Kuhn–Tucker conditions in Chapter 5. [It turns out to be much more convenient to extend (4.17) than the other three characterizations of stationarity.]

We finish this section by proving a proposition on the behaviour of the gradient of the objective function f on the solution set S^* to convex problems of the form (4.1). The below result shows that ∇f enjoys a stability property, and it also extends the result from the unconstrained case where the value of ∇f always is zero on the solution set.

Proposition 4.25 (invariance of ∇f on the solution set of convex programs) *Suppose that $S \subseteq \mathbb{R}^n$ is convex and that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and in C^1 on S. Then, the value of $\nabla f(\boldsymbol{x})$ is constant on the optimal solution set S^*.*

Further, suppose that $\boldsymbol{x}^* \in S^*$. Then,

$$S^* = \{\, \boldsymbol{x} \in S \mid \nabla f(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) = 0 \text{ and } \nabla f(\boldsymbol{x}) = \nabla f(\boldsymbol{x}^*) \,\}.$$

Proof. Let $\boldsymbol{x}^* \in S^*$. The definition of the convexity of f shows that

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \nabla f(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*), \qquad \boldsymbol{x} \in \mathbb{R}^n. \tag{4.18}$$

Let $\bar{\boldsymbol{x}} \in S^*$. Then, it follows that $\nabla f(\boldsymbol{x}^*)^{\mathrm{T}}(\bar{\boldsymbol{x}} - \boldsymbol{x}^*) = 0$. By substituting $\nabla f(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{x}^*$ with $\nabla f(\boldsymbol{x}^*)^{\mathrm{T}}\bar{\boldsymbol{x}}$ in (4.18) and using that $f(\boldsymbol{x}^*) = f(\bar{\boldsymbol{x}})$, we obtain that

$$f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) \geq \nabla f(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \bar{\boldsymbol{x}}), \qquad \boldsymbol{x} \in \mathbb{R}^n,$$

which is equivalent to the statement that $\nabla f(\bar{\boldsymbol{x}}) = \nabla f(\boldsymbol{x}^*)$. We are done. $\blacksquare$

4.5 Near-optimality in convex optimization

We will here utilize Theorem 4.23 in order to provide a measure of the distance to the optimal solution in terms of the value of f at any feasible point $\boldsymbol{x}$.

Let $\boldsymbol{x} \in S$, and suppose that f is convex on S. Suppose also that $\boldsymbol{x}^* \in S$ is an arbitrary globally optimal solution, which we suppose exists. From the necessary optimality conditions stated in Proposition 4.22(b), it is clear that unless $\boldsymbol{x}$ solves (4.9) there exists a $\boldsymbol{y} \in S$ such that $\nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}) < 0$, and hence $\boldsymbol{p} := \boldsymbol{y}-\boldsymbol{x}$ is a feasible descent direction.

Suppose now that

$$\bar{\boldsymbol{y}} \in \arg \operatorname*{minimum}_{\boldsymbol{y} \in S} \; z(\boldsymbol{y}) := \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}). \tag{4.19}$$

Consider the following string of inequalities and equalities:

$$\begin{aligned} f(\boldsymbol{x}) + z(\bar{\boldsymbol{y}}) &= f(\boldsymbol{x}) + \operatorname*{minimum}_{\boldsymbol{y} \in S} \; z(\boldsymbol{y}) \\ &\leq f(\boldsymbol{x}) + z(\boldsymbol{x}^*) \\ &\leq f(\boldsymbol{x}^*) \\ &\leq f(\boldsymbol{x}). \end{aligned}$$

The equality follows by definition; the first inequality stems from the fact that $\bar{\boldsymbol{y}}$ solves the linear minimization problem, while the vector $\boldsymbol{x}^*$ may not; the second inequality follows from the convexity of f on S [cf. Theorem 3.40(a)]; the final inequality follows from the global optimality of $\boldsymbol{x}^*$ and the feasibility of $\boldsymbol{x}$.

From the above, we obtain a closed interval wherein we know that the optimal value of the problem (4.9) lies. Let $f^* := \operatorname{minimum}_{\boldsymbol{x} \in S} f(\boldsymbol{x}) = f(\boldsymbol{x}^*)$. Then, for every $\boldsymbol{x} \in S$,

$$f^* \in [f(\boldsymbol{x}) + z(\bar{\boldsymbol{y}}), f(\boldsymbol{x})]. \tag{4.20}$$

Clearly, the length of the interval is defined by how far from zero the value of $z(\bar{\boldsymbol{y}})$ is. Suppose then that $z(\bar{\boldsymbol{y}}) \geq -\varepsilon$, for some small value $\varepsilon > 0$. (In an algorithm where a sequence $\{\boldsymbol{x}_k\}$ is constructed such that it converges to an optimal solution, this will eventually happen for every $\varepsilon > 0$.) Then, from the above we obtain that $f(\boldsymbol{x}^*) \geq f(\boldsymbol{x}) + z(\bar{\boldsymbol{y}}) \geq f(\boldsymbol{x}) - \varepsilon$; in short,

$$f(\boldsymbol{x}^*) \geq f(\boldsymbol{x}) - \varepsilon, \quad \text{or,} \quad f(\boldsymbol{x}) \leq f^* + \varepsilon. \tag{4.21}$$

We refer to a vector $\boldsymbol{x} \in S$ satisfying the inequality (4.21) as an *ε-optimal solution*. From the above LP problem we hence have a simple instrument for evaluating the quality of a feasible solution in our problem. Note, again, that convexity is a crucial property enabling this possibility.

As far as iterative algorithms go, it is quite often the case that for the problem (4.9) involving a convex feasible set the sequences $\{\boldsymbol{x}_k\}$ of iterates do not necessarily stay inside the feasible set S. The reason is that even if the constraints are convex inequalities it is difficult to check when one reaches the boundary of S. We mention however two cases where feasible algorithms (that is, those for which $\{\boldsymbol{x}_k\} \subset S$ holds) are viable:

(I) When S is a polyhedral set, then it is only a matter of solving a series of simple linear systems to check for the maximum step length along a feasible direction. Among the algorithms that actually are feasible we count the *simplex method* for linear programming (LP) problems, the *Frank–Wolfe method* which builds on the fact that the lower bounds and descent directions discussed above rely on solving such LP problems, and the *projection methods* which build on the property investigated in Exercise 4.5. More on these algorithms will be said in Chapter 12.

(II) When the set S has an interior point, we may replace the constraints with an *interior penalty function* which has an asymptote whenever approaching the boundary, thus automatically ensuring that iterates stay (strictly) feasible. More on a class of methods based on this penalty function is said in Chapter 13.

4.6 Applications

4.6.1 *Continuity of convex functions

A remarkable property of any convex function is that without any additional assumptions it can be shown to be *continuous* relative to any open convex set in the intersection of its effective domain and its affine hull.[4] We establish a special case below, in which relative interior is replaced by interior for simplicity.

Theorem 4.26 (continuity of convex functions) *Suppose that $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a convex function, and consider an open convex subset S of its effective domain. The function f is continuous on S.*

[4] In other words, it is continuous relative to any relatively open convex subset of its effective domain.

Proof. Let $\bar{\boldsymbol{x}} \in S$. To establish continuity of f at $\bar{\boldsymbol{x}}$, we must show that given $\varepsilon > 0$, there exists $\delta > 0$ such that $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq \delta$ implies that $|f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}})| \leq \varepsilon$. We establish this property in two parts, by showing that f is both lower and upper semi-continuous at $\bar{\boldsymbol{x}}$.

[upper semi-continuity] By the openness of S, there exists $\delta' > 0$ with $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq \delta$, implying $\boldsymbol{x} \in S$. Construct the value of the scalar γ as follows:

$$\gamma := \underset{i \in \{1,2,\ldots,n\}}{\text{maximum}} \left\{ \text{maximum} \left\{ f(\bar{\boldsymbol{x}} + \delta' \boldsymbol{e}_i) - f(\bar{\boldsymbol{x}}), f(\bar{\boldsymbol{x}} - \delta' \boldsymbol{e}_i) - f(\bar{\boldsymbol{x}}) \right\} \right\}, \tag{4.22}$$

where $\boldsymbol{e}_i$ is the i^{th} unit vector in $\mathbb{R}^n$. Note that $0 \leq \gamma < \infty$. Let now

$$\delta := \text{minimum} \left\{ \frac{\delta'}{n}, \frac{\varepsilon\delta'}{\gamma n} \right\}. \tag{4.23}$$

Choose an $\boldsymbol{x}$ with $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq \delta$. For every $i \in \{1, 2, \ldots, n\}$, if $x_i \geq \bar{x}_i$ then define $\boldsymbol{z}_i := \delta' \boldsymbol{e}_i$, otherwise $\boldsymbol{z}_i := -\delta' \boldsymbol{e}_i$. Then,

$$\boldsymbol{x} - \bar{\boldsymbol{x}} = \sum_{i=1}^{n} \alpha_i \boldsymbol{z}_i,$$

where $\alpha_i \geq 0$ for all i. Moreover,

$$\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| = \delta' \|\boldsymbol{\alpha}\|. \tag{4.24}$$

From (4.23), and since $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq \delta$, it follows that $\alpha_i \leq 1/n$ for all i. Hence, by the convexity of f and since $0 \leq \alpha_i n \leq 1$, we get

$$\begin{aligned}
f(\boldsymbol{x}) = f\left(\bar{\boldsymbol{x}} + \sum_{i=1}^{n} \alpha_i \boldsymbol{z}_i \right) &= f\left[\frac{1}{n} \sum_{i=1}^{n} (\bar{\boldsymbol{x}} + \alpha_i n \boldsymbol{z}_i) \right] \\
&\leq \frac{1}{n} \sum_{i=1}^{n} f(\bar{\boldsymbol{x}} + \alpha_i n \boldsymbol{z}_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} f[(1 - \alpha_i n)\bar{\boldsymbol{x}} + \alpha_i n(\bar{\boldsymbol{x}} + \boldsymbol{z}_i)] \\
&\leq \frac{1}{n} \sum_{i=1}^{n} [(1 - \alpha_i n) f(\bar{\boldsymbol{x}}) + \alpha_i n f(\bar{\boldsymbol{x}} + \boldsymbol{z}_i)].
\end{aligned}$$

Therefore, $f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) \leq \sum_{i=1}^{n} \alpha_i [f(\bar{\boldsymbol{x}} + \boldsymbol{z}_i) - f(\bar{\boldsymbol{x}})]$. From (4.22) it is obvious that $f(\bar{\boldsymbol{x}} + \boldsymbol{z}_i) - f(\bar{\boldsymbol{x}}) \leq \gamma$ for each i; and since $\alpha_i \geq 0$, it follows that

$$f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) \leq \gamma \sum_{i=1}^{n} \alpha_i. \tag{4.25}$$

Noting (4.23), (4.24), it follows that $\alpha_i \leq \varepsilon/n\gamma$, and (4.25) implies that $f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) \leq \varepsilon$. Hence, we have so far shown that $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq \delta$ implies that $f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) \leq \varepsilon$. By Definition 4.5(b), f hence is upper semi-continuous at $\bar{\boldsymbol{x}}$.

[lower semi-continuity] Let $\boldsymbol{y} := 2\bar{\boldsymbol{x}} - \boldsymbol{x}$, and note that $\|\boldsymbol{y} - \bar{\boldsymbol{x}}\| \leq \delta$. Therefore, as above,

$$f(\boldsymbol{y}) - f(\bar{\boldsymbol{x}}) \leq \varepsilon. \tag{4.26}$$

But $\bar{\boldsymbol{x}} = \frac{1}{2}\boldsymbol{y} + \frac{1}{2}\boldsymbol{x}$, and by the convexity of f,

$$f(\bar{\boldsymbol{x}}) \leq \frac{1}{2} f(\boldsymbol{y}) + \frac{1}{2} f(\boldsymbol{x})$$

follows. Combining this inequality with (4.26), it follows that $f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}) \leq \varepsilon$, whence Definition 4.5(b) applies. We are done. ∎

Note that convex functions need not be continuous everywhere; by the above theorem we know however that points of non-continuity must occur at the boundary of the effective domain of f. For example, check the continuity of the following function:

$$f(x) := \begin{cases} x^2, & \text{for } |x| < 1, \\ 2, & \text{for } |x| = 1. \end{cases}$$

4.6.2 The Separation Theorem

The previously established Weierstrass Theorem 4.6 will now be utilized together with the above variational inequality characterization (4.10) of stationary points in order to finally establish the Separation Theorem 3.24. For simplicity, we rephrase the theorem.

Theorem 4.27 (Separation Theorem) *Suppose that the set $S \subseteq \mathbb{R}^n$ is closed and convex, and that the point $\boldsymbol{y}$ does not lie in S. Then there exist a vector $\boldsymbol{\pi} \neq \boldsymbol{0}^n$ and $\alpha \in \mathbb{R}$ such that $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{y} > \alpha$ and $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{x} \leq \alpha$ for all $\boldsymbol{x} \in S$.*

Proof. We may assume that S is nonempty, and define a function $f : \mathbb{R}^n \to \mathbb{R}$ through $f(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2/2$, $\boldsymbol{x} \in \mathbb{R}^n$. Now by Weierstrass' Theorem 4.6) there exists a minimizer $\tilde{\boldsymbol{x}}$ of f over S, which by the first order necessary condition [see Proposition 4.22(b)] satisfies $(\boldsymbol{y} - \tilde{\boldsymbol{x}})^{\mathrm{T}}(\boldsymbol{x} - \tilde{\boldsymbol{x}}) \leq 0$ for all $\boldsymbol{x} \in S$ (since $-\nabla f(\tilde{\boldsymbol{x}}) = \boldsymbol{y} - \tilde{\boldsymbol{x}}$). Now setting $\boldsymbol{\pi} = \boldsymbol{y} - \tilde{\boldsymbol{x}}$ and $\alpha = (\boldsymbol{y} - \tilde{\boldsymbol{x}})^{\mathrm{T}}\tilde{\boldsymbol{x}}$ gives the result sought. ∎

A slightly different separation theorem will be used in the Lagrangian duality theory in Chapter 6. We state it without proof.

Theorem 4.28 (separation of convex sets) *Each pair of disjoint nonempty convex sets A and B in $\mathbb{R}^n$ can be separated by a hyperplane in $\mathbb{R}^n$, that is, there exists a vector $\boldsymbol{\pi} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ such that $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{x} \leq \alpha$ for all $\boldsymbol{x} \in A$ and $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{y} \geq \alpha$ for all $\boldsymbol{y} \in B$.* ■

Remark 4.29 The main difference between the Separation Theorems 3.24 and 4.28 is that in Theorem 3.24 there exists a hyperplane that in fact *strictly* separates the point $\boldsymbol{y}$ and the closed convex set C, that is, there exists a vector $\boldsymbol{\pi} \in \mathbb{R}^n$ and an $\alpha \in \mathbb{R}$ such that $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{y} > \alpha$ while $\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{x} < \alpha$ holds for all $\boldsymbol{x} \in C$. In Theorem 4.28, however, this is not true. Consider, for example, the sets $A = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_2 \leq 0 \,\}$ and $B = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 > 0;\ x_2 \geq 1/x_1 \,\}$. Then, the line $\{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_2 = 0 \,\}$ separates A and B, but the sets can not be strictly separated. ■

4.6.3 Euclidean projection

We will finish our discussions on the projection operation, which was defined in Section 3.4, by establishing an interesting continuity property.

Definition 4.30 (non-expansive operator) *Suppose that $S \subseteq \mathbb{R}^n$ is closed and convex. Let $\boldsymbol{f} : S \to S$ denote a vector-valued operator from S to S. We say that $\boldsymbol{f}$ is* non-expansive *if, as a result of applying the mapping $\boldsymbol{f}$, the distance between any two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in S does not increase.*

In other words, the operator $\boldsymbol{f}$ is non-expansive on S if

$$\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|, \qquad \boldsymbol{x}, \boldsymbol{y} \in S, \tag{4.27}$$

holds.

Theorem 4.31 (the projection operation is non-expansive) *Let S be a nonempty, closed and convex set in $\mathbb{R}^n$. For every $\boldsymbol{x} \in \mathbb{R}^n$, its projection $\operatorname{Proj}_S(\boldsymbol{x})$ is uniquely defined. The operator $\operatorname{Proj}_S : \mathbb{R}^n \to S$ is non-expansive on $\mathbb{R}^n$, and therefore in particular continuous.*

Proof. The uniqueness of the operation is the result of the fact that the objective function $\boldsymbol{x} \mapsto \|\boldsymbol{x} - \boldsymbol{z}\|$ (or, $\boldsymbol{x} \mapsto \|\boldsymbol{x} - \boldsymbol{z}\|^2$) is both weakly coercive and strictly convex on S, so there exists a unique optimal solution to the projection problem for every $\boldsymbol{z} \in \mathbb{R}^n$. (Cf. Weierstrass' Theorem 4.6 and Proposition 4.10, respectively.)

Next, take $\boldsymbol{x}^1, \boldsymbol{x}^2 \in \mathbb{R}^n$. Then, by the characterization (4.15) of the Euclidean projection,

$$\begin{aligned}
&[\operatorname{Proj}_S(\boldsymbol{x}^2) - \operatorname{Proj}_S(\boldsymbol{x}^1)]^{\mathrm{T}}(\boldsymbol{x}^1 - \operatorname{Proj}_S(\boldsymbol{x}^2)) \leq 0, \\
&[\operatorname{Proj}_S(\boldsymbol{x}^1) - \operatorname{Proj}_S(\boldsymbol{x}^2)]^{\mathrm{T}}(\boldsymbol{x}^2 - \operatorname{Proj}_S(\boldsymbol{x}^1)) \leq 0.
\end{aligned}$$

Summing the two inequalities yields

$$\begin{aligned}\|\mathrm{Proj}_S(\boldsymbol{x}^2) - \mathrm{Proj}_S(\boldsymbol{x}^1)\|^2 &\leq [\mathrm{Proj}_S(\boldsymbol{x}^2) - \mathrm{Proj}_S(\boldsymbol{x}^1)]^{\mathrm{T}}(\boldsymbol{x}^2 - \boldsymbol{x}^1) \\ &\leq \|\mathrm{Proj}_S(\boldsymbol{x}^2) - \mathrm{Proj}_S(\boldsymbol{x}^1)\| \cdot \|\boldsymbol{x}^2 - \boldsymbol{x}^1\|,\end{aligned}$$

that is, $\|\mathrm{Proj}_S(\boldsymbol{x}^2) - \mathrm{Proj}_S(\boldsymbol{x}^1)\| \leq \|\boldsymbol{x}^2 - \boldsymbol{x}^1\|$. Since this is true for every pair $(\boldsymbol{x}^1, \boldsymbol{x}^2) \in \mathbb{R}^n$, we have shown that the operator Proj_S is non-expansive on $\mathbb{R}^n$. In particular, non-expansive functions are continuous. (The proof of the latter is left as an exercise.) ∎

The theorem is illustrated in Figure 4.5.

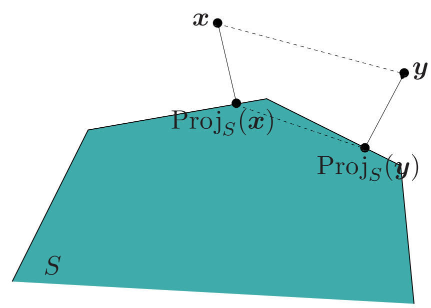


Figure 4.5: The projection operation is non-expansive.

4.6.4 Fixed point theorems

Fixed point theorems state properties of a problem of the following form: Suppose the mapping $\boldsymbol{f}$ is defined on a closed, convex set S in $\mathbb{R}^n$ and that $\boldsymbol{f}(\boldsymbol{x}) \subseteq S$ for every $\boldsymbol{x} \in S$. Is there an $\boldsymbol{x} \in S$ such that $\boldsymbol{f}$ maps $\boldsymbol{x}$ onto itself (that is, onto $\boldsymbol{x}$), or, in other words,

$$\exists \boldsymbol{x} \in S \text{ such that } \boldsymbol{x} \in \boldsymbol{f}(\boldsymbol{x})?$$

Such a point is called a *fixed point* of $\boldsymbol{f}$ over the set S. If the mapping $\boldsymbol{f}$ is single-valued rather than set-valued then the question boils down to:

$$\exists \boldsymbol{x} \in S \text{ such that } \boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x})?$$

Many questions in optimization and analysis can be reduced to the analysis of a fixed point problem. For example, an optimization problem can in some circumstances be reduced to a fixed point problem, in which case the question of the existence of solutions to the optimization problem can be answered by studying the fixed point problem. Further, the optimality conditions analyzed in Section 4.4 can be written as the solution to a fixed point problem; we can therefore equate the search for a stationary point with that of finding a fixed point of a particular function $\boldsymbol{f}$. This type of analysis is quite useful also when analyzing the convergence of iterative algorithms for optimization problems.

4.6.4.1 Theory

We begin by studying some classical fixed point theorems, and then we provide examples of the connections between the results in Section 4.4 with fixed point theory.

Definition 4.32 (contractive operator) *Let S be a nonempty, closed and convex set in $\mathbb{R}^n$. Let $\boldsymbol{f}$ be a mapping from S to S. We say that $\boldsymbol{f}$ is* contractive *on S if, as a result of applying the mapping $\boldsymbol{f}$, the distance between any two distinct vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in S decreases.*

In other words, the operator $\boldsymbol{f}$ is contractive on S if there exists $\alpha \in [0, 1)$ such that

$$\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})\| \leq \alpha \|\boldsymbol{x} - \boldsymbol{y}\|, \qquad \boldsymbol{x}, \boldsymbol{y} \in S, \tag{4.28}$$

holds. ■

Clearly, a contractive operator is non-expansive.

In the below result we utilize the notion of a *geometric convergence rate*; while its definition is in fact given in the result below, we also refer to Sections 6.5.1 and 11.10 for more detailed discussions on convergence rates.

Theorem 4.33 (fixed point theorems) *Let S be a nonempty, closed and convex set in $\mathbb{R}^n$.*

(a) [Banach's Theorem] *Let $\boldsymbol{f}$ be a contraction mapping from S to S. Then, $\boldsymbol{f}$ has a unique fixed point $\boldsymbol{x}^* \in S$. Further, for every initial vector $\boldsymbol{x}_0 \in S$, the iteration sequence $\{\boldsymbol{x}_k\}$ defined by the fixed-point iteration*

$$\boldsymbol{x}_{k+1} := \boldsymbol{f}(\boldsymbol{x}_k), \qquad k = 0, 1, \dots, \tag{4.29}$$

converges geometrically to the unique fixed point $\boldsymbol{x}^*$. In particular,

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq \alpha^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|, \qquad k = 0, 1, \dots .$$

(b) [Brouwer's Theorem] *Let S further be bounded, and assume merely that $\boldsymbol{f}$ is continuous. Then, $\boldsymbol{f}$ has a fixed point.*

Proof. (a) For any $\boldsymbol{x}_0 \in S$, consider the sequence $\{\boldsymbol{x}_k\}$ defined by (4.29). Then, for any $p \geq 1$,

$$\begin{aligned}\|\boldsymbol{x}_{k+p} - \boldsymbol{x}_k\| &\leq \sum_{i=1}^{p} \|\boldsymbol{x}_{k+i} - \boldsymbol{x}_{k+i-1}\| \\ &\leq (\alpha^{p-1} + \cdots + 1)\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| \leq [\alpha^k/(1-\alpha)]\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|.\end{aligned}$$

Hence, $\{\boldsymbol{x}_k\}$ is a Cauchy sequence and thus converges. By continuity, the limit point is the unique fixed point.

The convergence speed follows from the identification

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| = \|\boldsymbol{f}(\boldsymbol{x}_{k-1}) - \boldsymbol{f}(\boldsymbol{x}^*)\| \leq \alpha\|\boldsymbol{x}_{k-1} - \boldsymbol{x}^*\|, \qquad k = 1, 2, \ldots.$$

Applying this relation recursively yields the result.

(b) [Sketch] In short, the proof is to first establish that any C^1 function on the unit ball has a fixed point there. Extending the reasoning to merely continuous operators is possible, because of the Stone–Weierstrass Theorem (which states that for any continuous operator defined on the unit ball there is a sequence of C^1 functions defined on the unit ball that uniformly converges to it). Each of these functions can be established to have a fixed point, and because of the compactness of the unit ball, so does the merely continuous limit function. For our final argument, we can assume that the set S has a nonempty interior. Then there exists a homeomorphism[5] $\boldsymbol{h} : S \to B$, where B is the unit ball. Since the composite mapping $\boldsymbol{h} \circ \boldsymbol{f} \circ \boldsymbol{h}^{-1}$ is a continuous operator from B to B it has a fixed point $\boldsymbol{y}$ in B; therefore, $\boldsymbol{h}^{-1}(\boldsymbol{y})$ is a fixed point of $\boldsymbol{f}$. ∎

The result in (a) is due to Banach [Ban22]; the result in (b) is due to Brouwer [Bro09, Bro12], and Hadamard [Had10].

A special case in one variable of the result in (b) is illustrated in Figure 4.6.

4.6.4.2 Applications

Particularly the result of Theorem 4.33(b) is quite remarkably strong. We provide some sample consequences of it below. In each case, we ask the reader to find the pair $(S, \boldsymbol{f})$ defining the corresponding fixed point problem.

- [Mountaineering] You climb a mountain, following a trail, in six hours (noon to 6 PM). You camp on top overnight. Then at noon

[5] The given function $\boldsymbol{h}$ is a *homeomorphism* if it is a continuous operator which is *onto*—that is, its range, $\boldsymbol{h}(S)$, is identical to the set B defining its image set—and has a continuous inverse.

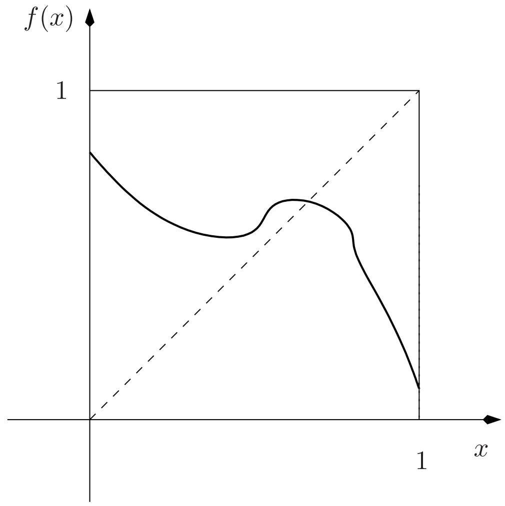


Figure 4.6: Consider the case $S = [0, 1]$, and a continuous function $f : S \to S$. Brouwer's Theorem states that there exists an $x^* \in S$ with $f(x^*) = x^*$. This is the same as saying that the continuous curve starting at $(0, f(0))$ and ending at $(1, f(1))$ must pass through the line $y = x$ inside the square.

the next day, you start descending. The descent is easier, and you make much better time. After an hour, you notice that your compass is missing, and you turn around and ascend a short distance, where you find your compass. You sit on a rock to admire the view. Then you descend the rest of the way. The entire descent takes four hours (noon to 4 PM). Along the trail there must then be a place where you were at the same place at the same time on both days.

- [Maps] Suppose you have two city maps over Gothenburg, which are not of the same scale. You crumple one of them up into a loose ball and place it on top of the other map entirely within the borders of the Gothenburg region on the flat map. Then, there is a point on the crumpled map (that represents the same place in Gothenburg on both maps) that is directly over its twin on the flat map. (A more simple problem is defined by a non-crumpled map and the city of Gothenburg itself; lay down the map anywhere in Gothenburg, and at least one point on the map will lie over that exact spot in real-life Gothenburg.)

- [Raking of gravel] Suppose you wish to rake the gravel in your garden; if the area is, say, circular, then any continuous raking will leave at least one tiny stone (which one is a function of time) in the same place.
- [Stirring coffee] Stirring the contents of a (convex) coffee cup in a continuous way, no matter how long you stir, some particle (which one is a function of time) will stay in the same position as it did before you began stirring.[6]
- [Meteorology] Even as the wind blows across the Earth there will be one location where the wind is perfectly vertical (or, perfectly calm). This fact actually implies the existence of cyclones; not to mention whorls, or crowns, in your hair no matter how you comb it. (The latter result also bears its own name: The Hairy Ball Theorem; cf. [BoL00, pp. 186–187].)

Applying fixed point theorems to our own development of this book, we take a look at the variational inequality (4.10). Rephrasing it in a more general form, the variational inequality problem is to find $\boldsymbol{x}^* \in S$ such that

$$\boldsymbol{f}(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0, \qquad \boldsymbol{x} \in S. \tag{4.30}$$

In order to turn it into a fixed point problem, we construct the following composite operator from $\mathbb{R}^n$ to S:

$$\boldsymbol{F} := (\boldsymbol{I}^n - \boldsymbol{f}) \circ \mathrm{Proj}_S,$$

or, in other words,

$$\boldsymbol{F}(\boldsymbol{x}) := \mathrm{Proj}_S(\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{x})), \qquad \boldsymbol{x} \in \mathbb{R}^n,$$

and consider finding a fixed point of $\boldsymbol{F}$ on S. Why is this operator a correct one? Because it is equivalent to the statement that

$$\mathrm{Proj}_S(\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{x}!$$

The special case for $\boldsymbol{f} = \nabla f$ is found in (4.11). Applying a fixed point theorem to the above problem then proves that the variational inequality problem (4.30) has solutions whenever $\boldsymbol{f}$ is continuous and S is nonempty, convex and compact. (Moreover, we have immediately found an iterative algorithm for the variational inequality problem: if the operator $\boldsymbol{x} \mapsto \mathrm{Proj}_S(\boldsymbol{x} - \alpha\boldsymbol{f}(\boldsymbol{x}))$ is contractive for some $\alpha > 0$, then it defines a convergent algorithm.)

[6] Ever wondered why adding lots of sugar does not always help improve the taste of coffee?

At the same time, we saw that the fixed point problem was defined through the same type of stationarity condition that we derived in Section 4.4 for differentiable optimization problems over convex sets. We have thereby also illustrated that stationarity in an optimization problem is intimately associated with fixed points of a particular operator.[7]

As an exercise, we consider the problem to find an $x \in \mathbb{R}$ such that $f(x) = 0$, where $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable. The classic Newton–Raphson algorithm has an iteration formula of the form

$$x_0 \in \mathbb{R}; \qquad x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \dots .$$

If we assume that there exists a zero at x^* at which $f'(x^*) > 0$, then by starting close enough to x^* we can prove that the above iteration formula defines a contraction, and hence we can establish local convergence. (Why?) Further analyses of Newton methods will be performed in Chapter 11.

A similar technique can be used to establish that a system of linear equations with a symmetric matrix is solvable by the classic Jacobi algorithm in numerical analysis if the matrix is diagonally dominant; this condition is equivalent to the Jacobi algorithm's algorithm-defining operator being a contraction. (Similar, but stronger, results can also be obtained for the Gauss–Seidel algorithm; cf. [Kre78, BeT89].)

An elegant application of fixed point theorems is the analysis of *matrix games*. The famous Minimax Theorem of von Neumann is associated with the existence of a saddle point of a function of the form $(\boldsymbol{v}, \boldsymbol{w}) \mapsto L(\boldsymbol{v}, \boldsymbol{w}) := \boldsymbol{v}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{w}$. Von Neumann's minimax theorem states that if V and W both are nonempty, convex and compact, then

$$\operatorname*{minimum}_{\boldsymbol{v} \in V} \operatorname*{maximum}_{\boldsymbol{w} \in W} \boldsymbol{v}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{w} = \operatorname*{maximum}_{\boldsymbol{w} \in W} \operatorname*{minimum}_{\boldsymbol{v} \in V} \boldsymbol{v}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{w}.$$

In order to prove this theorem we can use the above existence theorem for variational inequalities. Let

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{w} \end{pmatrix}; \quad \boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} -\boldsymbol{A}^{\mathrm{T}} \boldsymbol{v} \\ \boldsymbol{A} \boldsymbol{w} \end{pmatrix}; \quad S = V \times W.$$

It is a reasonably simple exercise to prove that the variational inequality (4.30) with the above identifications is equivalent to the saddle point conditions, which can also be written as the existence of a pair $(\boldsymbol{v}^*, \boldsymbol{w}^*) \in V \times W$ such that

$$(\boldsymbol{v}^*)^{\mathrm{T}} \boldsymbol{A} \boldsymbol{w} \leq (\boldsymbol{v}^*)^{\mathrm{T}} \boldsymbol{A} \boldsymbol{w}^* \leq \boldsymbol{v}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{w}^*, \qquad (\boldsymbol{v}, \boldsymbol{w}) \in V \times W;$$

[7] The book [Pat98] analyzes a large variety of optimization algorithms by utilizing this connection.

and we are done immediately.

Saddle point results will be returned to in the study of (Lagrangian) duality in the coming chapters, especially for linear programming (which was also von Neumann's special interest).

4.7 Notes and further reading

Most of the material of this chapter is elementary (as is relies mostly on the Taylor expansion of differentiable functions), and can be found in most basic books on nonlinear optimization, such as [Man69, Zan69, Avr76, BSS93, Ber99].

Weierstrass' Theorem 4.6 is the strongest existence result for optimal solutions that does not utilize convexity. The result is credited to Karl Weierstrass, but it was in fact known already by Bernard Bolzano in 1817 (although then only available in manuscript form); it has strong connections to the theorem of the existence of intermediate values as well as to that on the existence of limit points of every bounded sequence (now often referred to as the Bolzano–Weierstrass Theorem), and the notion of Cauchy sequences, often also credited to Weierstrass and Augustin-Louis Cauchy, respectively.

The Frank–Wolfe Theorem in Corollary 4.8 is found in [FrW56]. The stronger result in Theorem 4.7 is found in [Eav71, BlO72]. Proposition 4.25 on the invariance of the gradient on the solution set is found in [Man88, BuF91].

Fixed point theorems are developed in greater detail in [GrD03]. Non-cooperative game theory was developed in work by John von Neumann, together with Oskar Morgenstern (see [vNe28, vNM43]), and by John Nash [Nas50, Nas51].

4.8 Exercises

Exercise 4.1 (redundant constraints) Consider the problem to

$$\begin{aligned} &\text{minimize} \;\; f(\boldsymbol{x}), \\ &\text{subject to} \;\; g(\boldsymbol{x}) \leq b, \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are continuous functions, and $b \in \mathbb{R}$. Suppose that this problem has a globally optimal solution, $\boldsymbol{x}^*$, and that $g(\boldsymbol{x}^*) < b$ holds.

Claim: The vector $\boldsymbol{x}^*$ is also a globally optimal solution to the unconstrained problem to

$$\begin{aligned}&\text{minimize} \;\; f(\boldsymbol{x}),\\ &\text{subject to} \;\; \boldsymbol{x} \in \mathbb{R}^n.\end{aligned}$$

Is the claim true? ■

Exercise 4.2 (unconstrained optimization, exam 020826) Consider the unconstrained optimization problem to minimize the function

$$f(\boldsymbol{x}) := \frac{3}{2}(x_1^2 + x_2^2) + (1+a)x_1x_2 - (x_1 + x_2) + b$$

over $\mathbb{R}^2$, where a and b are real-valued parameters. Find all values of a and b such that the problem has a unique optimal solution. ■

Exercise 4.3 (spectral theory and unconstrained optimization) Let $\boldsymbol{A}$ be a symmetric $n \times n$ matrix. For $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \neq \boldsymbol{0}^n$, consider the function $\rho(\boldsymbol{x}) := \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}$, and the related optimization problem to

$$\underset{\boldsymbol{x} \neq \boldsymbol{0}^n}{\text{minimize}} \; \rho(\boldsymbol{x}). \tag{P}$$

Determine all the stationary points as well as the global minima in the minimization problem (P). Interpret the result in terms of linear algebra. ■

Exercise 4.4 (non-convex QP over subspaces) The Frank–Wolfe Theorem 4.8 can be further improved for some special cases of linear constraints. Suppose that $f(\boldsymbol{x}) := \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{q}^{\mathrm{T}}\boldsymbol{x}$, where $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ is a symmetric matrix and $\boldsymbol{q} \in \mathbb{R}^n$. Suppose further that the constraints are equalities, that is, that the ℓ constraints define the linear system $\boldsymbol{E}\boldsymbol{x} = \boldsymbol{d}$, where $\boldsymbol{E} \in \mathbb{R}^{\ell \times n}$ and $\boldsymbol{d} \in \mathbb{R}^\ell$. Note that the problem may not be convex, as we have not assumed that $\boldsymbol{Q}$ is positive semi-definite.

For this set-up, establish the following:

(a) Every locally optimal solution is a globally optimal solution.

(b) A locally [hence globally, by (a)] optimal solution exists if and only if f is lower bounded on $S := \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{E}\boldsymbol{x} = \boldsymbol{d} \,\}$. ■

Exercise 4.5 (descent from projection) Consider the problem (4.9), where f is in C^1 on the convex set S. Let $\boldsymbol{x} \in S$. Let $\alpha > 0$, and define

$$\boldsymbol{p} := \mathrm{Proj}_S[\boldsymbol{x} - \alpha\nabla f(\boldsymbol{x})] - \boldsymbol{x}.$$

Notice that $\boldsymbol{p}$ is a feasible direction at $\boldsymbol{x}$. Establish that

$$\nabla f(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{p} \leq -\frac{1}{\alpha} \|\boldsymbol{p}\|^2$$

holds. Hence, $\boldsymbol{p}$ is zero if and only if $\boldsymbol{x}$ is stationary [according to the characterization in (4.11)], and if $\boldsymbol{p}$ is non-zero then it defines a feasible descent direction with respect to f at $\boldsymbol{x}$. ∎

Exercise 4.6 (optimality conditions for a special problem) Suppose that $f \in C^1$ on the set $S := \{\, \boldsymbol{x} \in \mathbb{R}^n \mid x_j \geq 0, \quad j = 1, 2, \dots, n \,\}$, and consider the problem of finding a minimum of $f(\boldsymbol{x})$ over S. Develop the necessary optimality conditions for this problem in a compact form. ∎

Exercise 4.7 (optimality conditions for a special problem) Consider the problem to

$$\begin{aligned} \text{maximize} \quad & f(\boldsymbol{x}) := x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}, \\ \text{subject to} \quad & \sum_{j=1}^{n} x_j = 1, \\ & \quad x_j \geq 0, \qquad j = 1, \dots, n, \end{aligned}$$

where the values of a_j $(j = 1, \dots, n)$ are positive. Find a global maximum and show that it is unique. ∎

Exercise 4.8 (extensions of convexity, exam 040602) We have stressed that convexity is a crucial property of functions when analyzing optimization models in general and studying optimality conditions in particular. There are, however, certain properties of convex functions that are shared also by classes of non-convex functions. The purpose of this exercise is to relate the convex functions to two such classes of non-convex functions by means of some example properties.

Suppose that $S \subseteq \mathbb{R}^n$ and that $f : \mathbb{R}^n \to \mathbb{R}$ is continuous on S.

(a) Suppose further that f is in C^1 on S. We say that the function f is *pseudo-convex* on S if, for every $\boldsymbol{x}, \boldsymbol{y} \in S$,

$$\nabla f(\boldsymbol{x})^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{x}) \geq 0 \qquad \Longrightarrow \qquad f(\boldsymbol{y}) \geq f(\boldsymbol{x}).$$

Establish the following two statements: (1) if f is a convex function on S then f is pseudo-convex on S (that is, "convexity implies pseudo-convexity"); (2) the reverse statement ("pseudo-convexity implies convexity") is not true.

[*Hint:* On the statement (2) you may construct an explicit or graphical counter-example.]

[*Note:* Pseudo-convex functions were introduced by Mangasarian [Man65].]

(b) A well-known property of a differentiable convex function is its role in necessary and sufficient conditions for globally optimal solutions. Suppose now that S is convex. If f is a convex function on $\mathbb{R}^n$ which is in C^1 on S then Theorem 4.23 applies. Establish that the equivalence relation of this theorem still holds if the convexity of f on S is replaced by the pseudo-convexity of f on S.

(c) Let S be convex. We say that the function f is *quasi-convex* on S if its level sets are convex. In other words, f is quasi-convex on S if

$$\text{lev}_f^S(b) := \{\, \boldsymbol{x} \in S \mid f(\boldsymbol{x}) \leq b \,\}$$

is convex for every $b \in \mathbb{R}$.

Establish the following two statements for a function f which is in C^1 on S: (1) if f is convex function on S then f is quasi-convex on S (that is, "convexity implies quasi-convexity"); (2) the reverse statement ("quasi-convexity implies convexity") is not true.

[*Hint:* On the statement (2) you may construct an explicit or graphical counter-example.]

[*Note:* Pseudo-convex functions were introduced by De Finetti [DeF49].] ■

Exercise 4.9 (illustrations of fixed point results) (a) Let $S := \{\, x \in \mathbb{R} \mid x \geq 1 \,\}$ and $f(x) := x/2 + 1/x$. Show that f is a contraction and find the smallest value of α.

(b) In analysis, a usual condition for the convergence of an iteration $x_k = g(x_{k-1})$ is that g be continuously differentiable and

$$|g'(x)| \leq \alpha < 1.$$

Verify this by the use of Banach's Theorem 4.33(a).

(c) Show that a fixed-point iteration for calculating the square root of a given positive number c is

$$x_0 > 0; \qquad x_{k+1} = g(x_k) := \frac{1}{2}\left(x_k + \frac{c}{x_k}\right), \quad k = 0, 1, \ldots .$$

What condition do we get from (b)? Starting at $x_0 = 1$, calculate approximations x_1, x_2, x_3, x_4 of $\sqrt{2}$. ■

Optimality conditions

V

5.1 Relations between optimality conditions (OCs) and CQs at a glance

Optimality conditions are introduced as an attempt to construct an easily verifiable criterion that allows us to examine points in a feasible set, one after another, and classify them into optimal and non-optimal ones. Unfortunately, this is impossible in practice, and not only due to the fact that there are far too many feasible points, but also because it is impossible to construct such a universal criterion. It is usually possible to construct either practical (that is, computationally verifiable) conditions that admit some mistakes in the characterization, or perfect ones which are impossible to use in the computations. It is of course the first group that is of practical value for us, and it may further be classified into two distinct subgroups based on the type of mistakes allowed in the decision-making process. Namely, optimality conditions encountered in practice are divided into two classes, known as *necessary* and *sufficient* conditions.

Necessary conditions must be satisfied at every locally optimal point; on the other hand, we cannot guarantee that every point satisfying the necessary optimality conditions is indeed locally optimal. On the contrary, sufficient optimality conditions provide such guarantees; however, there are some locally optimal points that violate the optimality conditions. Arguably, it is much more important to be able to find a few candidates for local minima that can be further investigated by other means, than to eliminate some local (or even global) minima from the beginning. Therefore, this chapter is dedicated to the development of *necessary optimality conditions*. However, for *convex* optimization problems these conditions turn out to be *sufficient*.

Now, we can concentrate on what should be meant by easily verifiable conditions. A human being can immediately state whether a given point belongs to a simple set or not, by just glancing at a picture of it; for a numerical algorithm, a clear algebraic description of a set in terms of equalities and inequalities is vital. Therefore, we start our development with geometric optimality conditions (Section 5.3), to gain an understanding about the relationships between the gradient of the objective function and the feasible set that must hold at every local minimum point. Given a specific description of a feasible set in terms of inequalities, the geometric conditions immediately imply some relationships between the gradients of the objective functions and the constraints that are binding at the point under consideration (see Section 5.4); these conditions are known as the *Fritz–John optimality conditions*, and are rather weak (i.e., they can be satisfied by many points that have nothing in common with locally optimal points). However, if we assume an additional regularity of the system of inequalities and equalities that define our feasible set, then the geometric optimality conditions imply stronger conditions, known as the *Karush–Kuhn–Tucker optimality conditions* (see Section 5.5). The additional regularity assumptions are known under the name *constraint qualifications* (CQs), and they vary from very abstract and difficult to check, but enjoyed by many feasible sets (such as, e.g., Abadie's CQ, see Definition 5.23) to more specific, easily verifiable but also somewhat restrictive in many situations (such as the linear independence CQ (see Definition 5.41), or the Slater CQ, see Definition 5.38). In Section 5.8 we show that for convex problems the KKT conditions are sufficient for local, hence global, optimality.

The contents of this chapter are in principle summarized in the flow-chart in Figure 5.1. Various optimality conditions and constraint qualifications that are discussed in this chapter constitute the nodes of the flow-chart. Logical relationships between them are denoted with edges, and the direction of the arrow shows the direction of the logical implication; each implication is further labeled with the result that establishes it. We note that the KKT conditions "follow" from *both* geometric conditions and constraint qualifications satisfied at a given point; also, global optimality holds if *both* the KKT conditions are verified and the optimization problem is convex.

5.2 A note of caution

In this chapter we will discuss various *necessary* optimality conditions for a given point to be a local minimum to a nonlinear programming model. If the NLP is a convex program, any point satisfying these necessary

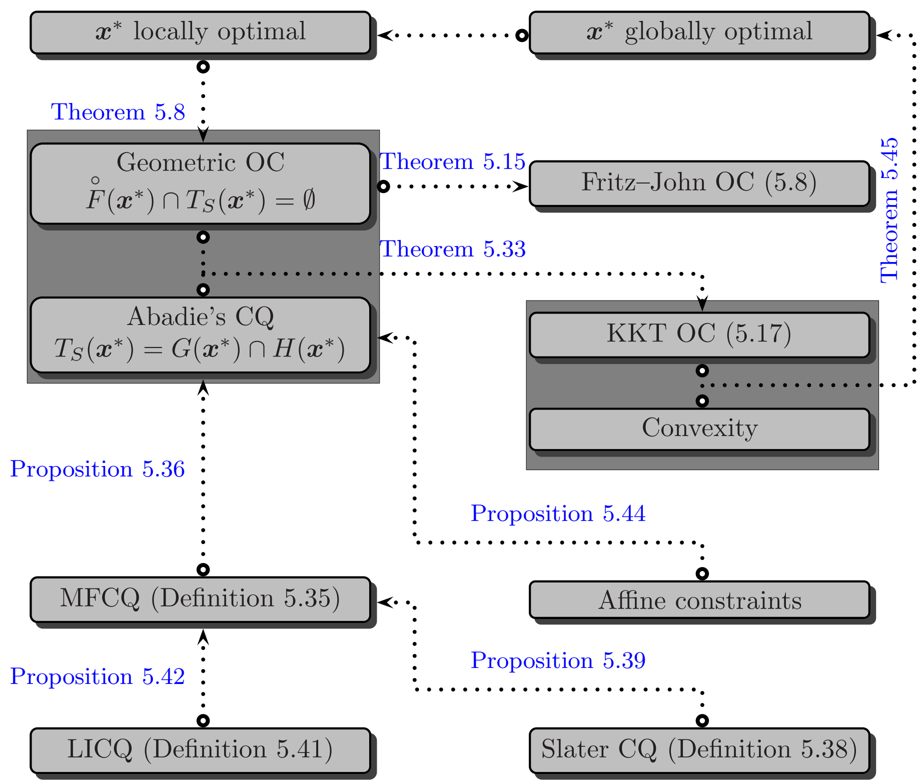


Figure 5.1: Relations between optimality conditions (OCs) and CQs at a glance.

optimality conditions is not only a local minimum, but actually a global minimum (see Section 5.8). Arguably, most NLP models that arise in real world applications tend to be nonconvex, and for such a problem, a point satisfying the necessary optimality conditions may not even be a local minimum. Algorithms for NLP are usually designed to converge to a point satisfying the necessary optimality conditions, and as mentioned earlier, one should not blindly accept such a point as an optimum solution to the problem without checking (e.g., using the second order necessary optimality conditions, see [BSS93, Section 4.4], or by means of some local search in the vicinity of the point) that it is at least better than all the other nearby points. Also, the system of necessary optimality conditions may have many solutions. Finding alternate solutions of this system, and selecting the best among them, usually leads to a good point to investigate further.

We will illustrate the importance of this with the story of US Air Force's controversial B-2 Stealth bomber program in the Reagan era of the 1980s. There were many design variables such as the various dimensions, the distribution of volume between the wing and the fuselage, flying speed, thrust, fuel consumption, drag, lift, air density, etc., that could be manipulated for obtaining the best range (i.e., the distance it can fly starting with full tanks, without refueling). The problem of maximizing the range subject to all the constraints was modeled as an NLP in a secret Air Force study going back to the 1940s. A solution to the necessary optimality conditions of this problem was found; it specified values for the design variables that put almost all of the total volume in the wing, leading to the *flying wing design* for the B-2 bomber. After spending billions of dollars, building test planes, etc., it was found that the design solution implemented works, but that its range was too low in comparison with other bomber designs being experimented subsequently in the US and abroad.

A careful review of the model was then carried out. The review indicated that all the formulas used, and the model itself, are perfectly valid. However, the model was a nonconvex NLP, and the review revealed a second solution to the system of necessary optimality conditions for it, besides the one found and implemented as a result of earlier studies. The second solution makes the wing volume much less than the total volume, and seems to maximize the range; while the first solution that is implemented for the B-2 bomber seems to actually minimize the range. (The second solution also looked like an airplane should, while the flying wing design was counter-intuitive.) In other words, the design implemented was the aerodynamically *worst* possible choice of configuration, leading to a very costly error.

For an account, see the research news item "Skeleton Alleged in the Stealth Bomber's Closet," *Science*, vol. 244, 12 May 1989 issue, pages 650–651.

5.3 Geometric optimality conditions

In this section we will discuss the optimality conditions for the following optimization problem [cf. (4.1)]:

$$\begin{aligned} &\text{minimize } f(\boldsymbol{x}), \\ &\text{subject to } \boldsymbol{x} \in S, \end{aligned} \tag{5.1}$$

where $S \subset \mathbb{R}^n$ is a nonempty closed set and $f : \mathbb{R}^n \to \mathbb{R}$ is a given differentiable function. Since we do not have any particular description of the feasible set S in terms of equality or inequality constraints, the optimality conditions will be based on purely geometrical ideas. Being quite general, the optimality conditions we will develop in this section are almost useless when it comes to computations, because they are also not very easy, even impossible, to verify for an optimization algorithm. Therefore, in the sections that follow, we will use an algebraic description of the set S and geometric optimality conditions to further develop classical Fritz–John and Karush–Kuhn–Tucker optimality conditions in the form of easily verifiable systems of equations and inequalities.

The basic idea behind the optimality conditions is that if the point $\boldsymbol{x}^* \in S$ is a point of local minimum for f over S, it should not be possible to draw a curve, or, more generally, a sequence of points, starting at the point $\boldsymbol{x}^*$ inside S, such that f decreases along it. Linearizing the objective function and the constraints along such curves, we eventually establish relationships between their gradients that are necessary to hold at points of local minima.

We start by defining the meaning of "possible to draw a curve starting at $\boldsymbol{x}^*$ inside S". Arguably, the simplest curves are the straight lines; the following definition gives exactly the set of lines that locally around $\boldsymbol{x}^*$ belong to S.

Definition 5.1 (cone of feasible directions) *Let $S \subset \mathbb{R}^n$ be a nonempty closed set. The* cone of feasible directions *for S at $\boldsymbol{x} \in \mathbb{R}^n$, known also as the radial cone, is defined as:*

$$R_S(\boldsymbol{x}) := \{\, \boldsymbol{p} \in \mathbb{R}^n \mid \exists\, \widetilde{\delta} > 0 \text{ such that } \boldsymbol{x} + \delta \boldsymbol{p} \in S, 0 \leq \delta \leq \widetilde{\delta} \,\}. \tag{5.2}$$

Thus, this is nothing else but the cone containing all feasible directions in the sense of Definition 4.19. ■

This cone is used in some optimization algorithms, but unfortunately it is too small to develop optimality conditions that are general enough. Therefore, we consider less intuitive, but bigger and more well-behaving sets (cf. Proposition 5.3 and the examples that follow).

Definition 5.2 (tangent cone) *Let $S \subset \mathbb{R}^n$ be a nonempty closed set. The* tangent cone *for S at $\boldsymbol{x} \in \mathbb{R}^n$ is defined as*

$$\begin{aligned} T_S(\boldsymbol{x}) := \{\, \boldsymbol{p} \in \mathbb{R}^n \mid \exists \{\boldsymbol{x}_k\} \subset S, \{\lambda_k\} \subset (0, \infty) : \lim_{k \to \infty} \boldsymbol{x}_k = \boldsymbol{x}, \\ \lim_{k \to \infty} \lambda_k (\boldsymbol{x}_k - \boldsymbol{x}) = \boldsymbol{p} \,\}. \end{aligned} \tag{5.3}$$

■

Thus, to construct a tangent cone we consider all the sequences $\{\boldsymbol{x}_k\}$ in S that converge to the given $\boldsymbol{x} \in \mathbb{R}^n$, and then calculate all the directions $\boldsymbol{p} \in \mathbb{R}^n$ that are tangential to the sequences at $\boldsymbol{x}$; such tangential vectors are described as the limits of $\{\lambda_k(\boldsymbol{x}_k - \boldsymbol{x})\}$ for arbitrary positive sequences $\{\boldsymbol{\lambda}_k\}$. Note that to generate a nonzero vector $\boldsymbol{p} \in T_S(\boldsymbol{x})$ the sequence $\{\boldsymbol{\lambda}_k\}$ must converge to $+\infty$.

While it is possible that $\operatorname{cl} R_S(\boldsymbol{x}) = T_S(\boldsymbol{x})$, or even that $R_S(\boldsymbol{x}) = T_S(\boldsymbol{x})$, in general we have only the following proposition, and examples that follow show that the two cones might be very different.

Proposition 5.3 (relationship between the radial and the tangent cones) *The tangent cone is a closed set, and the inclusion* $\operatorname{cl} R_S(\boldsymbol{x}) \subset T_S(\boldsymbol{x})$ *holds for every* $\boldsymbol{x} \in \mathbb{R}^n$.

Proof. Consider a sequence $\{\boldsymbol{p}_k\} \subset T_S(\boldsymbol{x})$, and assume that $\boldsymbol{p}_k \to \boldsymbol{p}$. Since every $\boldsymbol{p}_k \in T_S(\boldsymbol{x})$, there exists $\boldsymbol{x}_k \in S$ and $\lambda_k > 0$, such that $\|\boldsymbol{x}_k - \boldsymbol{x}\| < k^{-1}$ and $\|\lambda(\boldsymbol{x}_k - \boldsymbol{x}) - \boldsymbol{p}_k\| < k^{-1}$. Then, clearly, $\boldsymbol{x}_k \to \boldsymbol{x}$, and, by the triangle inequality, $\|\lambda(\boldsymbol{x}_k - \boldsymbol{x}) - \boldsymbol{p}\| \le \|\lambda(\boldsymbol{x}_k - \boldsymbol{x}) - \boldsymbol{p}_k\| + \|\boldsymbol{p}_k - \boldsymbol{p}\| \to 0$, which implies that $\boldsymbol{p} \in T_S(\boldsymbol{x})$ and thus the latter set is closed.

In view of the closedness of the tangent cone, it is enough to show the inclusion $R_S(\boldsymbol{x}) \subset T_S(\boldsymbol{x})$. Let $\boldsymbol{p} \in R_S(\boldsymbol{x})$. Then, for all large integers k it holds that $\boldsymbol{x} + k^{-1}\boldsymbol{p} \in S$, and, therefore, setting $\boldsymbol{x}_k = \boldsymbol{x} + k^{-1}\boldsymbol{p}$ and $\lambda_k = k$ we see that $\boldsymbol{p} \in T_S(\boldsymbol{x})$ as defined by Definition 5.2. ∎

Example 5.4 Let $S = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid -x_1 \le 0, (x_1 - 1)^2 + x_2^2 \le 1 \,\}$. Then, $R_S(\boldsymbol{0}^2) = \{\, \boldsymbol{p} \in \mathbb{R}^2 \mid p_1 > 0 \,\}$, and $T_S(\boldsymbol{0}^2) = \{\, \boldsymbol{p} \in \mathbb{R}^2 \mid p_1 \ge 0 \,\}$, i.e., $T_S(\boldsymbol{0}^2) = \operatorname{cl} R_S(\boldsymbol{0}^2)$ (see Figure 5.2). ∎

Example 5.5 (complementarity constraint) Let $S = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid -x_1 \le 0, -x_2 \le 0, x_1 x_2 \le 0 \,\}$. In this case, S is a (non-convex) cone, and $R_S(\boldsymbol{0}^2) = T_S(\boldsymbol{0}^2) = S$ (see Figure 5.3). ∎

Example 5.6 Let $S = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid -x_1^3 + x_2 \le 0, x_1^5 - x_2 \le 0, -x_2 \le 0 \,\}$. Then, $R_S(\boldsymbol{0}^2) = \emptyset$, $T_S(\boldsymbol{0}^2) = \{\, \boldsymbol{p} \in \mathbb{R}^2 \mid p_1 \ge 0, p_2 = 0 \,\}$ (see Figure 5.4). ∎

Example 5.7 Let $S = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid -x_2 \le 0, (x_1 - 1)^2 + x_2^2 = 1 \,\}$. Then, $R_S(\boldsymbol{0}^2) = \emptyset$, $T_S(\boldsymbol{0}^2) = \{\, \boldsymbol{p} \in \mathbb{R}^2 \mid p_1 = 0, p_2 \ge 0 \,\}$ (see Figure 5.5). ∎

We already know that f decreases along any descent direction (cf. Definition 4.14), and that for a vector $\boldsymbol{p} \in \mathbb{R}^n$ it is sufficient to verify the inequality $\nabla f(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{p} < 0$ to be a descent direction for f at $\boldsymbol{x}^* \in \mathbb{R}^n$ (see Proposition 4.15). Even though this condition is not necessary, it

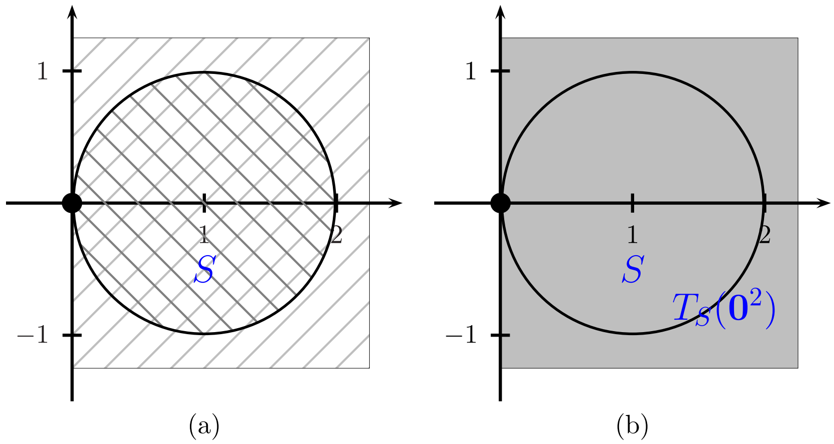


Figure 5.2: (a) The set S obtained as the intersection of two constraints; (b) the tangent cone $T_S(\mathbf{0}^2)$ (see Example 5.4).

is very easy to check in practice and therefore we will use it to develop optimality conditions. Therefore, it would be convenient to define a cone of such directions (which may be empty if $\nabla f(\boldsymbol{x}^*)$ happens to be $\mathbf{0}^n$):

$$\overset{\circ}{F}(\boldsymbol{x}^*) = \{\, \boldsymbol{p} \in \mathbb{R}^n \mid \nabla f(\boldsymbol{x}^*)^{\mathrm{T}} \boldsymbol{p} < 0 \,\}. \tag{5.4}$$

Now we have the necessary notation to state and prove the main theorem of this section.

Theorem 5.8 (geometric necessary optimality conditions) *Consider the optimization problem* (5.1)*. Then, for $\boldsymbol{x}^* \in S$ to be a local minimum of f over S it is necessary that $\overset{\circ}{F}(\boldsymbol{x}^*) \cap T_S(\boldsymbol{x}^*) = \emptyset$, where $\overset{\circ}{F}(\boldsymbol{x}^*)$ and $T_S(\boldsymbol{x}^*)$ are defined by* (5.4) *and Definition 5.2, respectively.*

Proof. Assume that $\boldsymbol{p} \in T_S(\boldsymbol{x}^*)$, i.e., $\exists \{\boldsymbol{x}_k\} \subset S$, and $\{\lambda_k\} \subset (0, \infty)$ such that $\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}^*$ and $\lim_{k\to\infty} \lambda_k(\boldsymbol{x}_k - \boldsymbol{x}^*) = \boldsymbol{p}$. Using the first order Taylor expansion (2.1) we get:

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) = \nabla f(\boldsymbol{x}^*)^{\mathrm{T}} (\boldsymbol{x}_k - \boldsymbol{x}^*) + o(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|) \geq 0,$$

where the last inequality holds for all enough large k by the local optimality of $\boldsymbol{x}^*$. Multiplying by $\lambda_k > 0$ and taking limit we get

$$\begin{aligned} 0 &\leq \lim_{k\to\infty} \left[\lambda_k \nabla f(\boldsymbol{x}^*)^{\mathrm{T}} (\boldsymbol{x}_k - \boldsymbol{x}^*) + \|\lambda_k(\boldsymbol{x}_k - \boldsymbol{x}^*)\| \frac{o(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|)}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|} \right] \\ &= \nabla f(\boldsymbol{x}^*)^{\mathrm{T}} \boldsymbol{p} + \|\boldsymbol{p}\| \cdot 0, \end{aligned}$$

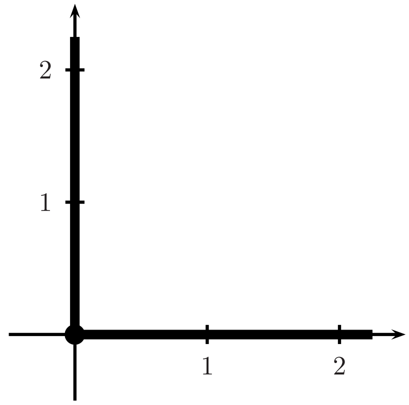


Figure 5.3: $S = R_S(\mathbf{0}^2) = T_S(\mathbf{0}^2)$ (see Example 5.5).

and thus $\boldsymbol{p} \notin \overset{\circ}{F}(\boldsymbol{x}^*)$. ∎

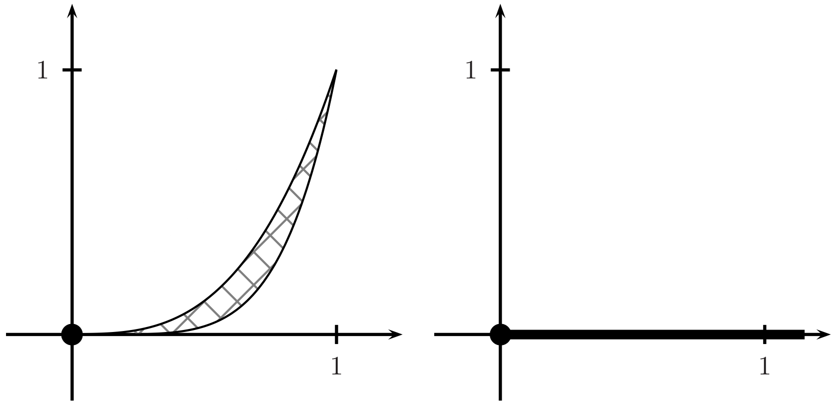
Combining Proposition 5.3 and Theorem 5.8 we get that for $\boldsymbol{x}^* \in S$ to be a local minimum of f over S it is necessary that $\overset{\circ}{F}(\boldsymbol{x}^*) \cap R_S(\boldsymbol{x}^*) = \emptyset$; but this statement is weaker than Theorem 5.8.

Example 5.9 Consider the differentiable (linear) function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(\boldsymbol{x}) = x_1$. Then, $\nabla f = (1, 0)^{\mathrm{T}}$, and $\overset{\circ}{F}(\mathbf{0}^2) = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 < 0 \,\}$. It is easy to see from geometric considerations that $\boldsymbol{x}^* = \mathbf{0}^2$ is a local (in fact, even global) minimum in either problem (5.1) with S given by Examples 5.4–5.7, and equally easy it is to check that the geometric necessary optimality condition $\overset{\circ}{F}(\mathbf{0}^2) \cap T_S(\mathbf{0}^2) = \emptyset$ is satisfied in all these examples (which is no surprise, in view of Theorem 5.8). ∎

5.4 The Fritz–John conditions

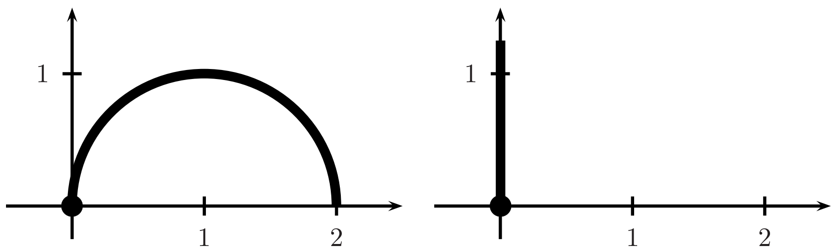
Theorem 5.8 gives a very elegant criterion for checking whether a given point $\boldsymbol{x}^* \in S$ is a candidate for a local minimum for the problem (5.1), but there is a catch: the set $T_S(\boldsymbol{x}^*)$ is close to impossible to compute for general sets S! Therefore, in this section we will use the algebraic characterization of the set S to compute other cones that we hope could approximate $T_S(\boldsymbol{x}^*)$ in many practical situations.

Namely, we assume that the set S is defined as the solution set of a system of differentiable inequality constraints defined by the functions



(a) (b)

Figure 5.4: (a) The set S; (b) the tangent cone $T_S(\mathbf{0}^2)$ (see Example 5.6).



(a) (b)

Figure 5.5: (a) The set S; (b) the tangent cone $T_S(\mathbf{0}^2)$ (see Example 5.7).

$g_i \in C^1(\mathbb{R}^n)$, $i = 1, \dots, m$:

$$S = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid g_i(\boldsymbol{x}) \leq 0, \quad i = 1, \dots, m \,\}. \tag{5.5}$$

We can always assume this structure, because any equality constraint $h(\boldsymbol{x}) = 0$ may be written in the form $h(\boldsymbol{x}) \leq 0 \wedge -h(\boldsymbol{x}) \leq 0$. Treating the equality constraints in this way we obtain the Fritz–John conditions, that however are somewhat too weak to be practical; on the positive side, it significantly simplifies the notation and does not affect the development of the KKT-conditions. Therefore, we keep this assumption for some time, and state the KKT system that specifically distinguishes between the inequality and equality constraints in Section 5.6. We will use the symbol $\mathcal{I}(\boldsymbol{x})$ to denote the index set of binding, or active, *inequality* constraints at $\boldsymbol{x} \in \mathbb{R}^n$ (see Definition 4.20), and $|\mathcal{I}(\boldsymbol{x})|$ to denote the cardinality of this set, i.e., the number of active inequality constraints at $\boldsymbol{x} \in \mathbb{R}^n$.

In order to compute approximations to the tangent cone $T_S(\boldsymbol{x})$, similarly to Example 4.21 we consider cones associated with the active constraints at a given point:

$$\overset{\circ}{G}(\boldsymbol{x}) = \{\, \boldsymbol{p} \in \mathbb{R}^n \mid \nabla g_i(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{p} < 0, i \in \mathcal{I}(\boldsymbol{x}) \,\}, \tag{5.6}$$

and

$$G(\boldsymbol{x}) = \{\, \boldsymbol{p} \in \mathbb{R}^n \mid \nabla g_i(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{p} \le 0, i \in \mathcal{I}(\boldsymbol{x}) \,\}. \tag{5.7}$$

The following proposition verifies that $\overset{\circ}{G}(\boldsymbol{x})$ is an "inner approximation" for $R_S(\boldsymbol{x})$ (and, therefore, for $T_S(\boldsymbol{x})$ as well, see Proposition 5.3), and $G(\boldsymbol{x})$ is an outer approximation for $T_S(\boldsymbol{x})$.

Lemma 5.10 *For every $\boldsymbol{x}^* \in \mathbb{R}^n$ it holds that $\overset{\circ}{G}(\boldsymbol{x}^*) \subset R_S(\boldsymbol{x}^*)$, and $T_S(\boldsymbol{x}^*) \subset G(\boldsymbol{x}^*)$.*

Proof. Let $\boldsymbol{p} \in \overset{\circ}{G}(\boldsymbol{x}^*)$. For every $i \notin \mathcal{I}(\boldsymbol{x}^*)$ the function g_i is continuous and $g_i(\boldsymbol{x}^*) < 0$; therefore $g_i(\boldsymbol{x}^* + \delta \boldsymbol{p}) < 0$ for all small $\delta > 0$. Moreover, by Proposition 4.15, $\boldsymbol{p}$ is a direction of descent for every g_i at $\boldsymbol{x}^*$, $i \in \mathcal{I}(\boldsymbol{x}^*)$, which means that $g_i(\boldsymbol{x}^* + \delta \boldsymbol{p}) < g_i(\boldsymbol{x}^*) = 0$ for all such i and all small $\delta > 0$. Thus, $\boldsymbol{p} \in R_S(\boldsymbol{x})$, and, hence, $\overset{\circ}{G}(\boldsymbol{x}^*) \subset R_S(\boldsymbol{x})$.

Now, let $\boldsymbol{p} \in T_S(\boldsymbol{x})$, i.e., $\exists \{\boldsymbol{x}_k\} \subset S$, and $\{\lambda_k\} \subset (0, \infty)$ such that $\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}^*$ and $\lim_{k\to\infty} \lambda_k(\boldsymbol{x}_k - \boldsymbol{x}^*) = \boldsymbol{p}$. Exactly as in the proof of Theorem 5.8, we use the first order Taylor expansion (2.1) of the functions g_i, $i \in \mathcal{I}(\boldsymbol{x}^*)$, to get:

$$0 \ge g_i(\boldsymbol{x}_k) = g_i(\boldsymbol{x}_k) - g_i(\boldsymbol{x}^*) = \nabla g_i(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x}_k - \boldsymbol{x}^*) + o(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|),$$

where the first inequality is by the feasibility of $\boldsymbol{x}_k$. Multiplying by $\lambda_k > 0$ and taking limit we get, for $i \in \mathcal{I}(\boldsymbol{x}^*)$,

$$\begin{aligned} 0 &\ge \lim_{k\to\infty} \left[\lambda_k \nabla g_i(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x}_k - \boldsymbol{x}^*) + \|\lambda_k(\boldsymbol{x}_k - \boldsymbol{x}^*)\| \frac{o(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|)}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|} \right] \\ &= \nabla g_i(\boldsymbol{x}^*)^{\mathrm{T}} \boldsymbol{p} + \|\boldsymbol{p}\| \cdot 0, \end{aligned}$$

and thus $\boldsymbol{p} \in G(\boldsymbol{x}^*)$. ■

Example 5.11 (Example 5.4 continued) In this example the set S is defined by the two inequality constraints $g_1(\boldsymbol{x}) = -x_1 \le 0$ and $g_2(\boldsymbol{x}) =$

$(x_1-1)^2+x_2^2-1 \leq 0$. Let us calculate $\overset{\circ}{G}(\mathbf{0}^2)$ and $G(\mathbf{0}^2)$. Both constraints are satisfied with equality at the given point, so that $\mathcal{I}(\boldsymbol{x}) = \{1,2\}$. Then, $\nabla g_1(\mathbf{0}^2) = (-1,0)^{\mathrm{T}}$, $\nabla g_2(\mathbf{0}^2) = (-2,0)^{\mathrm{T}}$, and thus $\overset{\circ}{G}(\mathbf{0}^2) = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 > 0 \,\} = R_S(\mathbf{0}^2)$, $G(\mathbf{0}^2) = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 \geq 0 \,\} = T_S(\mathbf{0}^2)$ in this case. ■

Example 5.12 (Example 5.5 continued) S is defined by the three inequality constraints $g_1(\boldsymbol{x}) = -x_1 \leq 0$, $g_2(\boldsymbol{x}) = -x_2 \leq 0$, $g_3(\boldsymbol{x}) = x_1x_2 \leq 0$, which are all binding at $\boldsymbol{x}^* = \mathbf{0}^2$; $\nabla g_1(\mathbf{0}^2) = (-1,0)^{\mathrm{T}}$, $\nabla g_2(\mathbf{0}^2) = (0,-1)^{\mathrm{T}}$, and $\nabla g_3(\mathbf{0}^2) = (0,0)^{\mathrm{T}}$. Therefore, $\overset{\circ}{G}(\mathbf{0}^2) = \emptyset \subsetneq R_S(\mathbf{0}^2)$, and $G(\mathbf{0}^2) = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 \geq 0, x_2 \geq 0 \,\} \supsetneq T_S(\mathbf{0}^2)$. ■

Example 5.13 (Example 5.6 continued) S is defined by the three inequality constraints $g_1(\boldsymbol{x}) = -x_1^3 + x_2 \leq 0$, $g_2(\boldsymbol{x}) = x_1^5 - x_2 \leq 0$, $g_3(\boldsymbol{x}) = -x_2 \leq 0$, which are all binding at $\boldsymbol{x}^* = \mathbf{0}^2$; $\nabla g_1(\mathbf{0}^2) = (0,1)^{\mathrm{T}}$, $\nabla g_2(\mathbf{0}^2) = (0,-1)^{\mathrm{T}}$, and $\nabla g_3(\mathbf{0}^2) = (0,-1)^{\mathrm{T}}$. Therefore, $\overset{\circ}{G}(\mathbf{0}^2) = \emptyset = R_S(\mathbf{0}^2)$, and $G(\mathbf{0}^2) = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_2 = 0 \,\} \supsetneq T_S(\mathbf{0}^2)$. ■

Example 5.14 (Example 5.7 continued) In this example, the set S is defined by the inequality constraint $g_1(\boldsymbol{x}) = -x_2 \leq 0$, and one equality constraint $h_1(\boldsymbol{x}) = (x_1-1)^2 + x_2^2 - 1 = 0$; we split the latter into two inequality constraints $g_2(\boldsymbol{x}) = h_1(\boldsymbol{x}) \leq 0$, and $g_3(\boldsymbol{x}) = -h_1(\boldsymbol{x}) \leq 0$. Thus, we end up with three binding inequality constraints at $\boldsymbol{x}^* = \mathbf{0}^2$; $\nabla g_1(\mathbf{0}^2) = (0,-1)^{\mathrm{T}}$, $\nabla g_2(\mathbf{0}^2) = (-2,0)^{\mathrm{T}}$, and $\nabla g_3(\mathbf{0}^2) = (2,0)^{\mathrm{T}}$. Therefore, $\overset{\circ}{G}(\mathbf{0}^2) = \emptyset = R_S(\mathbf{0}^2)$, and $G(\mathbf{0}^2) = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid x_1 = 0, x_2 \geq 0 \,\} = T_S(\mathbf{0}^2)$. ■

Now we are ready to establish the Fritz–John optimality conditions.

Theorem 5.15 (Fritz–John necessary optimality conditions) *Let the set S be defined by (5.5). Then, for $\boldsymbol{x}^* \in S$ to be a local minimum of f over S it is necessary that there exist multipliers $\mu_0 \in \mathbb{R}$, $\boldsymbol{\mu} \in \mathbb{R}^m$, such that*

$$\mu_0 \nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \mu_i \nabla g_i(\boldsymbol{x}^*) = \mathbf{0}^n, \tag{5.8a}$$

$$\mu_i g_i(\boldsymbol{x}^*) = 0, \quad i = 1, \dots, m, \tag{5.8b}$$

$$\mu_0, \mu_i \geq 0, \quad i = 1, \dots, m, \tag{5.8c}$$

$$(\mu_0, \boldsymbol{\mu}^{\mathrm{T}})^{\mathrm{T}} \neq \mathbf{0}^{m+1}. \tag{5.8d}$$

Proof. Combining the results of Lemma 5.10 with the geometric optimality conditions provided by Theorem 5.8, we conclude that there is no direction $\boldsymbol{p} \in \mathbb{R}^n$ such that $\nabla f(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{p} < 0$ and $\nabla g_i(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{p} < 0$, $i \in \mathcal{I}(\boldsymbol{x}^*)$. Define the matrix $\boldsymbol{A}$ with columns $\nabla f(\boldsymbol{x}^*)$, $\nabla g_i(\boldsymbol{x}^*)$, $i \in \mathcal{I}(\boldsymbol{x}^*)$; then the system $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{p} < \boldsymbol{0}^{1+|\mathcal{I}(\boldsymbol{x}^*)|}$ is unsolvable. By Farkas' Lemma (cf. Theorem 3.30) there exists a nonzero vector $\boldsymbol{\lambda} \in \mathbb{R}^{1+|\mathcal{I}(\boldsymbol{x}^*)|}$ such that $\boldsymbol{\lambda} \geq \boldsymbol{0}^{1+|\mathcal{I}(\boldsymbol{x}^*)|}$ and $\boldsymbol{A\lambda} = \boldsymbol{0}^n$. Now, let $(\mu_0, \boldsymbol{\mu}^{\mathrm{T}}_{\mathcal{I}(\boldsymbol{x}^*)})^{\mathrm{T}} = \boldsymbol{\lambda}$, and set $\mu_i = 0$ for $i \notin \mathcal{I}(\boldsymbol{x}^*)$. It is an easy exercise now to verify that so defined μ_0 and $\boldsymbol{\mu}$ satisfy the conditions (5.8). ■

Remark 5.16 (terminology) The solutions $(\mu_0, \boldsymbol{\mu})$ to the system (5.8) are known as *Lagrange multipliers* (or just *multipliers*) associated with a given candidate $\boldsymbol{x}^* \in \mathbb{R}^n$ for a local minimum. Note, that every multiplier (except μ_0) corresponds to some constraint in the algebraic representation of S. The conditions (5.8a) and (5.8c) are known as the *dual feasibility*, and (5.8b) as the *complementarity constraints*, respectively; this terminology will become more clear in Chapter 6. Owing to the complementarity constraints, the multipliers μ_i corresponding to *inactive* inequality constraints $i \notin \mathcal{I}(\boldsymbol{x}^*)$ must be zero. In general, the Lagrange multiplier μ_i bears the important information about how *sensitive* a particular local minimum is with respect to small changes in the constraint g_i. ■

In the following examples, as before, we assume that $f(\boldsymbol{x}) = x_1$, so that $\nabla f = (1,0)^{\mathrm{T}}$ and $\boldsymbol{x}^* = \boldsymbol{0}^2$ is the point of local minimum.

Example 5.17 (Example 5.4 continued) The Fritz–John system (5.8) at the point $\boldsymbol{x}^* = \boldsymbol{0}^2$ in this case reduces to:

$$\begin{cases} \mu_0 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} -1 & -2 \\ 0 & 0 \end{pmatrix} \boldsymbol{\mu} = \boldsymbol{0}^2, \\ \qquad\qquad (\mu_0, \boldsymbol{\mu}^{\mathrm{T}})^{\mathrm{T}} \gneqq \boldsymbol{0}^3, \end{cases}$$

where $\boldsymbol{\mu} \in \mathbb{R}^2$ is a vector of Lagrange multipliers for the inequality constraints. We do not write the complementarity constraints (5.8b), because in our case all three constraints are active, and therefore the equation (5.8b) is automatically satisfied for all $\boldsymbol{\mu}$. The solutions to this system are described as pairs $(\mu_0, \boldsymbol{\mu})$, with $\boldsymbol{\mu} = (\mu_1, 2^{-1}(\mu_0 - \mu_1))^{\mathrm{T}}$, for every $\mu_0 > 0$, $0 \leq \mu_1 \leq \mu_0$. There are infinitely many Lagrange multipliers, that even form an unbounded set, but μ_0 must always be positive. ■

Example 5.18 (Example 5.5 continued) Similarly to the previous example, the Fritz–John system (5.8) at the point $\boldsymbol{x}^* = \mathbf{0}^2$ in this case reduces to:

$$\begin{cases} \mu_0 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix} \boldsymbol{\mu} = \mathbf{0}^2, \\ \qquad\qquad\qquad (\mu_0, \boldsymbol{\mu}^{\mathrm{T}})^{\mathrm{T}} \gneqq \mathbf{0}^4, \end{cases}$$

where $\boldsymbol{\mu} \in \mathbb{R}^3$ is a vector of Lagrange multipliers for the inequality constraints. The solution to the Fritz–John system is every pair $(\mu_0, \boldsymbol{\mu})$ with $\boldsymbol{\mu} = (\mu_0, 0, \mu_3)^{\mathrm{T}}$ for every $\mu_0 \geq 0$, $\mu_3 \geq 0$ such that either of them is strictly bigger than zero. That is, there are infinitely many Lagrange multipliers, that even form an unbounded set, and it is possible for μ_0 to assume the value zero. ■

Example 5.19 (Example 5.6 continued) The Fritz–John system (5.8) at the point $\boldsymbol{x}^* = \mathbf{0}^2$ in this case reduces to:

$$\begin{cases} \mu_0 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 1 & -1 & -1 \end{pmatrix} \boldsymbol{\mu} = \mathbf{0}^2, \\ \qquad\qquad\qquad (\mu_0, \boldsymbol{\mu}^{\mathrm{T}})^{\mathrm{T}} \gneqq \mathbf{0}^4, \end{cases}$$

where $\boldsymbol{\mu} \in \mathbb{R}^3$ is a vector of Lagrange multipliers for the inequality constraints. Thus, $\mu_0 = 0$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_1 - \mu_2)^{\mathrm{T}}$ for every $\mu_1 > 0$, $0 \leq \mu_2 \leq \mu_1$. That is, there are infinitely many Lagrange multipliers, that even form an unbounded set, and μ_0 must assume the value zero. ■

Example 5.20 (Example 5.7 continued) The Fritz–John system (5.8) at the point $\boldsymbol{x}^* = \mathbf{0}^2$ in this case reduces to:

$$\begin{cases} \mu_0 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & 0 \end{pmatrix} \boldsymbol{\mu} = \mathbf{0}^2, \\ \qquad\qquad\qquad (\mu_0, \boldsymbol{\mu}^{\mathrm{T}})^{\mathrm{T}} \gneqq \mathbf{0}^4, \end{cases}$$

where $\boldsymbol{\mu} \in \mathbb{R}^3$ is a vector of Lagrange multipliers for the inequality constraints. The solution to the Fritz–John system is every pair $(\mu_0, \boldsymbol{\mu})$ with $\boldsymbol{\mu} = (0, \mu_2, \mu_2 - 2^{-1}\mu_0)^{\mathrm{T}}$ for every $\mu_2 > 0$, $0 \leq \mu_0 \leq 2\mu_2$. That is, there are infinitely many Lagrange multipliers, that even form an unbounded set, and it is possible for μ_0 to assume the value zero. ■

The fact that μ_0 may be zero in the system (5.8) essentially means that the objective function f plays no role in the optimality conditions. This is of course a rather unexpected and unwanted situation, and the rest of the chapter is in principle dedicated to describing how one can avoid it.

Since the cone of feasible directions $R_S(\boldsymbol{x})$ may be a bad approximation of the tangent cone $T_S(\boldsymbol{x})$, so may $\overset{\circ}{G}(\boldsymbol{x})$ owing to Lemma 5.10. Therefore, in the most general case we cannot improve on the conditions (5.8); however, it is possible to improve upon (5.8) if we assume that the set S is "regular" in some sense, i.e., that either $\overset{\circ}{G}(\boldsymbol{x})$ or $G(\boldsymbol{x})$ is a tight enough approximation of $T_S(\boldsymbol{x})$. Requirements of this type are called *constraint qualifications*, and they will be discussed in more detail in Section 5.7. However, to get a feeling of what can be achieved with a regular constraint sets S, we show that the multiplier μ_0 in the system (5.8) cannot vanish (i.e., KKT conditions hold, see Section 5.5) if the constraint qualification $\overset{\circ}{G}(\boldsymbol{x}^*) \neq \emptyset$ holds (which is quite a restrictive one, in view of Example 5.22; however, see the similar, but much weaker, assumption MFCQ in Section 5.7 dedicated to constraint qualifications).

Proposition 5.21 (KKT optimality conditions – preview) *Assume the conditions of Theorem 5.8, and assume that* $\overset{\circ}{G}(\boldsymbol{x}^*) \neq \emptyset$*. Then, the multiplier* μ_0 *in* (5.8) *cannot be zero; dividing all equations by* μ_0 *we may assume that it equals one.*

Proof. Assume that $\mu_0 = 0$ in (5.8), and define the matrix $\boldsymbol{A}$ with columns $\nabla g_i(\boldsymbol{x}^*)$, $i \in \mathcal{I}(\boldsymbol{x}^*)$. Since $\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{0}^n$, $\boldsymbol{\mu} \geq \boldsymbol{0}^{|\mathcal{I}(\boldsymbol{x}^*)|}$, and $\boldsymbol{\mu} \neq \boldsymbol{0}^{|\mathcal{I}(\boldsymbol{x}^*)|}$, the system $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{p} < \boldsymbol{0}^{|\mathcal{I}(\boldsymbol{x}^*)|}$ is unsolvable (see Farkas' Lemma, Theorem 3.30), i.e., $\overset{\circ}{G}(\boldsymbol{x}^*) = \emptyset$. ∎

Example 5.22 Out of the four Examples 5.4–5.7, only the first one verifies the condition $\overset{\circ}{G}(\boldsymbol{x}^*) \neq \emptyset$ assumed in Proposition 5.21, while as we see later (and as Examples 5.17–5.20 may suggest), three out of four problems admit solutions to the corresponding KKT systems. ∎

5.5 The Karush–Kuhn–Tucker conditions

In this section we develop the famous and classic Karush–Kuhn–Tucker optimality conditions for constrained optimization problems with inequality constraints, which are essentially the Fritz–John conditions (5.8) with the additional requirement $\mu_0 \neq 0$. We establish these conditions as before, for inequality constrained problems (5.5) (which we do without any loss of generality or sharpness of the theory), and then discuss the possible modifications of the conditions if one wants to specifically distinguish between equality and inequality constraints in Section 5.6. Abadie's *constraint qualification* (see Definition 5.23) which we impose is very abstract and extremely general (this is *almost* the weakest condition one can require); of course it is impossible to check it when it comes to practical problems. Therefore, in Section 5.7 we list some computationally verifiable assumptions that all imply Abadie's constraint qualification.

We start with a formal definition.

Definition 5.23 (Abadie's constraint qualification) *We say that at the point $\boldsymbol{x} \in S$ Abadie's constraint qualification holds if $T_S(\boldsymbol{x}) = G(\boldsymbol{x})$, where $T_S(\boldsymbol{x})$ is defined by Definition 5.2 and $G(\boldsymbol{x})$ by (5.7).* ■

Example 5.24 Out of the four Examples 5.4–5.7, the first and the last satisfy Abadie's constraint qualification (see Examples 5.11–5.14). ■

Then, we are ready to prove the main theorem in this chapter.

Theorem 5.25 (Karush–Kuhn–Tucker optimality conditions) *Assume that at a given point $\boldsymbol{x}^* \in S$ Abadie's constraint qualification holds. Then, for $\boldsymbol{x}^* \in S$ to be a local minimum of f over S it is necessary that there exists $\boldsymbol{\mu} \in \mathbb{R}^m$, such that*

$$\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \mu_i \nabla g_i(\boldsymbol{x}^*) = \mathbf{0}^n, \tag{5.9a}$$

$$\mu_i g_i(\boldsymbol{x}^*) = 0, \quad i = 1, \dots, m, \tag{5.9b}$$

$$\boldsymbol{\mu} \geq \mathbf{0}^m, \tag{5.9c}$$

the system bearing the name of Karush–Kuhn–Tucker optimality conditions.

Proof. By Theorem 5.8 we have that $\overset{\circ}{F}(\boldsymbol{x}^*) \cap T_S(\boldsymbol{x}^*) = \emptyset$, which due to our assumptions implies that $\overset{\circ}{F}(\boldsymbol{x}^*) \cap G(\boldsymbol{x}^*) = \emptyset$.

As in the proof of Theorem 5.15, construct a matrix $\boldsymbol{A}$ with columns $\nabla g_i(\boldsymbol{x}^*)$, $i \in \mathcal{I}(\boldsymbol{x}^*)$. Then, the system $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{p} \leq \mathbf{0}^{|\mathcal{I}(\boldsymbol{x}^*)|}$ and $-\nabla f(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{p} > 0$ has no solutions. By Farkas' Lemma (cf. Theorem 3.30), the system $\boldsymbol{A}\boldsymbol{\xi} = -\nabla f(\boldsymbol{x}^*)$, $\boldsymbol{\xi} \geq \mathbf{0}^{|\mathcal{I}(\boldsymbol{x}^*)|}$ has a solution. Thus, define $\boldsymbol{\mu}_{\mathcal{I}(\boldsymbol{x}^*)} = \boldsymbol{\xi}$, and $\mu_i = 0$, for $i \notin \mathcal{I}(\boldsymbol{x}^*)$. Then, then so defined $\boldsymbol{\mu}$ verifies the KKT conditions (5.9). ■

Remark 5.26 (terminology) Similar to the case of the Fritz–John necessary optimality conditions, the solutions $\boldsymbol{\mu}$ to the system (5.9) are known as *Lagrange multipliers* (or just *multipliers*) associated with a given candidate $\boldsymbol{x}^* \in \mathbb{R}^n$ for a local minimum. The conditions (5.9a) and (5.9c) are known as the *dual feasibility*, and (5.9b) as the *complementarity constraints*, respectively; this terminology will become more clear in Chapter 6. Owing to the complementarity constraints, the multipliers μ_i corresponding to *inactive* inequality constraints $i \notin \mathcal{I}(\boldsymbol{x}^*)$ must be zero. In general, the Lagrange multiplier μ_i bears the important information about how *sensitive* a particular local minimum is with respect to small changes in the constraint g_i. ∎

Remark 5.27 (geometric interpretation) The system of equations and inequalities defining (5.9) can (and should) be interpreted geometrically as $-\nabla f(\boldsymbol{x}^*) \in N_S(\boldsymbol{x}^*)$ (see Figure 5.6), the latter cone being the *normal cone* to S at $\boldsymbol{x}^* \in S$ (see Definition 4.24); according to the figure, the normal cone to S at $\boldsymbol{x}^*$ is furthermore spanned by the gradients of the active constraints at $\boldsymbol{x}^*$.

Notice the specific roles played by the different parts of the system (5.9) in this respect: the complementarity conditions (5.9b) force μ_i to be equal to 0 for the inactive constraints, whence the summation in the left-hand side of the linear system (5.9a) involves the active constraints only. Further, the sign conditions in (5.9c) ensures that each vector $\mu_i \nabla g_i(\boldsymbol{x}^*)$, $i \in \mathcal{I}(\boldsymbol{x}^*)$, is an *outward normal* to S at $\boldsymbol{x}^*$. ∎

Remark 5.28 Note that in the unconstrained case the KKT system (5.9) reduces to the single requirement $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}^n$, which we have already encountered in Theorem 4.13.

It is possible to further develop the KKT-theory (with some technical complications) for two times differentiable functions as it has been done for the unconstrained case in Theorem 4.16. We refer the interested reader to [BSS93, Section 4.4]. ∎

Example 5.29 (Example 5.4 continued) In this example Abadie's constraint qualification is fulfilled, therefore the KKT-system must be solvable. Indeed, the system

$$\begin{cases} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} -1 & -2 \\ 0 & 0 \end{pmatrix} \boldsymbol{\mu} = \boldsymbol{0}^2, \\ \qquad\qquad\qquad\qquad \boldsymbol{\mu} \geq \boldsymbol{0}^2, \end{cases}$$

possesses solutions $\boldsymbol{\mu} = (\mu_1, 2^{-1}(1-\mu_1))^{\mathrm{T}}$ for every $0 \leq \mu_1 \leq 1$. Therefore, there are infinitely many multipliers, that all belong to a bounded set. ∎

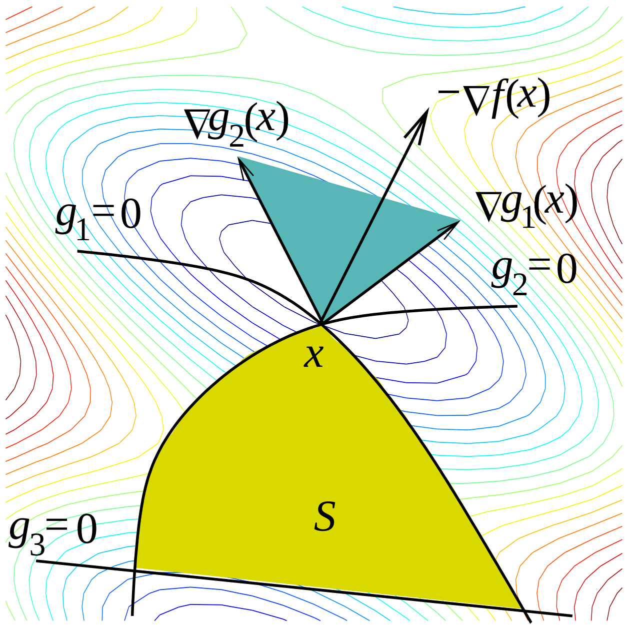


Figure 5.6: Geometrical interpretation of the KKT system.

Example 5.30 (Example 5.5 continued) This is one of the rare cases when Abadie's constraint qualification is violated, and nevertheless the KKT system happens to be solvable:

$$\begin{cases} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix} \boldsymbol{\mu} = \mathbf{0}^2, \\ \boldsymbol{\mu} \geq \mathbf{0}^3, \end{cases}$$

admits solutions $\boldsymbol{\mu} = (1, 0, \mu_3)^{\mathrm{T}}$ for every $\mu_3 \geq 0$. That is, the set of Lagrange multipliers is still unbounded in this case. ■

Example 5.31 (Example 5.6 continued) Since, for this example, in the Fritz–John system the multiplier μ_0 is necessarily zero, the KKT system admits no solutions:

$$\begin{cases} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 1 & -1 & -1 \end{pmatrix} \boldsymbol{\mu} = \mathbf{0}^2, \\ \boldsymbol{\mu} \geq \mathbf{0}^3, \end{cases}$$

is clearly inconsistent. In this example the very basic Abadie's constraint qualifications is violated, and therefore the lack of KKT multipliers should not be a big surprise. ■

Example 5.32 (Example 5.7 continued) This example also satisfies Abadie's constraint qualification, therefore the KKT-system is necessarily solvable:

$$\begin{cases} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & 0 \end{pmatrix} \boldsymbol{\mu} = \mathbf{0}^2, \\ \boldsymbol{\mu} \geq \mathbf{0}^3, \end{cases}$$

admits the solutions $\boldsymbol{\mu} = (0, \mu_2, \mu_2 - 2^{-1})^{\mathrm{T}}$, for all $\mu_2 \geq 2^{-1}$. The set of Lagrange multipliers is unbounded in this case, but this is because we have split the original equality constraint into two inequalities. In Section 5.6 we formulate the KKT-system that keeps the original equality-representation of the set, and thus reduce the number of multipliers to just one! ■

5.6 Proper treatment of equality constraints

Now we consider both inequality and equality constraints, i.e., we assume that the feasible set S is given by

$$S = \{ \boldsymbol{x} \in \mathbb{R}^n \mid g_i(\boldsymbol{x}) \leq 0, \quad i = 1, \dots, m, \\ h_j(\boldsymbol{x}) = 0, \quad j = 1, \dots, \ell \}, \tag{5.10}$$

instead of (5.5), where $g_i \in C^1(\mathbb{R}^n)$, $i = 1, \dots, m$, and $h_j \in C^1(\mathbb{R}^n)$, $j = 1, \dots, \ell$. As it was done in Section 5.4, we write S using only inequality constraints, by defining the functions $\widetilde{g}_i \in C^1(\mathbb{R}^n)$, $i = 1, \dots, m + 2\ell$, via:

$$\widetilde{g}_i = \begin{cases} g_i, & i = 1, \dots, m, \\ h_{i-m}, & i = m+1, \dots, m+\ell, \\ -h_{i-m-\ell}, & i = m+\ell+1, \dots, m+2\ell, \end{cases} \tag{5.11}$$

so that

$$S = \{ \boldsymbol{x} \in \mathbb{R}^n \mid \widetilde{g}_i(\boldsymbol{x}) \leq 0, \quad i = 1, \dots, m + 2\ell \}. \tag{5.12}$$

Now, let $\widetilde{G}(\boldsymbol{x})$ be defined by (5.7) for the inequality representation (5.12) of S. We will use the old notation $G(\boldsymbol{x})$ for the cone defined only by the gradients of the functions defining the *inequality* constraints active at $\boldsymbol{x}$ in the representation (5.10), and in addition define the null space of the matrix defined by the gradients of the functions defining the *equality* constraints:

$$H(\boldsymbol{x}) = \{ \boldsymbol{p} \in \mathbb{R}^n \mid \nabla h_i(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{p} = 0, \quad i = 1, \dots, \ell \}.. \tag{5.13}$$

Since all inequality constraint functions $\widetilde{g}_i$, $i = m+1, \dots, m+2\ell$, are necessarily active at every $\boldsymbol{x} \in S$, it holds that

$$\widetilde{G}(\boldsymbol{x}) = G(\boldsymbol{x}) \cap H(\boldsymbol{x}), \tag{5.14}$$

and thus Abadie's constraint qualification (see Definition 5.23) for the set (5.10) may be equivalently written as

$$T_S(\boldsymbol{x}) = G(\boldsymbol{x}) \cap H(\boldsymbol{x}). \tag{5.15}$$

Assuming that the latter constraint qualification holds we can write the KKT-system (5.9) for $\boldsymbol{x}^* \in S$, corresponding to the inequality representation (5.12) (see Theorem 5.25):

$$\sum_{i=1}^{m} \mu_i \nabla g_i(\boldsymbol{x}^*) + \sum_{i=m+1}^{m+\ell} \mu_i \nabla h_{i-m}(\boldsymbol{x}^*) - \sum_{i=m+\ell+1}^{m+2\ell} \mu_i \nabla h_{i-m-\ell}(\boldsymbol{x}^*)$$

$$+\nabla f(\boldsymbol{x}^*) = \mathbf{0}^n, \tag{5.16a}$$

$$\mu_i g_i(\boldsymbol{x}^*) = 0, \quad i = 1, \dots, m, \tag{5.16b}$$

$$\mu_i h_{i-m}(\boldsymbol{x}^*) = 0, \quad i = m+1, \dots, m+\ell, \tag{5.16c}$$

$$-\mu_i h_{i-m-\ell}(\boldsymbol{x}^*) = 0, \quad i = m+\ell+1, \dots, m+2\ell, \tag{5.16d}$$

$$\boldsymbol{\mu} \geq \mathbf{0}^{m+2\ell}. \tag{5.16e}$$

Define the pair of vectors $(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\lambda}}) \in \mathbb{R}^m \times \mathbb{R}^\ell$ as $\widetilde{\mu}_i = \mu_i$, $i = 1, \dots, m$; $\widetilde{\lambda}_j = \mu_{m+j} - \mu_{m+\ell+j}$, $j = 1, \dots, \ell$. We also note that the equations (5.16c) and (5.16d) are superfluous, because $\boldsymbol{x}^* \in S$ implies that $h_j(\boldsymbol{x}^*) = 0$, $j = 1, \dots, m$. Therefore, we get the following system for $(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\lambda}})$, known as the KKT necessary optimality conditions for the sets represented by differentiable equality and inequality constraints:

$$\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \widetilde{\mu}_i \nabla g_i(\boldsymbol{x}^*) + \sum_{j=1}^{\ell} \widetilde{\lambda}_j \nabla h_j(\boldsymbol{x}^*) = \mathbf{0}^n, \tag{5.17a}$$

$$\widetilde{\mu}_i g_i(\boldsymbol{x}^*) = 0, \quad i = 1, \dots, m, \tag{5.17b}$$

$$\widetilde{\boldsymbol{\mu}} \geq \mathbf{0}^m. \tag{5.17c}$$

Thus, we have established the following theorem.

Theorem 5.33 (KKT optimality conditions for inequality and equality constraints) *Assume that at a given point $\boldsymbol{x}^* \in S$ Abadie's constraint qualification (5.15) holds, where S is given by (5.10). Then, for this point to* *be a local minimum of a differentiable function f over S it is necessary that there exists a pair of vectors $(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\lambda}}) \in \mathbb{R}^m \times \mathbb{R}^\ell$, such that the system* (5.17) *is satisfied.* ■

Example 5.34 (Example 5.32 revisited) Let us write the system of KKT-conditions for the original representation of the set with one inequality and one equality constraint (see Example 5.14). As has already been mentioned, Abadie's constraint qualification is satisfied, therefore the KKT-system is necessarily solvable:

$$\begin{cases} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \mu_1 \begin{pmatrix} 0 \\ -1 \end{pmatrix} + \lambda_1 \begin{pmatrix} -2 \\ 0 \end{pmatrix} = \mathbf{0}^2, \\ \mu_1 \geq 0, \end{cases}$$

which admits the unique solution $\mu_1 = 0$, $\lambda_1 = 1/2$. ■

5.7 Constraint qualifications

In this section we discuss conditions on the functions involved in the representation (5.10) of a given feasible set S, that all imply Abadie's constraint qualification (5.15).

5.7.1 Mangasarian–Fromovitz CQ (MFCQ)

Definition 5.35 (Mangasarian–Fromovitz CQ) *We say that at the point $\boldsymbol{x} \in S$, where S is given by* (5.10), *the* Mangasarian–Fromovitz *CQ holds if the gradients $\nabla h_j(\boldsymbol{x})$ of the functions h_j, $j = 1, \dots, \ell$, defining the equality constraints, are linearly independent, and the intersection $\overset{\circ}{G}(\boldsymbol{x}) \cap H(\boldsymbol{x})$ is not empty (see Proposition 5.21 for the proof of KKT conditions in the case of inequality-constrained problem).* ■

We state the following result without a "real" proof, but we outline the ideas.

Proposition 5.36 *The MFCQ implies Abadie's CQ.*

Proof.[sketch] Since the gradients $\nabla h_j(\boldsymbol{x})$, $j = 1, \dots, \ell$, are linearly independent, it can be shown that $\mathrm{cl}(\overset{\circ}{G}(\boldsymbol{x}) \cap H(\boldsymbol{x})) \subset T_S(\boldsymbol{x})$ (in the absence of equality constraints, it follows directly from Lemma 5.10).

Furthermore, from Lemma 5.10 applied to the inequality representation of S, i.e., to $\widetilde{G}(\boldsymbol{x})$ defined by (5.14), we know that $T_S(\boldsymbol{x}) \subset (G(\boldsymbol{x}) \cap H(\boldsymbol{x}))$.

Finally, since $\mathring{G}(\boldsymbol{x}) \cap H(\boldsymbol{x}) \neq \emptyset$, it can be shown that $\mathrm{cl}(\mathring{G}(\boldsymbol{x}) \cap H(\boldsymbol{x})) = G(\boldsymbol{x}) \cap H(\boldsymbol{x})$. ■

Example 5.37 Since MFCQ implies Abadie's constraint qualification, Example 5.5 and 5.6 must necessarily violate it. On the other hand, both Examples 5.4 and 5.7 verify it (since they also satisfy stronger constraint qualifications, see Example 5.40 and 5.43). ■

5.7.2 Slater CQ

Definition 5.38 (Slater CQ) *We say that the system of constraints describing the feasible set S via* (5.10) *satisfies the* Slater CQ, *if the functions g_i, $i = 1, \dots, m$, defining the inequality constraints are convex, the functions h_j, $j = 1, \dots, \ell$, defining the equality constraints are affine with linearly independent gradients $\nabla h_j(\boldsymbol{x})$, $j = 1, \dots, \ell$, and, finally, that there exists $\bar{\boldsymbol{x}} \in S$ such that $g_i(\bar{\boldsymbol{x}}) < 0$, for all $i \in \{1, \dots, m\}$.* ■

Proposition 5.39 *The Slater CQ implies the MFCQ.*

Proof. Suppose the Slater CQ holds at $\boldsymbol{x}^* \in S$. By the convexity of the inequality constraints we get:

$$0 > g_i(\bar{\boldsymbol{x}}) = g_i(\bar{\boldsymbol{x}}) - g_i(\boldsymbol{x}^*) \geq \nabla g_i(\boldsymbol{x}^*)^{\mathrm{T}}(\bar{\boldsymbol{x}} - \boldsymbol{x}^*),$$

for all $i \in \mathcal{I}(\boldsymbol{x}^*)$. Furthermore, since the equality constraints are affine, we have that

$$0 = h_j(\bar{\boldsymbol{x}}) - h_j(\boldsymbol{x}^*) = \nabla h_j(\boldsymbol{x}^*)^{\mathrm{T}}(\bar{\boldsymbol{x}} - \boldsymbol{x}^*),$$

$j = 1, \dots, \ell$. Then, $\bar{\boldsymbol{x}} - \boldsymbol{x}^* \in G(\boldsymbol{x}^*) \cap H(\boldsymbol{x}^*)$. ■

Example 5.40 Only Example 5.4 verifies Slater CQ (which in particular explains why it satisfies MFCQ as well, see Example 5.37). ■

5.7.3 Linear independence CQ (LICQ)

Definition 5.41 (LICQ) *We say that at the point $\boldsymbol{x} \in S$, where S is given by* (5.10), *the* linear independence CQ *holds if the gradients $\nabla g_i(\boldsymbol{x})$ of the functions g_i, $i \in \mathcal{I}(\boldsymbol{x})$, defining the active inequality constraints, as well as the gradients $\nabla h_j(\boldsymbol{x})$ of the functions h_j, $j = 1, \dots, \ell$, defining the equality constraints, are linearly independent.* ■

Proposition 5.42 *The LICQ implies the MFCQ.*

Proof.[sketch] Assume that $\mathring{G}(\boldsymbol{x}^*) \cap H(\boldsymbol{x}^*) = \emptyset$, i.e., the system $\boldsymbol{G}^{\mathrm{T}}\boldsymbol{p} < \mathbf{0}^{|\mathcal{I}(\boldsymbol{x}^*)|}$ and $\boldsymbol{H}^{\mathrm{T}}\boldsymbol{p} = \mathbf{0}^{\ell}$ is unsolvable, where $\boldsymbol{G}$ and $\boldsymbol{H}$ are the matrices having the gradients of the active inequality and equality constraints, respectively, as their columns. Using a separation result similar to Farkas' Lemma (cf. Theorem 3.30) it can be shown that the system $\boldsymbol{G}\boldsymbol{\mu} + \boldsymbol{H}\boldsymbol{\lambda} = \mathbf{0}^n$, $\boldsymbol{\mu} \geq \mathbf{0}^{|\mathcal{I}(\boldsymbol{x}^*)|}$ has a nonzero solution $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathbb{R}^{|\mathcal{I}(\boldsymbol{x}^*)|+\ell}$, which contradicts the linear independence assumption on the gradients. ■

In fact, the solution $(\boldsymbol{\mu}, \boldsymbol{\lambda})$ to the KKT system (5.17), if one exists, is necessarily unique in this case, and therefore LICQ is a rather strong assumption in many practical situations.

Example 5.43 Only Example 5.7 in the original description using both inequality and equality constraints verifies the LICQ (which in particular explains why it satisfies the MFCQ, see Example 5.37, and why the Lagrange multipliers are unique in this case, see Example 5.34). ■

5.7.4 Affine constraints

Assume that both the functions g_i, $i = 1, \dots, m$, defining the inequality constraints and the functions h_j, $j = 1, \dots, \ell$, defining the equality constraints in the representation (5.10), are affine. Then, the radial cone $R_S(\boldsymbol{x})$ (see Definition 5.1) is equal to $G(\boldsymbol{x}) \cap H(\boldsymbol{x})$ (see Example 4.21). Owing to the inclusions $R_S(\boldsymbol{x}) \subset T_S(\boldsymbol{x})$ (Proposition 5.3) and $T_S(\boldsymbol{x}) \subset \widetilde{G}(\boldsymbol{x}) = G(\boldsymbol{x}) \cap H(\boldsymbol{x})$ (Lemma 5.10), where $\widetilde{G}(\boldsymbol{x})$ was defined in Section 5.6 (cf. (5.12) and the discussion thereafter), Abadie's CQ (5.15) holds in this case.

Thus, the following claim is established.

Proposition 5.44 *If all (inequality and equality) constraints are affine, then Abadie's CQ is satisfied.* ■

5.8 Sufficiency of KKT–conditions under convexity

In general, the KKT necessary conditions do not imply local optimality, as has been mentioned before (see, e.g., an example right after the proof of Theorem 4.13). However, if the optimization problem (5.1) is convex, then the KKT conditions are *sufficient* for global optimality.

Theorem 5.45 (sufficiency of the KKT conditions for convex problems) *Assume that the problem* (5.1) *with the feasible set S given by* (5.10) *is convex, i.e., the objective function f as well as the functions g_i, $i = 1, \dots, m$, are convex, and the functions h_j, $j = 1, \dots, \ell$, are affine. Assume further that for $\boldsymbol{x}^* \in S$ the KKT conditions* (5.17) *are satisfied. Then, $\boldsymbol{x}^*$ is a globally optimal solution of the problem* (5.1).

Proof. Choose an arbitrary $\boldsymbol{x} \in S$. Then, by the convexity of the functions g_i, $i = 1, \dots, m$, it holds that

$$-\nabla g_i(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) \geq g_i(\boldsymbol{x}^*) - g_i(\boldsymbol{x}) = -g_i(\boldsymbol{x}) \geq 0, \tag{5.18}$$

for all $i \in \mathcal{I}(\boldsymbol{x}^*)$, and using the affinity of the functions h_j, $j = 1, \dots, \ell$, we get that

$$-\nabla h_j(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) = h_j(\boldsymbol{x}^*) - h_j(\boldsymbol{x}) = 0, \tag{5.19}$$

for all $j = 1, \dots, \ell$. Using the convexity of the objective function, equations (5.17a) and (5.17b), non-negativity of the Lagrange multipliers μ_i, $i \in \mathcal{I}(\boldsymbol{x}^*)$, and equations (5.18) and (5.19) we obtain the inequality

$$\begin{aligned} f(\boldsymbol{x}) - f(\boldsymbol{x}^*) &\geq \nabla f(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) \\ &= -\sum_{i \in \mathcal{I}(\boldsymbol{x}^*)} \mu_i \nabla g_i(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) - \sum_{j=1}^{\ell} \lambda_j \nabla h_j(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0. \end{aligned}$$

Since the point $\boldsymbol{x} \in S$ was arbitrary, this shows the global optimality of $\boldsymbol{x}^*$ in (5.1). ■

Theorem 5.45 combined with the necessity of the KKT conditions under appropriate CQ leads to the following statement.

Corollary 5.46 Assume that the problem (5.1) is *convex* and verifies the Slater CQ (Definition 5.38). Then, for $\boldsymbol{x}^* \in S$ to be a globally optimal solution of (5.1) it is both *necessary and sufficient* to verify the system (5.17).

Not surprisingly, without the Slater constraint qualification the KKT conditions remain only sufficient (i.e., they are unnecessarily strong), as the following example demonstrates.

Example 5.47 Consider the optimization problem to

$$\begin{aligned} &\text{minimize } x_1, \\ &\text{subject to } \begin{cases} x_1^2 + x_2 \leq 0, \\ \quad\; -x_2 \leq 0, \end{cases} \end{aligned}$$

which is convex but has only one feasible point $\mathbf{0}^2 \in \mathbb{R}^2$. At this unique point both the inequality constraints are active, and thus the Slater CQ is violated, which however does not contradict the global optimality of $\mathbf{0}^2$. It is easy to check that the KKT system

$$\begin{cases} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix} \boldsymbol{\mu} = \mathbf{0}^2, \\ \qquad\qquad\qquad\qquad \boldsymbol{\mu} \geq \mathbf{0}^2, \end{cases}$$

is unsolvable, and therefore the KKT conditions are not necessary without a CQ even for convex problems. ■

5.9 Applications and examples

Example 5.48 Consider a symmetric square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, and the optimization problem

$$\begin{aligned} &\text{minimize } -\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}, \\ &\text{subject to } \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} \leq 1. \end{aligned}$$

The only constraint of this problem is convex; furthermore, $(\mathbf{0}^n)^{\mathrm{T}}\mathbf{0}^n = 0 < 1$, and thus Slater's CQ (Definition 5.38) is verified. Therefore, the KKT conditions are necessary for the local optimality in this problem. We will find all the possible KKT points, and then choose a globally optimal point among them.

$\nabla(-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}) = -2\boldsymbol{A}\boldsymbol{x}$ ($\boldsymbol{A}$ is symmetric), and $\nabla(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}) = 2\boldsymbol{x}$. Thus, the KKT system is as follows: $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} \leq 1$ and

$$\begin{aligned} -2\boldsymbol{A}\boldsymbol{x} + 2\mu\boldsymbol{x} &= \mathbf{0}^n, \\ \mu &\geq 0, \\ \mu(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} - 1) &= 0. \end{aligned}$$

From the first two equations we immediately see that either $\boldsymbol{x} = \mathbf{0}^n$, or the pair $(\mu, \boldsymbol{x})$ is respectively a nonnegative eigenvalue and a corresponding eigenvector of $\boldsymbol{A}$. In the former case, from the complementarity condition we deduce that $\mu = 0$.

Thus, we can characterize the KKT-points of the problem into the following groups:

1. Let $\mu_1, \dots, \mu_k$ be all the *positive* eigenvalues of $\boldsymbol{A}$ (if any), and define $X_i = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = 1, \boldsymbol{A}\boldsymbol{x} = \mu_i\boldsymbol{x} \,\}$ to be the set of corresponding eigenvectors of length 1, $i = 1, \dots, k$. Then, $(\boldsymbol{x}, \mu_i)$

is a KKT-point with the corresponding multiplier for every $\boldsymbol{x} \in X_i$, $i = 1, \ldots, k$. Moreover, $-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = -\mu_i\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = -\mu_i < 0$, for every $\boldsymbol{x} \in X_i$, $i = 1, \ldots, k$.

2. Define also $X_0 = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} \leq 1, \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}^n \,\}$. Then, the pair $(\boldsymbol{x}, 0)$ is a KKT-point with the corresponding multiplier for every $\boldsymbol{x} \in X_0$. We note that if the matrix $\boldsymbol{A}$ is nonsingular, then $X_0 = \{\boldsymbol{0}^n\}$. In any case, $-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = 0$ for every $\boldsymbol{x} \in X_0$.

Therefore, if the matrix $\boldsymbol{A}$ has any positive eigenvalue, then the global minima points of the problem we consider are the eigenvectors of length one, corresponding to the largest positive eigenvalue; otherwise, every vector that satisfies $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}^n$ is globally optimal. ■

Example 5.49 Similarly to the previous example, consider the following equality-constrained minimization problem associated with a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$:

$$\begin{aligned} &\text{minimize } -\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}, \\ &\text{subject to } \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = 1. \end{aligned}$$

The gradient of the only equality constraint equals $2\boldsymbol{x}$, and since $\boldsymbol{0}^n$ is infeasible, LICQ is satisfied by this problem (Definition 5.41), and the KKT conditions are necessary for local optimality. In this case, the KKT system is extremely simple: $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = 1$ and

$$-2\boldsymbol{A}\boldsymbol{x} + 2\lambda\boldsymbol{x} = \boldsymbol{0}^n.$$

Let $\lambda_1 < \lambda_2 < \cdots < \lambda_k$ denote all distinct eigenvalues of $\boldsymbol{A}$, and define as before $X_i = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = 1, \boldsymbol{A}\boldsymbol{x} = \lambda_i\boldsymbol{x} \,\}$ to be the set of corresponding eigenvectors of length 1, $i = 1, \ldots, k$. Then, $(\boldsymbol{x}, \lambda_i)$ is a KKT-point with the corresponding multiplier for every $\boldsymbol{x} \in X_i$, $i = 1, \ldots, k$. Furthermore, since $-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = -\lambda_i$ for every $\boldsymbol{x} \in X_i$, $i = 1, \ldots, k$, it holds that every $\boldsymbol{x} \in X_k$, that is, every eigenvector corresponding to the *largest* eigenvalue, is globally optimal.

Considering two problems corresponding to $\boldsymbol{A}$ and $-\boldsymbol{A}$, we may deduce that $\|\boldsymbol{A}\| = \max_{1 \leq i \leq n}\{\, |\lambda_i| \,\}$, a very well known fact in linear algebra. ■

Example 5.50 Consider the problem of finding the projection of a given point $\boldsymbol{x}^*$ onto the hyperplane $\{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \,\}$, where $\boldsymbol{A} \in \mathbb{R}^{k \times n}$, $\boldsymbol{b} \in \mathbb{R}^k$. Thus, we consider the following minimization problem with affine constraints (so that the KKT conditions are necessary for the local optimality, see Section 5.7.4):

$$\begin{aligned}
&\text{minimize } \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*), \\
&\text{subject to } \boldsymbol{Ax} = \boldsymbol{b}.
\end{aligned}$$

The KKT-system in this case is written as follows:

$$\begin{aligned}
\boldsymbol{Ax} &= \boldsymbol{b}, \\
(\boldsymbol{x} - \boldsymbol{x}^*) + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} &= \boldsymbol{0}^n,
\end{aligned}$$

for some $\boldsymbol{\lambda} \in \mathbb{R}^k$. Pre-multiplying the last equation with $\boldsymbol{A}$, and using the fact that $\boldsymbol{Ax} = \boldsymbol{b}$ we get:

$$\boldsymbol{AA}^{\mathrm{T}}\boldsymbol{\lambda} = \boldsymbol{Ax}^* - \boldsymbol{b}.$$

Substituting an arbitrary solution of this equation into the KKT-system, we calculate $\boldsymbol{x}$ via

$$\boldsymbol{x} = \boldsymbol{x}^* - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}.$$

It can be shown that the vector $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}$ is constant for every Lagrange multiplier $\boldsymbol{\lambda}$, so using this formula we obtain the globally optimal solution to our minimization problem.

Now assume that the columns of $\boldsymbol{A}^{\mathrm{T}}$ are linearly independent, i.e., LICQ holds. Then, the matrix $\boldsymbol{AA}^{\mathrm{T}}$ is nonsingular, and the multiplier $\boldsymbol{\lambda}$ is therefore unique:

$$\boldsymbol{\lambda} = (\boldsymbol{AA}^{\mathrm{T}})^{-1}(\boldsymbol{Ax}^* - \boldsymbol{b}).$$

Substituting this into the KKT-system, we finally obtain the well-known formula for calculating the projection:

$$\boldsymbol{x} = \boldsymbol{x}^* - \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{AA}^{\mathrm{T}})^{-1}(\boldsymbol{Ax}^* - \boldsymbol{b}).$$

■

5.10 Notes and further reading

One cannot overemphasize the importance of Karush–Kuhn–Tucker type optimality conditions for any development in optimization. We essentially follow the ideas presented in [BSS93, Chapters 4 and 5]; an alternative presentation may be found in [Ber99, Chapter 3]. The original papers by Fritz John [Joh48], and Kuhn and Tucker [KuT51] might also be interesting.

Various forms of constraint qualifications play especially important role in studies of parametric optimization problems [Fia83, BoS00]. Original presentation of constraint qualifications, some of which we considered in this chapter, may be found in the works of Arrow, Hurwitz, and Uzawa [AHU61], Abadie [Aba67], Mangasarian and Fromowitz [MaF67], Guignard [Gui69], Zangwill [Zan69], and Evans [Eva70].

5.11 Exercises

Exercise 5.1 Consider the following problem:

$$\begin{aligned} &\text{minimize } f(\boldsymbol{x}) = 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10_x2, \\ &\text{subject to } \quad x_1^2 + x_2^2 \le 5, \\ &\qquad\qquad 3x_1 + x_2 \le 6. \end{aligned}$$

Verify whether the point $\boldsymbol{x}^0 = (2,1)^{\mathrm{T}}$ is a KKT point for this problem. Is this an optimal solution? Which CQ are satisfied at the point $\boldsymbol{x}^0$? ■

Exercise 5.2 (optimality conditions, exam 020529) (a) Consider the following optimization problem:

$$\begin{aligned} &\text{minimize } x^2, \\ &\text{subject to } \sin(x) \le -1. \end{aligned} \tag{5.20}$$

Find every locally and every globally optimal solution to this problem. Write down the KKT conditions. Are they necessary/sufficient for this problem?

(b) Do the locally/globally optimal solutions to the problem (5.20) satisfy the FJ optimality conditions?

(c) Question the usefulness of the FJ optimality conditions by finding a point (x, y), which satisfies these conditions for the problem:

$$\begin{aligned} &\text{minimize } y, \\ &\text{subject to } \begin{cases} x^2 + y^2 \le 1, \\ \qquad x^3 \ge y^4, \end{cases} \end{aligned}$$

but, nevertheless, is neither the local nor the global minimum for this problem. ■

Exercise 5.3 Consider the following *linear* programming problem:

$$\begin{aligned} &\text{minimize } \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}, \\ &\text{subject to } \boldsymbol{A}\boldsymbol{x} \ge \boldsymbol{b}. \end{aligned}$$

State the KKT-conditions for this problem. Verify that at every KKT-point $\boldsymbol{x}$ the following equality is verified:

$$\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\lambda},$$

where $\boldsymbol{\lambda}$ is a vector of KKT-multipliers. ■

Exercise 5.4 (optimality conditions, exam 020826) (a) Consider the nonlinear programming problem with equality constraints:

$$\begin{aligned} &\text{minimize } f(\boldsymbol{x}),\\ &\text{subject to } \begin{cases} h_1(\boldsymbol{x}) &= 0,\\ \vdots \\ h_m(\boldsymbol{x}) &= 0, \end{cases} \end{aligned} \tag{5.21}$$

where f, $h_1, \dots, h_m$ are continuously differentiable functions.

Show that the problem (5.21) is equivalent to the following problem with one inequality constraint:

$$\begin{aligned} &\text{minimize } f(\boldsymbol{x}),\\ &\text{subject to } \sum_{i=1}^{m} \big(h_i(\boldsymbol{x})\big)^2 \leq 0. \end{aligned} \tag{5.22}$$

Show (by giving a formal argument or an illustrative example) that the KKT conditions for the latter problem are not necessary for local optimality.

Can Slater's CQ or LICQ be satisfied for the problem (5.22)?

(b) Consider the unconstrained minimization problem

$$\text{minimize } \max\big\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\big\},$$

where $f_1 : \mathbb{R}^n \to \mathbb{R}$, $f_2 : \mathbb{R}^n \to \mathbb{R}$ are in C^1.

Show that if $\boldsymbol{x}^*$ is a local minimum for this problem, then there exist $\mu_1, \mu_2 \in \mathbb{R}$ such that

$$\mu_1 \geq 0, \mu_2 \geq 0, \quad \mu_1 \nabla f_1(\boldsymbol{x}^*) + \mu_2 \nabla f_2(\boldsymbol{x}^*) = 0, \quad \mu_1 + \mu_2 = 1,$$

and $\mu_i = 0$ if $f_i(\boldsymbol{x}^*) < \max\big\{ f_1(\boldsymbol{x}^*), f_2(\boldsymbol{x}^*) \big\}$, $i = 1, 2$. ■

Exercise 5.5 Consider the following optimization problem:

$$\begin{aligned} &\text{minimize } \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x},\\ &\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \end{aligned}$$

Assume that the matrix $\boldsymbol{A}$ has full row rank. Find the globally optimal solution to this problem. ■

Exercise 5.6 Consider the following optimization problem:

$$\begin{aligned} &\text{minimize } \sum_{j=1}^{n} c_j x_j, \\ &\text{subject to } \begin{cases} \displaystyle\sum_{j=1}^{n} x_j^2 \leq 1, \\ -x_j \leq 0, \quad j = 1, \dots, n. \end{cases} \end{aligned} \tag{5.23}$$

Assume that $\min\{c_1, \dots, c_n\} < 0$, and let us introduce KKT multipliers $\lambda \geq 0$ and $\mu_j \geq 0$, $j = 1, \dots, n$ for the inequality constraints.

(a) Show that the equalities

$$\begin{aligned} x_j^* &= \min\{0, c_j\}/(2\lambda^*), \qquad j = 1, \dots, n, \\ \lambda^* &= \frac{1}{2}\left(\sum_{j=1}^{n}[\min\{0, c_j\}]^2\right)^{1/2}, \\ \mu_j^* &= \max\{0, c_j\}, \qquad j = 1, \dots, n, \end{aligned}$$

define a KKT point for (5.23).

(b) Show that there is only one locally optimal solution to the problem (5.23). ■

Exercise 5.7 (optimality conditions, exam 040308) Consider the following optimization problem:

$$\begin{aligned} &\text{minimize } f(x, y) = \frac{1}{2}(x-2)^2 + \frac{1}{2}(y-1)^2, \\ &\text{subject to } \begin{cases} x - y \geq 0, \\ y \geq 0, \\ y(x - y) = 0, \end{cases} \end{aligned} \tag{5.24}$$

where $x, y \in \mathbb{R}$.

(a) Find *all* points of global and local minima (you may do this graphically), as well as *all* KKT-points. Is this a convex problem? Are the KKT optimality conditions necessary and/or sufficient for local optimality *in this problem*?

(b) Demonstrate that LICQ is violated at *every feasible point* of the problem (5.24). Show that instead of solving the problem (5.24) we can solve *two* convex optimization problems that furthermore verify some constraint qualification, and then choose the best point out of the two.

(c) Generalize the procedure from the previous part to more general optimization problems:

$$\begin{aligned} &\text{minimize } \ g(\boldsymbol{x}), \\ &\text{subject to } \begin{cases} \boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{x} \geq b_i, & i = 1, \dots, n, \\ x_i \geq 0, & i = 1, \dots, n, \\ x_i(\boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{x} - b_i) = 0, & i = 1, \dots, n, \end{cases} \end{aligned}$$

where $\boldsymbol{x} = (x_1, \dots, x_n)^{\mathrm{T}} \in \mathbb{R}^n$, $\boldsymbol{a}_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$, $i = 1, \dots, n$, and $g : \mathbb{R}^n \to \mathbb{R}$ is a convex differentiable function. ∎

Exercise 5.8 Determine the values of the parameter c for which the point $(x, y) = (4, 3)$ is an optimal solution to the following problem:

$$\begin{aligned} &\text{minimize } \ cx + y, \\ &\text{subject to } \begin{cases} x^2 + y^2 \leq 25, \\ x - y \leq 1, \end{cases} \end{aligned}$$

where $x, y \in \mathbb{R}$. ∎

Exercise 5.9 Consider the following optimization problem:

$$\begin{aligned} &\text{minimize } \ f(\boldsymbol{x}) = \sum_{j=1}^{n} \frac{x_j^2}{c_j}, \\ &\text{subject to } \begin{cases} \displaystyle\sum_{j=1}^{n} x_j = D, \\ x_j \geq 0, \quad j = 1, \dots, n, \end{cases} \end{aligned}$$

where $c_j > 0$, $j = 1, \dots, n$, and $D > 0$. Find the unique globally optimal solution to this problem. ∎

Lagrangian duality

VI

This chapter collects some basic results on Lagrangian duality, in particular as it applies to convex programs with no duality gap.

6.1 The relaxation theorem

Given the problem to find

$$f^* := \operatorname*{infimum}_{\boldsymbol{x}} f(\boldsymbol{x}), \tag{6.1a}$$

$$\text{subject to } \boldsymbol{x} \in S, \tag{6.1b}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a given function and $S \subseteq \mathbb{R}^n$, we define a *relaxation* to (6.1) to be a problem of the following form: to find

$$f_R^* := \operatorname*{infimum}_{\boldsymbol{x}} f_R(\boldsymbol{x}), \tag{6.2a}$$

$$\text{subject to } \boldsymbol{x} \in S_R, \tag{6.2b}$$

where $f_R : \mathbb{R}^n \to \mathbb{R}$ is a function with the property that $f_R \leq f$ on S, and where $S_R \supseteq S$. For this pair of problems, we have the following basic result.

Theorem 6.1 (Relaxation Theorem) (a) [relaxation] $f_R^* \leq f^*$.

(b) [infeasibility] *If* (6.2) *is infeasible, then so is* (6.1).

(c) [optimal relaxation] *If the problem* (6.2) *has an optimal solution,* $\boldsymbol{x}_R^*$, *for which it holds that*

$$\boldsymbol{x}_R^* \in S \qquad \textit{and} \qquad f_R(\boldsymbol{x}_R^*) = f(\boldsymbol{x}_R^*), \tag{6.3}$$

then $\boldsymbol{x}_R^*$ *is an optimal solution to* (6.1) *as well.*

Proof. The result in (a) is obvious, as every solution feasible in (6.1) is both feasible in (6.2) and has a lower objective value in the latter problem.

The result in (b) follows for similar reasons.

For the result in (c), we note that

$$f(\boldsymbol{x}_R^*) = f_R(\boldsymbol{x}_R^*) \le f_R(\boldsymbol{x}) \le f(\boldsymbol{x}), \qquad \boldsymbol{x} \in S,$$

from which the result follows. ■

This basic result will be utilized both in this chapter and later on to motivate why Lagrangian relaxation, objective function linearization and penalization are relaxations, and to derive optimality conditions and algorithms based on them.

6.2 Lagrangian duality

In this section we formulate the Lagrangian dual problem and establish its convexity. The Weak Duality Theorem is also established, and we introduce the terms "Lagrange multiplier" and "duality gap."

6.2.1 Lagrangian relaxation and the dual problem

Consider the optimization problem to find

$$f^* := \operatorname*{infimum}_{\boldsymbol{x}} f(\boldsymbol{x}), \tag{6.4a}$$

$$\text{subject to } \boldsymbol{x} \in X, \tag{6.4b}$$

$$g_i(\boldsymbol{x}) \le 0, \qquad i = 1, \dots, m, \tag{6.4c}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}$ ($i = 1, 2, \dots, m$) are given functions, and $X \subseteq \mathbb{R}^n$.

For this problem, we assume that

$$-\infty < f^* < \infty, \tag{6.5}$$

that is, that f is bounded from below and that the problem has at least one feasible solution.

For an arbitrary vector $\boldsymbol{\mu} \in \mathbb{R}^m$, we define the *Lagrange function*

$$L(\boldsymbol{x}, \boldsymbol{\mu}) := f(\boldsymbol{x}) + \sum_{i=1}^m \mu_i g_i(\boldsymbol{x}) = f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}). \tag{6.6}$$

We call the vector $\boldsymbol{\mu}^* \in \mathbb{R}^m$ a *Lagrange multiplier* if it is non-negative and if $f^* = \inf_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \boldsymbol{\mu}^*)$ holds.

Theorem 6.2 (Lagrange multipliers and global optima) *Let $\boldsymbol{\mu}^*$ be a Lagrange multiplier. Then, $\boldsymbol{x}^*$ is an optimal solution to* (6.4) *if and only if $\boldsymbol{x}^*$ is feasible in* (6.4) *and*

$$\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in X} L(\boldsymbol{x}, \boldsymbol{\mu}^*), \qquad \textit{and} \qquad \mu_i^* g_i(\boldsymbol{x}^*) = 0, \quad i = 1, \dots, m. \tag{6.7}$$

Proof. If $\boldsymbol{x}^*$ is an optimal solution to (6.4), then it is in particular feasible, and

$$f^* = f(\boldsymbol{x}^*) \geq L(\boldsymbol{x}^*, \boldsymbol{\mu}^*) \geq \operatorname*{infimum}_{\boldsymbol{x}\in X} L(\boldsymbol{x}, \boldsymbol{\mu}^*),$$

where the first inequality stems from the feasibility of $\boldsymbol{x}^*$ and the definition of a Lagrange multiplier. The second part of that definition implies that $f^* = \inf_{\boldsymbol{x}\in X} L(\boldsymbol{x}, \boldsymbol{\mu}^*)$, so that equality holds throughout in the above line of inequalities. Hence, (6.7) follows.

Conversely, if $\boldsymbol{x}^*$ is feasible and (6.7) holds, then by the use of the definition of a Lagrange multiplier,

$$f(\boldsymbol{x}^*) = L(\boldsymbol{x}^*, \boldsymbol{\mu}^*) = \operatorname*{minimum}_{\boldsymbol{x}\in X} L(\boldsymbol{x}, \boldsymbol{\mu}^*) = f^*,$$

so $\boldsymbol{x}^*$ is a global optimum. ∎

Let

$$q(\boldsymbol{\mu}) := \operatorname*{infimum}_{\boldsymbol{x}\in X} L(\boldsymbol{x}, \boldsymbol{\mu}) \tag{6.8}$$

be the *Lagrangian dual function*, defined by the infimum value of the Lagrange function over X; the *Lagrangian dual problem* is to

$$\operatorname*{maximize}_{\boldsymbol{\mu}} q(\boldsymbol{\mu}), \tag{6.9a}$$

$$\text{subject to } \boldsymbol{\mu} \geq \mathbf{0}^m. \tag{6.9b}$$

For some $\boldsymbol{\mu}$, $q(\boldsymbol{\mu}) = -\infty$ is possible; if this is true for all $\boldsymbol{\mu} \geq \mathbf{0}^m$, then

$$q^* := \operatorname*{supremum}_{\boldsymbol{\mu}\geq\mathbf{0}^m} q(\boldsymbol{\mu})$$

equals $-\infty$.

The *effective domain* of q is

$$D_q := \{\, \boldsymbol{\mu} \in \mathbb{R}^m \mid q(\boldsymbol{\mu}) > -\infty \,\}.$$

Theorem 6.3 (convex dual problem) *The effective domain D_q of q is convex, and q is concave on D_q.*

Proof. Let $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{\mu}, \bar{\boldsymbol{\mu}} \in \mathbb{R}^m$, and $\alpha \in [0,1]$. We have that

$$L(\boldsymbol{x}, \alpha\boldsymbol{\mu} + (1-\alpha)\bar{\boldsymbol{\mu}}) = \alpha L(\boldsymbol{x}, \boldsymbol{\mu}) + (1-\alpha) L(\boldsymbol{x}, \bar{\boldsymbol{\mu}}).$$

Take the infimum over $\boldsymbol{x} \in X$ on both sides; then,

$$\begin{aligned}
\inf_{\boldsymbol{x}\in X} L(\boldsymbol{x}, \alpha\boldsymbol{\mu} + (1-\alpha)\bar{\boldsymbol{\mu}}) &= \inf_{\boldsymbol{x}\in X} \{\alpha L(\boldsymbol{x}, \boldsymbol{\mu}) + (1-\alpha) L(\boldsymbol{x}, \bar{\boldsymbol{\mu}})\} \\
&\geq \inf_{\boldsymbol{x}\in X} \alpha L(\boldsymbol{x}, \boldsymbol{\mu}) + \inf_{\boldsymbol{x}\in X} (1-\alpha) L(\boldsymbol{x}, \bar{\boldsymbol{\mu}}) \\
&= \alpha \inf_{\boldsymbol{x}\in X} L(\boldsymbol{x}, \boldsymbol{\mu}) + (1-\alpha) \inf_{\boldsymbol{x}\in X} L(\boldsymbol{x}, \bar{\boldsymbol{\mu}}),
\end{aligned}$$

due to the fact that $\alpha \in [0,1]$, and that the sum of infimum values may be smaller than the infimum of the sum, since in the former case we have the possibility to choose different optimal solutions in the two problems. Hence,

$$q(\alpha\boldsymbol{\mu} + (1-\alpha)\bar{\boldsymbol{\mu}}) \geq \alpha q(\boldsymbol{\mu}) + (1-\alpha) q(\bar{\boldsymbol{\mu}})$$

holds. This inequality has two implications: if $\boldsymbol{\mu}$ and $\bar{\boldsymbol{\mu}}$ belong to D_q, then so does $\alpha\boldsymbol{\mu} + (1-\alpha)\bar{\boldsymbol{\mu}}$, so D_q is convex, and further, q is concave on D_q. ∎

That the Lagrangian dual problem always is convex (we indeed maximize a concave function!) is very good news, because it means that it can be solved efficiently. What remains of course is to show how a Lagrangian dual optimal solution can be used to generate a primal optimal solution.

Next, we establish that every feasible point in the Lagrangian *dual* problem always underestimates the objective function value of every feasible point in the *primal* problem; hence, also their optimal values have this relationship.

Theorem 6.4 (Weak Duality Theorem) *Let $\boldsymbol{x}$ and $\boldsymbol{\mu}$ be feasible in* (6.4) *and* (6.9), *respectively. Then,*

$$q(\boldsymbol{\mu}) \leq f(\boldsymbol{x}).$$

In particular,

$$q^* \leq f^*$$

holds.

If $q(\boldsymbol{\mu}) = f(\boldsymbol{x})$, then the pair $(\boldsymbol{x}, \boldsymbol{\mu})$ is optimal in its respective problem.

Proof. For all $\boldsymbol{\mu} \geq \boldsymbol{0}^m$ and $\boldsymbol{x} \in X$ with $\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}^m$,

$$q(\boldsymbol{\mu}) = \operatorname*{infimum}_{\boldsymbol{z}\in X} L(\boldsymbol{z}, \boldsymbol{\mu}) \leq f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) \leq f(\boldsymbol{x}),$$

so

$$q^* = \operatorname*{supremum}_{\boldsymbol{\mu} \geq \mathbf{0}^m} q(\boldsymbol{\mu}) \leq \operatorname*{infimum}_{\boldsymbol{x} \in X : \boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}^m} f(\boldsymbol{x}) = f^*.$$

The result follows. ∎

Weak duality is also a consequence of the Relaxation Theorem: For any $\boldsymbol{\mu} \geq \mathbf{0}^m$, let

$$S := X \cap \{ \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}^m \}, \tag{6.10a}$$

$$S_R := X, \tag{6.10b}$$

$$f_R := L(\boldsymbol{\mu}, \cdot). \tag{6.10c}$$

Then, the weak duality statement is the result in Theorem 6.1(a), whence Lagrangian relaxation is a relaxation in terms of the definition in Section 6.1.

If our initial assumption (6.5) is false, then what does weak duality imply? Suppose that $f^* = -\infty$. Then, weak duality implies that $q(\boldsymbol{\mu}) = -\infty$ for all $\boldsymbol{\mu} \geq \mathbf{0}^m$, that is, the dual problem is infeasible. Suppose then that $X \neq \emptyset$ but that $X \cap \{ \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}^m \}$ is empty. Then, $f^* = \infty$, by convention. The dual function satisfies $q(\boldsymbol{\mu}) < \infty$ for all $\boldsymbol{\mu} \geq \mathbf{0}^m$, but it is possible that $q^* = -\infty$, $-\infty < q^* < \infty$, or $q^* = \infty$ (see [Ber99, Figure 5.1.8]). For linear programs, $\infty < q^* < \infty$ implies $\infty < f^* < \infty$, see below.

If $q^* = f^*$, we say that there is *no duality gap*. If there exists a Lagrange multiplier vector, then by the weak duality theorem, this implies that there is no duality gap. The converse is not true in general: there may be cases where no Lagrange multiplier exists even when there is no duality gap; in that case though, the Lagrangian dual problem cannot have an optimal solution, as implied by the following result.

Proposition 6.5 (duality gap and the existence of Lagrange multipliers) (a) *If there is no duality gap, then the set of Lagrange multipliers equals the set of optimal dual solutions (which however may be empty).*

(b) *If there is a duality gap, then there are no Lagrange multipliers.*

Proof. By definition, a vector $\boldsymbol{\mu}^* \geq \mathbf{0}^m$ is a Lagrange multiplier if and only if $f^* = q(\boldsymbol{\mu}^*) \leq q^*$, the equality following from the definition of $q(\boldsymbol{\mu}^*)$ and the inequality from the definition of q^* as the supremum of $q(\boldsymbol{\mu})$ over $\mathbb{R}^m_+$. By weak duality, this relation holds if and only if there is no duality gap and $\boldsymbol{\mu}^*$ is an optimal dual solution. ∎

Before moving on, we remark on the *statement* of the problem (6.4). There are several ways in which the original set of constraints of the problem can be placed either within the definition of the *ground set* X (which is kept intact), or within the explicit constraints defined by the functions g_i (which are Lagrangian relaxed). How to distinguish between the two, that is, how to decide whether a constraint should be kept or be Lagrangian relaxed, depends on several factors. For example, keeping more constraints within X may result in a smaller duality gap, and with fewer multipliers also result in a simpler Lagrangian dual problem. On the other hand, the Lagrangian subproblem defining the dual function simultaneously becomes more complex and difficult to solve. There are no immediate rules to follow, but experimentation and experience.

6.2.2 Global optimality conditions

The following result characterizes every optimal primal and dual solution. It is applicable only in the presence of Lagrange multipliers; in other words, the system (6.11) is consistent if and only if there exists a Lagrange multiplier and there is no duality gap.

Theorem 6.6 (global optimality conditions in the absence of a duality gap) *The vector $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ is a pair of optimal primal solution and Lagrange multiplier if and only if*

$$\boldsymbol{\mu}^* \geq \mathbf{0}^m, \qquad \text{(Dual feasibility)} \tag{6.11a}$$

$$\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \boldsymbol{\mu}^*), \qquad \text{(Lagrangian optimality)} \tag{6.11b}$$

$$\boldsymbol{x}^* \in X, \quad \boldsymbol{g}(\boldsymbol{x}^*) \leq \mathbf{0}^m, \qquad \text{(Primal feasibility)} \tag{6.11c}$$

$$\mu_i^* g_i(\boldsymbol{x}^*) = 0, \quad i = 1, \dots, m. \qquad \text{(Complementary slackness)} \tag{6.11d}$$

Proof. Suppose that (6.11) is satisfied. We apply the Relaxation Theorem 6.1, as follows. Consider the identification in (6.10), with $\boldsymbol{\mu} = \boldsymbol{\mu}^* \geq \mathbf{0}^m$ and $\boldsymbol{x}_R = \boldsymbol{x}^*$. We then note the following equivalences:

1. The relaxed solution, $\boldsymbol{x}_R$, is a Lagrangian optimal solution. (Lagrangian optimality is fulfilled.)
2. That $\boldsymbol{x}_R \in S$ means that it is feasible in the primal problem. (Primal feasibility is fulfilled.)
3. That $f^* = f_R^*$ means that $(\boldsymbol{\mu}^*)^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}^*) = 0$. (Complementary slackness is fulfilled.)

To conclude, if the conditions in (6.11) are satisfied, then Theorem 6.1 implies that the vector $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ is a pair of optimal primal solution and Lagrange multiplier.

Conversely, if $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ is a pair of optimal primal solution and Lagrange multiplier, then they are obviously primal and dual feasible, respectively. The last two equations in (6.11) follow from Theorem 6.2. ■

Theorem 6.7 (global optimality and saddle points) *The vector $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ is a pair of optimal primal solution and Lagrange multiplier if and only if $\boldsymbol{x}^* \in X$, $\boldsymbol{\mu}^* \geq \mathbf{0}^m$, and $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ is a* saddle point *of the Lagrangian function on $X \times \mathbb{R}^m_+$, that is,*

$$L(\boldsymbol{x}^*, \boldsymbol{\mu}) \leq L(\boldsymbol{x}^*, \boldsymbol{\mu}^*) \leq L(\boldsymbol{x}, \boldsymbol{\mu}^*), \qquad (\boldsymbol{x}, \boldsymbol{\mu}) \in X \times \mathbb{R}^m_+, \tag{6.12}$$

holds.

Proof. We establish that (6.11) and (6.12) are equivalent. The result then follows from Theorem 6.6. The first inequality in (6.12) is equivalent to

$$-\boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{\mu} - \boldsymbol{\mu}^*) \geq 0, \qquad \boldsymbol{\mu} \in \mathbb{R}^m_+, \tag{6.13}$$

for the given $\boldsymbol{\mu}^* \in \mathbb{R}^m_+$. This variational inequality is equivalent to stating that

$$\mathbf{0}^m \geq \boldsymbol{g}(\boldsymbol{x}^*) \perp \boldsymbol{\mu}^* \geq \mathbf{0}^m, \tag{6.14}$$

where $\perp$ denotes orthogonality: that is, for any vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{a} \perp \boldsymbol{b}$ means that $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b} = 0$. Because of the sign restrictions posed on $\boldsymbol{\mu}$ and $\boldsymbol{g}$, that is, the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, the relation $\boldsymbol{a} \perp \boldsymbol{b}$ actually means that not only does it hold that $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b} = 0$ but in fact $a_i \cdot b_i = 0$ must hold for all $i = 1, \dots, n$.[1] This complementarity system is, again, for the given $\boldsymbol{\mu}^* \in \mathbb{R}^m_+$, the same as (6.11a), (6.11c) and (6.11d). The second inequality in (6.12) is equivalent to (6.11b). ■

The above two results also state that the set of primal–dual solutions $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ is a Cartesian product set, that is, that every primal optimal solution $\boldsymbol{x}^*$ can be obtained through the system (6.11b) given any dual optimal solution $\boldsymbol{\mu}^*$, and vice versa.

[1] We establish the equivalence between (6.13) and (6.14) as follows. (Notice that the result is an extension to that of the optimality condition of the line search problem from 1 variable to m variables, except for the fact that the first is for a minimization problem while we here are dealing with a maximization problem; the proof is in fact a natural extension of that given in a footnote in Section 11.3.1.)

First, suppose that (6.14) is fulfilled. Then, $-\boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{\mu} - \boldsymbol{\mu}^*) = -\boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{\mu} \geq 0$, for all $\boldsymbol{\mu} \geq \mathbf{0}^m$, that is, (6.13) is fulfilled. Conversely, suppose that (6.13) is fulfilled. Setting $\boldsymbol{\mu} = \mathbf{0}^m$ yields that $\boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{\mu}^* \geq 0$. On the other hand, the choice $\boldsymbol{\mu} = 2\boldsymbol{\mu}^*$ yields that $-\boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{\mu}^* \geq 0$. Hence, $\boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{\mu}^* = 0$ holds. Last, let $\boldsymbol{\mu} = \boldsymbol{\mu}^* + \boldsymbol{e}_i$, where $\boldsymbol{e}_i$ is the ith unit vector in $\mathbb{R}^m$. Then, $-\boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{\mu} - \boldsymbol{\mu}^*) = -g_i(\boldsymbol{x}^*) \geq 0$. Since this is true for all $i \in \{1, 2, \dots, m\}$ we have obtained that $-\boldsymbol{g}(\boldsymbol{x}^*) \geq \mathbf{0}^m$, that is, $\boldsymbol{g}(\boldsymbol{x}^*) \leq \mathbf{0}^m$. We are done.

We note that structurally similar results to the above two propositions which are valid for the general problem (6.4) with any size of the duality gap can be found in [LaP05].[2]

We note finally a *practical* connection between the KKT system (5.9) and the above system (6.11). The practical use of the KKT system is normally to investigate whether a primal vector $\boldsymbol{x}$—obtained perhaps from a solver for our problem—is a candidate for a locally optimal solution; in other words, we have access to $\boldsymbol{x}$ and generate a vector $\boldsymbol{\mu}$ of Lagrange multipliers in the investigation of the KKT system (5.9). In contrast, the system (6.11) is normally investigated in the reverse order; we formulate and solve the Lagrangian dual problem, thereby obtaining an optimal dual vector $\boldsymbol{\mu}$. Starting from that vector, we investigate the global optimality conditions stated in (6.11) to obtain, if possible, an optimal primal vector $\boldsymbol{x}$. In the section to follow, we show when this is possible, and provide strong connections between the systems (5.9) and (6.11) in the convex and differentiable case.

6.2.3 Strong duality for convex programs

So far the results have been rather non-technical to achieve: the convexity of the Lagrangian dual problem comes with very few assumptions on the original, primal problem, and the characterization of the primal–dual set of optimal solutions is simple and also quite easily established. In order to establish *strong duality*, that is, to establish sufficient conditions under which there is no duality gap, however takes much more. In particular, as is the case with the KKT conditions we need regularity conditions (that is, constraint qualifications), and we also need to utilize separation theorems such as Theorem 4.28. Most importantly, however, is that strong duality is deeply associated with the convexity of the original problem, and it is in particular under convexity that the primal and dual optimal solutions are linked through the global optimality conditions provided in the previous section. We begin by concentrating on the inequality constrained case, proving this result in detail. We will also specialize the result to quadratic and linear optimization problems.

Consider the inequality constrained *convex* program (6.4), where $f : \mathbb{R}^n \to \mathbb{R}$ and g_i $(i = 1, \dots, m)$ are convex functions and $X \subseteq \mathbb{R}^n$ is a convex set. For this problem, we introduce the following regularity

[2] The system (6.11) is there appended with two relaxation parameters which measure, respectively, the near-optimality of $\boldsymbol{x}^*$ in the Lagrangian subproblem [that is, the ε-optimality in (6.11b)], and the violation of the complementarity conditions (6.11d). The saddle point condition (6.12) is similarly perturbed, and at an optimal solution, the sum of these two parameters equals the duality gap.

condition, due to Slater (cf. Definition 5.38): that

$$\exists \boldsymbol{x} \in X \text{ with } \boldsymbol{g}(\boldsymbol{x}) < \mathbf{0}^m. \tag{6.15}$$

Theorem 6.8 (Strong Duality, inequality constrained convex programs) *Suppose that* (6.5) *and Slater's constraint qualification* (6.15) *hold for the convex problem* (6.4).

(a) *There is no duality gap and there exists at least one Lagrange multiplier* $\boldsymbol{\mu}^*$. *Moreover, the set of Lagrange multipliers is bounded and convex.*

(b) *If the infimum in* (6.4) *is attained at some* $\boldsymbol{x}^*$, *then the pair* $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ *satisfies the global optimality conditions* (6.11).

(c) *If further* f *and* $\boldsymbol{g}$ *are differentiable at* $\boldsymbol{x}^*$, *then the condition* (6.11b) *can equivalently be written as the variational inequality*

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\mu}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0, \qquad \boldsymbol{x} \in X. \tag{6.16}$$

If, in addition, X *is open (such as is the case when* $X = \mathbb{R}^n$*), then this reduces to the condition that*

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\mu}^*) = \nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(\boldsymbol{x}^*) = \mathbf{0}^n, \tag{6.17}$$

and the global optimality conditions (6.11) *reduce to the Karush–Kuhn–Tucker conditions stated in Theorem 5.25.*

Proof. (a) We begin by establishing the existence of a Lagrange multiplier (and the presence of a zero duality gap).[3]

First, we consider the following subset of $\mathbb{R}^{m+1}$:

$$A := \{(z_1, \dots, z_m, w)^{\mathrm{T}} \mid \exists \boldsymbol{x} \in X \text{ with } g_i(\boldsymbol{x}) \leq z_i,\ i = 1, \dots, m;\ f(\boldsymbol{x}) \leq w\}.$$

It is elementary to show that A is convex.

Next, we observe that $((\mathbf{0}^m)^{\mathrm{T}}, f^*)^{\mathrm{T}}$ is not an interior point of A; otherwise, for some $\varepsilon > 0$ the point $((\mathbf{0}^m)^{\mathrm{T}}, f^* - \varepsilon)^{\mathrm{T}} \in A$ holds, which would contradict the definition of f^*. Therefore, by the (possibly non-proper) separation result in Theorem 4.28, we can find a hyperplane passing through $((\mathbf{0}^m)^{\mathrm{T}}, f^*)^{\mathrm{T}}$ such that A lies in one of the two corresponding halfspaces. In particular, there then exists a vector $(\boldsymbol{\mu}^{\mathrm{T}}, \beta)^{\mathrm{T}} \neq ((\mathbf{0}^m)^{\mathrm{T}}, 0)^{\mathrm{T}}$ such that

$$\beta f^* \leq \beta w + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{z}, \qquad (\boldsymbol{z}^{\mathrm{T}}, w)^{\mathrm{T}} \in A. \tag{6.18}$$

[3] This result is Proposition [Ber99, 5.3.1], whose proof we also utilize.

This implies that

$$\beta \geq 0; \qquad \boldsymbol{\mu} \geq \mathbf{0}^m, \tag{6.19}$$

since we have for each $(\boldsymbol{z}^{\mathrm{T}}, w)^{\mathrm{T}} \in A$ that $(\boldsymbol{z}^{\mathrm{T}}, w+\gamma)^{\mathrm{T}} \in A$ and $(z_1, \dots, z_{i-1}, z_i + \gamma, z_{i+1}, \dots, z_m, w)^{\mathrm{T}} \in A$ for all $\gamma > 0$ and $i = 1, \dots, m$.

We claim that $\beta > 0$ in fact holds. Indeed, if it was not the case, then $\beta = 0$ and (6.18) then implies that $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{z} \geq 0$ for every pair $(\boldsymbol{z}^{\mathrm{T}}, w)^{\mathrm{T}} \in A$. But since $(\boldsymbol{g}(\bar{\boldsymbol{x}})^{\mathrm{T}}, 0)^{\mathrm{T}} \in A$ [where $\bar{\boldsymbol{x}}$ is such that it satisfies the Slater condition (6.15)], we would obtain that $0 \leq \sum_{i=1}^m \mu_i g_i(\bar{\boldsymbol{x}})$ which in view of $\boldsymbol{\mu} \geq \mathbf{0}^m$ [cf. (6.19)] and the assumption that $\bar{\boldsymbol{x}}$ satisfies the Slater condition (6.15) implies that $\boldsymbol{\mu} = \mathbf{0}^m$. This means, however, that $(\boldsymbol{\mu}^{\mathrm{T}}, \beta)^{\mathrm{T}} = ((\mathbf{0}^m)^{\mathrm{T}}, 0)^{\mathrm{T}}$, arriving at a contradiction. We may therefore claim that $\beta > 0$. We further, without any loss of generality, assume that $\beta = 1$.

Thus, since $(\boldsymbol{g}(\boldsymbol{x})^{\mathrm{T}}, f(\boldsymbol{x}))^{\mathrm{T}} \in A$ for every $\boldsymbol{x} \in X$, (6.18) yields that

$$f^* \leq f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{x}), \qquad \boldsymbol{x} \in X.$$

Taking the infimum over $\boldsymbol{x} \in X$ and using the fact that $\boldsymbol{\mu} \geq \mathbf{0}^m$ we obtain

$$f^* \leq \operatorname*{infimum}_{\boldsymbol{x} \in X} \{ f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{x}) \} = q(\boldsymbol{\mu}) \leq \operatorname*{supremum}_{\boldsymbol{\mu} \geq \mathbf{0}^m} q(\boldsymbol{\mu}) = q^*.$$

Using the Weak Duality Theorem 6.4 it follows that $\boldsymbol{\mu}$ is a Lagrange multiplier vector, and there is no duality gap. This part of the proof is now done.

Take any vector $\bar{\boldsymbol{x}} \in X$ satisfying (6.15). By the definition of a Lagrange multiplier, $f^* \leq L(\bar{\boldsymbol{x}}, \boldsymbol{\mu}^*)$ holds, which implies that

$$\sum_{i=1}^m \mu_i^* \leq \frac{[f(\bar{\boldsymbol{x}}) - f^*]}{\min_{i=1,\dots,m}\{-g_i(\bar{\boldsymbol{x}})\}}.$$

By the non-negativity of $\boldsymbol{\mu}^*$, boundedness follows. As by Proposition 6.5(a) the set of Lagrange multipliers is the same as the set of optimal solutions to the dual problem (6.9), convexity follows from the identification of the dual solution set with the set of vectors $\boldsymbol{\mu} \in \mathbb{R}^m_+$ for which

$$q(\boldsymbol{\mu}) \geq q^*$$

holds. This is the upper level set for q at the level q^*; this set is convex, by the concavity of q (cf. Theorem 6.3 and Proposition 3.44).

(b) The result follows from Theorem 6.6.

(c) The first part follows from Theorem 4.23, as the Lagrangian function $L(\cdot, \boldsymbol{\mu}^*)$ is convex. The second part follows by identification. ∎

Consider next the extension of the inequality constrained convex program (6.4) in which we seek to find

$$f^* := \underset{\boldsymbol{x}}{\text{infimum}}\; f(\boldsymbol{x}), \tag{6.20a}$$
$$\text{subject to } \boldsymbol{x} \in X, \tag{6.20b}$$
$$g_i(\boldsymbol{x}) \leq 0, \qquad i = 1, \dots, m, \tag{6.20c}$$
$$\boldsymbol{\varepsilon}_j^{\mathrm{T}} \boldsymbol{x} - d_j = 0, \qquad j = 1, \dots, \ell, \tag{6.20d}$$

under the same conditions as stated following (6.4), and where $\boldsymbol{\varepsilon}_j \in \mathbb{R}^n$ $(j = 1, \dots, \ell)$. For this problem, we replace the Slater condition (6.15) with the following (cf. [BSS93, Theorem 6.2.4]):

$$\exists \boldsymbol{x} \in X \text{ with } \boldsymbol{g}(\boldsymbol{x}) < \boldsymbol{0}^m \text{ and } \boldsymbol{0}^m \in \operatorname{int} \{\, \boldsymbol{E}\boldsymbol{x} - \boldsymbol{d} \mid \boldsymbol{x} \in X \,\}, \tag{6.21}$$

where $\boldsymbol{E}$ is the $m \times n$ matrix with rows $\boldsymbol{\varepsilon}_j^{\mathrm{T}}$ and $\boldsymbol{d} = (d_j)_{j \in \{1,\dots,\ell\}} \in \mathbb{R}^m$.

Note that in the statement (6.21), the "int" can be stricken whenever X is polyhedral, so that the latter part simply states that $\boldsymbol{E}\boldsymbol{x} = \boldsymbol{d}$.

For this problem, the Lagrangian dual problem is to find

$$q^* := \underset{(\boldsymbol{\mu},\boldsymbol{\lambda})}{\text{supremum}}\; q(\boldsymbol{\mu}, \boldsymbol{\lambda}), \tag{6.22a}$$
$$\text{subject to } \boldsymbol{\mu} \geq \boldsymbol{0}^m, \tag{6.22b}$$

where

$$q(\boldsymbol{\mu}, \boldsymbol{\lambda}) := \underset{\boldsymbol{x}}{\text{infimum}}\; L(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{\lambda}^{\mathrm{T}} (\boldsymbol{E}\boldsymbol{x} - \boldsymbol{d}), \tag{6.23a}$$
$$\text{subject to } \boldsymbol{x} \in X. \tag{6.23b}$$

Theorem 6.9 (Strong Duality, general convex programs) *Suppose that in addition to* (6.5), *Slater's constraint qualification* (6.21) *holds for the problem* (6.4).

(a) *There is no duality gap and there exists at least one Lagrange multiplier pair* $(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$.

(b) *If the infimum in* (6.20) *is attained at some* $\boldsymbol{x}^*$, *then the triple* $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ *satisfies the global optimality conditions*

$$\boldsymbol{\mu}^* \geq \boldsymbol{0}^m, \quad \textit{(Dual feasibility)} \tag{6.24a}$$
$$\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*), \quad \textit{(Lagrangian optimality)} \tag{6.24b}$$
$$\boldsymbol{x}^* \in X,\ \boldsymbol{g}(\boldsymbol{x}^*) \leq \boldsymbol{0}^m,\ \boldsymbol{E}\boldsymbol{x}^* = \boldsymbol{d}, \quad \textit{(Primal feasibility)} \tag{6.24c}$$
$$\mu_i^* g_i(\boldsymbol{x}^*) = 0,\ i = 1, \dots, m. \quad \textit{(Complementary slackness)} \tag{6.24d}$$

(c) *If further* f *and* $\boldsymbol{g}$ *are differentiable at* $\boldsymbol{x}^*$, *then the condition* (6.24b) *can equivalently be written as*

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0, \qquad \boldsymbol{x} \in X. \tag{6.25}$$

If, in addition, X *is open (such as is the case when* $X = \mathbb{R}^n$*), then this reduces to the condition that*

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) = \nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(\boldsymbol{x}^*) + \sum_{j=1}^{\ell} \lambda_j^* \boldsymbol{\varepsilon}_j = \boldsymbol{0}^n, \tag{6.26}$$

and the global optimality conditions (6.24) *reduce to the Karush–Kuhn–Tucker conditions stated in Theorem 5.33.*

Proof. The proof is similar to that of Theorem 6.8. ∎

We finally consider a special case where automatically a regularity condition holds.

Consider the linearly constrained convex program to find

$$f^* := \operatorname*{infimum}_{\boldsymbol{x}} f(\boldsymbol{x}), \tag{6.27a}$$

$$\text{subject to } \boldsymbol{x} \in X, \tag{6.27b}$$

$$\boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{x} - b_i \leq 0, \qquad i = 1, \dots, m, \tag{6.27c}$$

$$\boldsymbol{\varepsilon}_j^{\mathrm{T}} \boldsymbol{x} - d_j = 0, \qquad j = 1, \dots, \ell, \tag{6.27d}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and $X \subseteq \mathbb{R}^n$ is polyhedral.

Theorem 6.10 (Strong Duality, linear constraints) *If* (6.5) *holds for the problem* (6.27), *then there is no duality gap and there exists at least one Lagrange multiplier.*

Proof. Again, the proof is similar to that of Theorem 6.8, except that no additional regularity conditions are needed.[4] ∎

The existence of a multiplier [which by Proposition 6.5 and the absence of a duality gap implies the existence of an optimal solution to the dual problem (6.9)] does not imply the existence of an optimal solution to the primal problem (6.27) without any additional assumptions (take the minimization of $f(x) := 1/x$ over $x \geq 1$ for example). However, when f is either weakly coercive, quadratic or linear, the existence results are stronger; see the primal existence results in Theorems 4.6, 4.7, and 6.11 below, for example.

[4] For a detailed proof, see [Ber99, Proposition 5.2.1]. (The special case where f is moreover differentiable is covered in [Ber99, Proposition 3.4.2].)

For convex programs, the Lagrange multipliers defined in this section, and those that appear in the Karush–Kuhn–Tucker conditions, are identical. We establish this result for two classes of convex programs below.

Next, we specialize the above to linear and quadratic programs.

6.2.4 Strong duality for linear and quadratic programs

The following result will be established and analyzed in detail in Chapter 10 on linear programming duality (cf. Theorem 10.6), but can in fact also be established similarly to above. Its proof will however be relegated to that of Theorem 10.6.

Theorem 6.11 (Strong Duality, linear programs) *Assume, in addition to the conditions of Theorem* 6.10, *that f is linear, so that* (6.27) *is a linear program. Then, the primal and dual problems have optimal solutions and there is no duality gap.*

Proof. The result follows by applying Farkas' Lemma 3.30 and the Weak Duality Theorem 6.4, or by analyzing an optimal Simplex tableau. Detailed proofs are found in [BSS93, Theorem 2.7.3] or [Ber99, Proposition 5.2.2], for example. ∎

The above result states a strong duality result for a general linear program. The dual problem is however not explicit. We next develop an explicit Lagrangian dual problem for a linear program. Again, more details on this problem will be covered later.

Consider the linear program to

$$\operatorname*{minimize}_{\boldsymbol{x}} \ \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}, \tag{6.28a}$$

$$\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{6.28b}$$

$$\boldsymbol{x} \geq \boldsymbol{0}^n, \tag{6.28c}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{c} \in \mathbb{R}^n$, and $\boldsymbol{b} \in \mathbb{R}^m$. If we let $X := \mathbb{R}^n_+$, then the Lagrangian dual problem is to

$$\operatorname*{maximize}_{\boldsymbol{\lambda}\in\mathbb{R}^m} \ \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\lambda}, \tag{6.29a}$$

$$\text{subject to } \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} \leq \boldsymbol{c}. \tag{6.29b}$$

The reason why we can write it in this form is that

$$q(\boldsymbol{\lambda}) := \operatorname*{infimum}_{\boldsymbol{x}\geq\boldsymbol{0}^n} \left\{ \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) \right\} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\lambda} + \operatorname*{infimum}_{\boldsymbol{x}\geq\boldsymbol{0}^n} (\boldsymbol{c} - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda})^{\mathrm{T}}\boldsymbol{x},$$

so that

$$q(\boldsymbol{\lambda}) = \begin{cases} \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\lambda}, & \text{if } \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} \leq \boldsymbol{c}, \\ -\infty, & \text{otherwise.} \end{cases}$$

(The infimum is attained at zero if and only if these inequalities are satisfied; otherwise, the inner problem is unbounded below.)

Further, why is it that $\boldsymbol{\lambda}$ here is not restricted in sign? Suppose we were to split the system $\boldsymbol{Ax} = \boldsymbol{b}$ into an inequality system of the form

$$\begin{aligned} \boldsymbol{Ax} &\leq \boldsymbol{b}, \\ -\boldsymbol{Ax} &\leq -\boldsymbol{b}. \end{aligned}$$

Let

$$\begin{pmatrix} \boldsymbol{\mu}^+ \\ \boldsymbol{\mu}^- \end{pmatrix}$$

be the corresponding vector of multipliers, and take the Lagrangian dual for this formulation. Then, we would have a Lagrange function of the form

$$(\boldsymbol{x}, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-) \mapsto L(\boldsymbol{x}, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-) := \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + (\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-)^{\mathrm{T}}(\boldsymbol{b} - \boldsymbol{Ax}),$$

and since $\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-$ can take on any value in $\mathbb{R}^m$ we can simply replace it with the unrestricted vector $\boldsymbol{\lambda} \in \mathbb{R}^m$. This is what has been done above, and it motivates why the multiplier for an equality constraint never is sign restricted; the same was the case, as we saw in Section 5.6, for the multipliers in the KKT conditions.

As applied to this problem, Theorem 6.11 states that if both the primal or dual problems have feasible solutions, then they both have optimal solutions, satisfying strong duality ($\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}^* = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\lambda}^*$). On the other hand, if any of the two problems has an unbounded solution, then the other problem is infeasible.

Consider next the quadratic programming problem to

$$\underset{\boldsymbol{x}}{\text{minimize}} \left\{ \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Qx} + \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \right\}, \tag{6.30a}$$

$$\text{subject to } \boldsymbol{Ax} \leq \boldsymbol{b}, \tag{6.30b}$$

where $\boldsymbol{Q}$ is a positive definite $n \times n$ matrix. We develop an explicit dual problem under this assumption on $\boldsymbol{Q}$.

Lagrangian relaxing the inequality constraints, we obtain that the inner problem in $\boldsymbol{x}$ is solved by letting

$$\boldsymbol{x} = -\boldsymbol{Q}^{-1}(\boldsymbol{c} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu}). \tag{6.31}$$

Substituting this expression into the Lagrangian function yields the Lagrangian dual problem to

$$\underset{\boldsymbol{\mu}}{\text{maximize}} \left\{ -\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu} - (\boldsymbol{b} + \boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{c})^{\mathrm{T}}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{Q}^{-1}\boldsymbol{c} \right\}, \tag{6.32a}$$

$$\text{subject to } \boldsymbol{\mu} \geq \boldsymbol{0}^m, \tag{6.32b}$$

Strong duality follows for this convex primal–dual pair of quadratic programs, in much the same way as for linear programming.

Proposition 6.12 (Strong Duality, quadratic programs) *For the primal–dual pair of convex quadratic programs* (6.30), (6.32), *the following holds:*

(a) *If both problems have feasible solutions, then both problems also have optimal solutions, and the primal problem* (6.30) *also has a unique optimal solution, given by* (6.31) *for any optimal Lagrange multiplier, and in the two problems the optimal values are equal.*

(b) *If either of the two problems has an unbounded solution, then the other one is infeasible.*

(c) *Suppose that $\boldsymbol{Q}$ is positive semi-definite, and that* (6.5) *holds. Then, both the problem* (6.30) *and its Lagrangian dual have nonempty, closed and convex sets of optimal solutions, and their optimal values are equal.* ■

In the result (a) it is important to note that the Lagrangian dual problem (6.32) is not necessarily strictly convex; the matrix $\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}$ need not be positive definite, especially so when $\boldsymbol{A}$ does not have full rank. The result (c) extends the strong duality result from linear programming, since $\boldsymbol{Q}$ in (c) can be the zero matrix. In the case of (c) we of course cannot write the Lagrangian dual problem in the form of (6.32) because $\boldsymbol{Q}$ is not invertible.

6.3 Illustrative examples

6.3.1 Two numerical examples

Example 6.13 (an explicit, differentiable dual problem) Consider the problem to

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & f(\boldsymbol{x}) := x_1^2 + x_2^2, \\ \text{subject to} \quad & x_1 + x_2 \geq 4, \\ & x_j \geq 0, \qquad j = 1, 2. \end{aligned}$$

We consider the first constraint to be the complicated one, and hence define $g(\boldsymbol{x}) := -x_1 - x_2 + 4$ and let $X := \{ (x_1, x_2) \mid x_j \geq 0,\ j = 1, 2 \}$.

Then, the Lagrangian dual function is

$$\begin{aligned} q(\mu) &= \underset{\boldsymbol{x} \in X}{\operatorname{minimum}}\ L(\boldsymbol{x}, \mu) := f(\boldsymbol{x}) - \mu(x_1 + x_2 - 4) \\ &= 4\mu + \underset{\boldsymbol{x} \in X}{\operatorname{minimum}}\ \{x_1^2 + x_2^2 - \mu x_1 - \mu x_2\} \\ &= 4\mu + \underset{x_1 \geq 0}{\operatorname{minimum}}\ \{x_1^2 - \mu x_1\} + \underset{x_2 \geq 0}{\operatorname{minimum}}\ \{x_2^2 - \mu x_2\}, \qquad \mu \geq 0. \end{aligned}$$

For a fixed $\mu \geq 0$, the minimum is attained at $x_1(\mu) = \frac{\mu}{2}, x_2(\mu) = \frac{\mu}{2}$.

Substituting this expression into $q(\mu)$, we obtain that $q(\mu) = f(\boldsymbol{x}(\mu)) - \mu(x_1(\mu) + x_2(\mu) - 4) = 4\mu - \frac{\mu^2}{2}$.

Note that q is strictly concave, and it is differentiable everywhere (due to the fact that f, g are differentiable and $\boldsymbol{x}(\mu)$ is unique), by Danskin's Theorem 6.16(d).

We then have that $q'(\mu) = 4 - \mu = 0 \iff \mu = 4$. As $\mu = 4 \geq 0$, it is the optimum in the dual problem! $\mu^* = 4; \boldsymbol{x}^* = (x_1(\mu^*), x_2(\mu^*))^{\mathrm{T}} = (2, 2)^{\mathrm{T}}$.

Also: $f(\boldsymbol{x}^*) = q(\mu^*) = 8$.

This is an example where the dual function is differentiable, and therefore we can utilize Proposition 6.25(c). In this case, the optimum $\boldsymbol{x}^*$ is also unique, so it is automatically given as $\boldsymbol{x}^* = \boldsymbol{x}(\mu)$. ■

Example 6.14 (an implicit, non-differentiable dual problem) Consider the linear programming problem to

$$\begin{aligned} \underset{\boldsymbol{x}}{\operatorname{minimize}}\ & f(\boldsymbol{x}) := -x_1 - x_2, \\ \text{subject to}\ & 2x_1 + 2x_2 \leq 3, \\ & 0 \leq x_1 \leq 2, \\ & 0 \leq x_2 \leq 1. \end{aligned}$$

Check that the optimal solution is that $\boldsymbol{x}^* = (3/2, 0)^{\mathrm{T}}, f(\boldsymbol{x}^*) = -3/2$.

Consider Lagrangian relaxing the first constraint, obtaining

$$\begin{aligned} L(\boldsymbol{x}, \mu) &= -x_1 - x_2 + \mu(2x_1 + 4x_2 - 3); \\ q(\mu) &= -3\mu + \underset{0 \leq x_1 \leq 2}{\operatorname{minimum}}\ \{(-1 + 2\mu)x_1\} + \underset{0 \leq x_2 \leq 1}{\operatorname{minimum}}\ \{(-1 + 4\mu)x_2\} \\ &= \begin{cases} -3 + 5\mu, & 0 \leq \mu \leq 1/4, \\ -2 + \ \mu, & 1/4 \leq \mu \leq 1/2, \\ \quad - 3\mu, & 1/2 \leq \mu. \end{cases} \end{aligned}$$

Check that $\mu^* = 1/2$, and hence that $q(\mu^*) = -3/2$. For linear programs, we have strong duality, but how do we obtain the optimal primal solution from μ^*? It is clear that q is non-differentiable at μ^*. Let us utilize the characterization given in the system (6.11).

First, at μ^*, it is clear that $X(\mu^*)$ is the set $\{\, \binom{2\alpha}{0} \mid 0 \leq \alpha \leq 1 \,\}$. Among the subproblem solutions, we next have to find one that is primal feasible as well as complementary.

Primal feasibility means that $2 \cdot 2\alpha + 2 \cdot 0 \leq 3 \iff \alpha \leq 3/4$.

Further, complementarity means that $\mu^* \cdot (2x_1^* + 4x_2^* - 3) = 0 \iff \alpha = 3/4$, since $\mu^* \neq 0$. We conclude that the only primal vector $\boldsymbol{x}$ that satisfies the system (6.11) together with the dual optimal solution $\mu^* = 1/2$ is $\boldsymbol{x}^* = (3/2, 0)^{\mathrm{T}}$. Check finally that $f^* = q^*$. ∎

In the first example, the Lagrangian dual function is differentiable since $\boldsymbol{x}(\mu)$ is unique. The second one shows that otherwise, there may be kinks in the function q where there are alternative solutions $\boldsymbol{x}(\mu)$; as a result, to obtain a primal optimal solution becomes more complex. The Dantzig–Wolfe algorithm, for example, represents a means by which to automatize the process that we have just shown; the algorithm generates extreme points of $X(\boldsymbol{\mu})$ algorithmically, and constructs the best feasible convex combination thereof, obtaining a primal–dual optimal solution in a finite number of iterations for linear programs.

6.3.2 An application to combinatorial optimization

Lagrangian relaxation has shown to be remarkably efficient for some combinatorial optimization problems. This is surprising when taking into account that such problems are integer or mixed-integer problems, which suffer from non-zero duality gaps in general. What then lies behind their popularity?

- One can show that Lagrangian relaxation of an integer program is always at least as good as that of a *continuous relaxation*[5] (in the sense that f_R is higher for Lagrangian relaxation than a continuous relaxation);
- Together with heuristics for finding primal feasible solution, surprisingly good feasible solutions are often found;
- The Lagrangian relaxed problems can be made computationally much simpler than the original problem, while still keeping a lot of the structure of the original problem.

[5] The continuous relaxation amounts to removing the integrality conditions, replacing, for example, $x_j \in \{0, 1\}$ by $x_j \in [0, 1]$.

The traveling salesman problem We provide an example, taken from an application of the traveling salesman problem.

Let c_{ij} denote the distance from city i to city j, with $i < j$, and $i, j \in \mathcal{N} = \{1, 2, \dots, n\}$, and

$$x_{ij} = \begin{cases} 1, & \text{if link } (i,j) \text{ is part of the TSP tour,} \\ 0, & \text{otherwise.} \end{cases}$$

With these definitions, the complete, undirected traveling salesman problem (TSP) is to

$$\begin{aligned}
\underset{\boldsymbol{x}}{\text{minimize}} \quad & \sum_{i=1}^{n} \sum_{j=1:j\neq i}^{n} c_{ij} x_{ij}, && && \text{(6.33a)} \\
\text{subject to} \quad & \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} x_{ij} \leq |\mathcal{S}| - 1, && \mathcal{S} \subset \mathcal{N}, && \text{(6.33b)} \\
& \sum_{i=1}^{n} \sum_{j=1:j\neq i}^{n} x_{ij} = n, && && \text{(6.33c)} \\
& \sum_{i=1}^{n} x_{ij} = 2, && j \in \mathcal{N}, && \text{(6.33d)} \\
& x_{ij} \in \{0, 1\}, && i, j \in \mathcal{N}. && \text{(6.33e)}
\end{aligned}$$

The constraints have the following interpretation: (6.33b) implies that there can be no *sub-tours*, that is, a tour where fewer than n cities are visited (that is, if $\mathcal{S} \subset \mathcal{N}$ then there can be at most $|\mathcal{S}| - 1$ links between nodes in the set $\mathcal{S}$, where $|\mathcal{S}|$ is the cardinality–number of members of–the set $\mathcal{S}$); (6.33c) implies that in total n cities must be visited; and (6.33d) implies that each city is connected to two others, such that we make sure to arrive from one city and leave for the next.

This problem is NP-hard, which implies that there is no known polynomial algorithm for solving it. We resort therefore to the use of relaxation techniques, in particular Lagrangian relaxation. We have more than one alternative relaxation to perform: If we Lagrangian relax the tree constraints (6.33b) and (6.33c) the remaining problem is a *2-matching* problem; it can be solved in polynomial time. If we instead Lagrangian relax the degree constraints (6.33d) for every node except for one node the remaining problem is a *1-MST* problem, that is, a special type of minimum spanning tree problem.

The following definition is classic: a *Hamiltonian path (respectively cycle)* is a path (respectively, cycle) which passes every node in the graph exactly once.

Every Hamiltonian cycle is a Hamiltonian path from a node s to another node, t, followed by a link (t, s); a subgraph which consists of a spanning tree plus an extra link such that all nodes have degree two. This is then a feasible solution to the TSP.

A *1-MST* problem is the problem to find an MST in the graph that excludes node s, followed by the addition of the two least expensive links from node s to that tree. If all nodes happen to get degree two, then the 1-MST solution is a traveling salesman tour (that is, Hamiltonian cycle). The idea behind solving the Lagrangian dual problem is then to find proper multiplier values such that the Lagrangian relaxation will produce feasible solutions.

Lagrangian relaxation of the traveling salesman problem Suppose that we Lagrangian relax the degree constraints (3), except for node 1.

The subproblem is the following: an 1-MST with the objective function (note that $\lambda_1 = 0$, corresponding to the starting node of the tour)

$$\begin{aligned} q(\boldsymbol{\lambda}) &= \operatorname*{minimum}_{\boldsymbol{x}} \sum_{i=1}^{n} \sum_{j=1:j\neq i}^{n} c_{ij}x_{ij} + \sum_{j=2}^{n} \lambda_j \Big(2 - \sum_{i=1:i\neq j}^{n} x_{ij}\Big) \\ &= 2\sum_{j=1}^{n} \lambda_j + \operatorname*{minimum}_{\boldsymbol{x}} \sum_{i=1}^{n} \sum_{j=1:j\neq i}^{n} (c_{ij} - \lambda_i - \lambda_j)x_{ij}. \end{aligned}$$

We see immediately the role of the Lagrange multipliers: a high (low) value of the multiplier λ_j makes node j attractive (unattractive) in the above 1-MST problem, and will therefore lead to more (less) links being attached to it. When solving the Lagrangian dual problem, we will use the class of subgradient optimization methods, an overview of which is found in Section 6.5; it is a kind of steepest ascent method for the problem to maximize the function q over $\mathbb{R}^n$, but where gradients are replaced by subgradients.

What is the updating step? It is as usual an update in the direction of a subgradient, that is, the direction of

$$h_i(\boldsymbol{x}(\boldsymbol{\lambda})) := 2 - \sum_{i=1:i\neq j}^{n} x_{ij}, \qquad i = 1, \dots, n,$$

where the value of $x_{ij} \in \{0, 1\}$ is the solution to the 1-MST solution with link costs $c_{ij} - \lambda_i - \lambda_j$. We see from the formula for the direction that it is such that

$$\lambda_j := \lambda_j + \alpha \left(2 - \sum_{i=1:i\neq j}^{n} x_{ij} \right), \qquad j = 2, \dots, n,$$

where $\alpha > 0$ is a step length. It is interesting to investigate what the update means:

$$\text{current degree at node } j : \begin{cases} > 2 \Longrightarrow \lambda_j \downarrow \text{(link cost } \uparrow\text{)} \\ = 2 \Longrightarrow \lambda_j - \text{(link cost constant)} \\ < 2 \Longrightarrow \lambda_j \uparrow \text{(link cost } \downarrow\text{)} \end{cases}$$

In other words, the updating formula in a subgradient method is such that the link cost in the 1-MST subproblem is shifted upwards (downwards) if there are too many (too few) links connected to node j in the 1-MST. We are hence adjusting the *node prices* of the nodes in such a way as to influence the 1-MST problem to always choose 2 links per node to connect to.

A feasibility heuristic A feasibility heuristic takes the optimal solution from the Lagrangian minimization problem over $\boldsymbol{x}$ and adjusts it such that a feasible solution to the original problem is constructed. As one cannot predict if, or when, a primal feasible solution will be found directly from the subproblem, the heuristic will provide a solution that can be used in place of an optimal one, should one not be found. Moreover, as we know from Lagrangian duality theory, we then have access to both lower and upper bounds on the optimal value f^* of the original problem, and so we have a quality measure of the feasible solutions found.

A feasibility heuristic which can be used together with our Lagrangian heuristic is as follows.

Identify a path in the 1-MST with many links. Then form a subgraph with the remaining nodes and find a path that passes all of them. Put the two paths together in the best way. The resulting path is a Hamiltonian cycle, that is, a feasible solution. We then have both a lower bound [the value of $q(\boldsymbol{\lambda})$] and an upper bound (the original cost of this heuristically constructed traveling salesman tour), and this interval can be used as a quality measure of the feasible solution at termination.

The Philips example In 1987–1988 an MSc project was performed at the department of mathematics at Linköping University, in cooperation with the company Philips, Norrköping. The project was initiated with the goal to improve the current practice of solving a production planning problem.

The problem was as follows: Philips produces circuit boards, perhaps several hundreds or thousands of the same type. There is a new batch of patterns (holes) to be drilled every day, and perhaps even several such batches per day.

In order to speed up the production process the drilling machine is connected to a microcomputer that selects the ordering of the holes to be drilled automatically, given their coordinates. The algorithm for performing the sorting used to be a simple sorting operation that found, for every fixed x-coordinate, the corresponding y-coordinates and sorted them in increasing order. The movement of the drill was therefore from left to right, and for each fixed x-coordinate the movement was vertical. The time it took to drill the holes on one circuit board was, however, far too long, simply because the drill traveled around a lot without performing any tasks, thus following a path that was too long. (On the other hand, the actual ordering was very fast to produce!) All in all, though, the complete batch production took too long because of the poorly planned drill movement.

At the beginning of the project it was observed that the production planning problem actually is a traveling salesman problem, where the cities are the holes to be drilled, and the distances between them correspond to the Euclidean distances between them. Therefore, an efficient TSP heuristic was devised and implemented, for use in conjunction with the microcomputer. In fact, it was based on precisely the above Lagrangian relaxation, a subgradient optimization method, and a graph-search type heuristic of the form discussed above.

A typical run with the algorithm took a few minutes, and was always stopped after a fixed number of subgradient iterations; the generation of feasible solutions with the above-mentioned graph search technique was performed at every Kth iteration, where K was set to a value strictly larger than one. (Moreover, feasible solutions were not generated during the first iterations of the dual procedure, because of the poor quality of $\boldsymbol{\lambda}_k$ for low values of k; it is often the case that the traveling salesman tour resulting from the heuristic is better when the multipliers are near-optimal in the Lagrangian dual problem.)

In one of the examples implemented it was found that the optimal path length was in the order to 2 meters, and that the upper and lower bounds produced lead to the conclusion that the relative error of the path length of the best feasible solution found was *less than 7 %*, a quite good result, also showing that the duality gap for the problem at hand (together with the Lagrangian relaxation chosen) is quite small.

After implementing the new procedure, Philips could report an increase in production by some 70 %. Hence, the slightly longer time it took to produce the production plan, that is, the traveling salesman tour for the drill to follow, was more than well compensated by the fact that the drilling could be done much faster.

Here is hence an interesting case where Lagrangian relaxation helped to solve a large-scale, complex and difficult problem by utilizing problem structure.

6.4 *Differentiability properties of the dual function

We have established that the Lagrangian dual problem (6.9) is a convex one, and further that under some circumstances we can generate a dual optimal solution that has the same objective value q^* as the optimal value of the original problem f^*. We now turn to study the Lagrangian dual problem in detail, and in particular how it can be solved efficiently. First, we will establish when the dual function q is differentiable. We will see that differentiability holds only in some special cases, in which we can recognize the workings of the so-called *Lagrange multiplier method*. In practice, the function q will be non-differentiable, and then this classic method will fail. This means that we must devise a more general numerical method which is not based on gradients but rather *subgradients*. This type of algorithm is the topic of the next section, while we here begin by studying the topic of subgradients of convex functions in general.

6.4.1 Sub-differentiability of convex functions

Throughout this section we suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, and study its sub-differentiability properties. We will later on apply our findings to the Lagrangian dual function q, or, rather, its negative $-q$. We first remark that a finite convex function is automatically continuous (cf. Theorem 4.26), in fact even Lipschitz continuous on every bounded subset of $\mathbb{R}^n$.

Definition 6.15 (subgradient) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. We say that a vector $\boldsymbol{p} \in \mathbb{R}^n$ is a* subgradient *of f at $\boldsymbol{x} \in \mathbb{R}^n$ if*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{p}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}), \qquad \boldsymbol{y} \in \mathbb{R}^n. \tag{6.34}$$

The set of such vectors $\boldsymbol{p}$ defines the subdifferential *of f at $\boldsymbol{x}$, and is denoted $\partial f(\boldsymbol{x})$.* ∎

Notice the connection to the characterization of a convex function in C^1 in Theorem 3.40(a). The difference between them is that $\boldsymbol{p}$ is not unique at a non-differentiable point. (Just as the gradient has a role in supporting hyperplanes to the graph of a convex function in C^1, the role of the subgradients is the same; at a non-differentiable point there are more then one supporting hyperplane to the graph of f.)

We illustrate this in Figure 6.1.

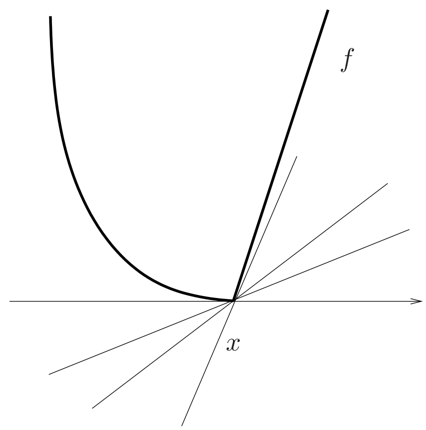


Figure 6.1: Three possible slopes of the convex function f at x.

Notice also that a global minimum $\boldsymbol{x}^*$ of f over $\mathbb{R}^n$ is characterized by the inclusion that $\mathbf{0}^n \in \partial f(\boldsymbol{x}^*)$, and recognize, again, the similarity to the C^1 case.

We list some additional basic results for convex functions next. Proofs will not be given here, we refer to the convex analysis text by Rockafellar [Roc70].

Proposition 6.16 (properties of a convex function) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function.*

(a) [boundedness of $\partial f(\boldsymbol{x})$] *For every $\boldsymbol{x} \in \mathbb{R}^n$, $\partial f(\boldsymbol{x})$ is a nonempty, convex, and compact set. If X is bounded then $\cup_{\boldsymbol{x}\in X}\, \partial f(\boldsymbol{x})$ is bounded.*

(b) [closedness of ∂f] *The subdifferential mapping $\boldsymbol{x} \rightrightarrows \partial f(\boldsymbol{x})$ is* closed*; in other words, if $\{\boldsymbol{x}_k\}$ is a sequence of vectors in $\mathbb{R}^n$ converging to $\boldsymbol{x}$, and $\boldsymbol{p}_k \in \partial f(\boldsymbol{x}_k)$ holds for every k, then the sequence $\{\boldsymbol{p}_k\}$ of subgradients is bounded and every limit point thereof belongs to $\partial f(\boldsymbol{x})$.*

(c) [directional derivative and differentiability] *For every $\boldsymbol{x} \in \mathbb{R}^n$, the directional derivative of f at $\boldsymbol{x}$ in the direction of $\boldsymbol{d} \in \mathbb{R}^n$ satisfies*

$$f'(\boldsymbol{x}; \boldsymbol{d}) = \underset{\boldsymbol{p}\in\partial f(\boldsymbol{x})}{\operatorname{maximum}}\ \boldsymbol{p}^{\mathrm{T}}\boldsymbol{d}. \tag{6.35}$$

In particular, f is differentiable at $\boldsymbol{x}$ with gradient $\nabla f(\boldsymbol{x})$ if and only if it has $\nabla f(\boldsymbol{x})$ as its unique subgradient at $\boldsymbol{x}$; in that case, $f'(\boldsymbol{x}; \boldsymbol{d}) = \nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{d}$.

(d) [Danskin's Theorem—directional derivatives of a convex max function] *Let Z be a compact subset of $\mathbb{R}^m$, and let $\phi : \mathbb{R}^n \times Z \to \mathbb{R}$ be continuous and such that $\phi(\cdot, \boldsymbol{z}) : \mathbb{R}^n \to \mathbb{R}$ is convex for each $\boldsymbol{z} \in Z$. Let the function $f : \mathbb{R}^n \to \mathbb{R}$ be given by*

$$f(\boldsymbol{x}) := \underset{\boldsymbol{z} \in Z}{\operatorname{maximum}}\ \phi(\boldsymbol{x}, \boldsymbol{z}), \qquad \boldsymbol{x} \in \mathbb{R}^n. \tag{6.36}$$

The function f then is convex on $\mathbb{R}^n$ and has a directional derivative at $\boldsymbol{x}$ in the direction of $\boldsymbol{d}$ equal to

$$f'(\boldsymbol{x}; \boldsymbol{d}) := \underset{\boldsymbol{z} \in Z(\boldsymbol{x})}{\operatorname{maximum}}\ \phi'(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{d}), \tag{6.37}$$

where $\phi'(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{d})$ is the directional derivative of $\phi(\cdot, \boldsymbol{z})$ at $\boldsymbol{x}$ in the direction of $\boldsymbol{d}$, and $Z(\boldsymbol{x}) := \{\, \boldsymbol{z} \in \mathbb{R}^m \mid \phi(\boldsymbol{x}, \boldsymbol{z}) = f(\boldsymbol{x}) \,\}$.

In particular, if $Z(\boldsymbol{x})$ contains a single point $\bar{\boldsymbol{z}}$ and $\phi(\cdot, \bar{\boldsymbol{z}})$ is differentiable at $\boldsymbol{x}$, then f is differentiable at $\boldsymbol{x}$, and $\nabla f(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \phi(\boldsymbol{x}, \bar{\boldsymbol{z}})$, where $\nabla_{\boldsymbol{x}} \phi(\boldsymbol{x}, \bar{\boldsymbol{z}})$ is the vector with components $\frac{\partial \phi(\boldsymbol{x}, \bar{\boldsymbol{z}})}{\partial x_i}$, $i = 1, \dots, n$.

If further $\phi(\cdot, \boldsymbol{z})$ is differentiable for all $\boldsymbol{z} \in Z$ and $\nabla_{\boldsymbol{x}} \phi(\boldsymbol{x}, \cdot)$ is continuous on Z for each $\boldsymbol{x}$, then

$$\partial f(\boldsymbol{x}) = \operatorname{conv} \{\, \nabla_{\boldsymbol{x}} \phi(\boldsymbol{x}, \boldsymbol{z}) \mid \boldsymbol{z} \in Z(x) \,\}, \qquad \boldsymbol{x} \in \mathbb{R}^n,$$

holds.

Proof. (a) This is a special case of [Roc70, Theorem 24.7].

(b) This is [Roc70, Theorem 24.5].

(c) This is [Roc70, Theorem 23.4 and 25.1].

(d) This is [Ber99, Proposition B.25]. ∎

Figure 6.2 illustrates the subdifferential of a convex function.

6.4.2 Differentiability of the Lagrangian dual function

We consider the inequality constrained problem (6.4), where we make the following standing assumption:

$$f, g_i \, (i = 1, \dots, m) \text{ are continuous; } X \text{ is nonempty and compact.} \tag{6.38}$$

Under this assumption, the set of solutions to the Lagrangian subproblem,

$$X(\boldsymbol{\mu}) := \arg \underset{\boldsymbol{x} \in X}{\operatorname{minimum}}\ L(\boldsymbol{x}, \boldsymbol{\mu}), \qquad \boldsymbol{\mu} \in \mathbb{R}^m, \tag{6.39}$$

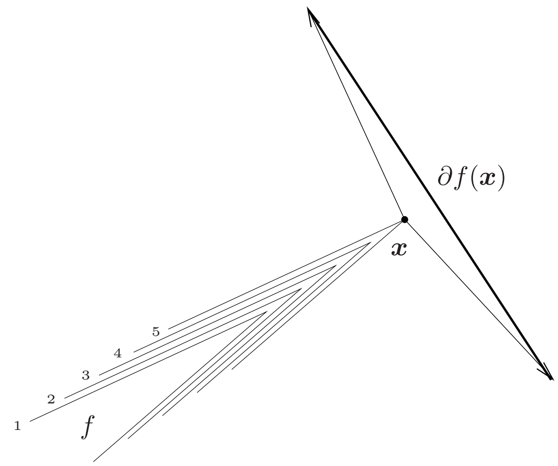


Figure 6.2: The subdifferential of a convex function f at $\boldsymbol{x}$.

is nonempty and compact for any choice of dual vector $\boldsymbol{\mu}$. We first develop the sub-differentiability properties of the associated dual function q, stated in (6.8). The first result strengthens Theorem 6.3(a) under these additional assumptions.

Proposition 6.17 (sub-differentiability of the dual function) *Suppose that, in the problem* (6.4), (6.38) *holds.*

(a) *The dual function* (6.8) *is finite, continuous and concave on* $\mathbb{R}^m$. *If its supremum over* $\mathbb{R}^m_+$ *is attained, then the optimal solution set therefore is closed and convex.*

(b) *The mapping* $\boldsymbol{\mu} \rightrightarrows X(\boldsymbol{\mu})$ *is closed on* $\mathbb{R}^m$. *If* $X(\bar{\boldsymbol{\mu}})$ *is the singleton set* $\{\bar{\boldsymbol{x}}\}$ *for some* $\bar{\boldsymbol{\mu}} \in \mathbb{R}^m$, *and for some sequence* $\mathbb{R}^m \supset \{\boldsymbol{\mu}_k\} \to \bar{\boldsymbol{\mu}}$, $\boldsymbol{x}_k \in X(\boldsymbol{\mu}_k)$ *for all* k, *then* $\{\boldsymbol{x}_k\} \to \bar{\boldsymbol{x}}$.

(c) *Let* $\boldsymbol{\mu} \in \mathbb{R}^m$. *If* $\boldsymbol{x} \in X(\boldsymbol{\mu})$, *then* $\boldsymbol{g}(\boldsymbol{x})$ *is a subgradient to* q *at* $\boldsymbol{\mu}$, *that is,* $\boldsymbol{g}(\boldsymbol{x}) \in \partial q(\boldsymbol{\mu})$.

(d) *Let* $\boldsymbol{\mu} \in \mathbb{R}^m$. *Then,*

$$\partial q(\boldsymbol{\mu}) = \operatorname{conv} \{ \boldsymbol{g}(\boldsymbol{x}) \mid \boldsymbol{x} \in X(\boldsymbol{\mu}) \}.$$

The set $\partial q(\boldsymbol{\mu})$ *is convex and compact. Moreover, if* U *is a bounded set, then* $\cup_{\boldsymbol{\mu} \in U} \partial q(\boldsymbol{\mu})$ *is also bounded.*

(e) *The directional derivative of* q *at* $\boldsymbol{\mu} \in \mathbb{R}^m$ *in the direction of* $\boldsymbol{d} \in \mathbb{R}^m$ *is*

$$q'(\boldsymbol{\mu}; \boldsymbol{d}) = \underset{\boldsymbol{\gamma} \in \partial q(\boldsymbol{\mu})}{\operatorname{minimum}} \, \boldsymbol{d}^{\mathrm{T}} \boldsymbol{\gamma}.$$

Proof. (a) Theorem 6.3(a) stated the concavity of q on its effective domain. Weierstrass' Theorem 4.6 states that q is finite on $\mathbb{R}^m$, which is then also its effective domain. The continuity of q follows from that of *any* finite concave function, as we have already stated.[6] The closedness property of the solution set complements that of Theorem 6.8(a), and is a direct consequence of the continuity of q (the upper level set then automatically is a closed set).

(b) Let $\{\boldsymbol{\mu}_k\}$ be a sequence of vectors in $\mathbb{R}^m$, and let $\boldsymbol{x}_k \in X(\boldsymbol{\mu}_k)$ be arbitrary. Let $\boldsymbol{x}$ be arbitrary in X, and let further $\bar{\boldsymbol{x}} \in X$ be an arbitrary limit point of $\{\boldsymbol{x}_k\}$ (at least exists from the compactness of X). From the property that for all k,

$$L(\boldsymbol{x}_k, \boldsymbol{\mu}_k) \leq L(\boldsymbol{x}, \boldsymbol{\mu}_k),$$

follows by the continuity of L that, in the limit of k in the subsequence in which $\{\boldsymbol{x}_k\}$ converges to $\bar{\boldsymbol{x}}$,

$$L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\mu}}) \leq L(\boldsymbol{x}, \bar{\boldsymbol{\mu}}),$$

so that $\bar{\boldsymbol{x}} \in X(\bar{\boldsymbol{\mu}})$, as desired. The special case of a singleton set $X(\bar{\boldsymbol{\mu}})$ follows.

(c) Let $\bar{\boldsymbol{\mu}} \in \mathbb{R}^m$ be arbitrary. We have that

$$\begin{aligned} q(\bar{\boldsymbol{\mu}}) &= \operatorname*{infimum}_{\boldsymbol{y} \in X} L(\boldsymbol{y}, \bar{\boldsymbol{\mu}}) \leq f(\boldsymbol{x}) + \bar{\boldsymbol{\mu}}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) \\ &= f(\boldsymbol{x}) + (\bar{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) = q(\boldsymbol{\mu}) + (\bar{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}), \end{aligned}$$

which implies that $\boldsymbol{g}(\boldsymbol{x}) \in \partial q(\boldsymbol{\mu})$.

(d) The inclusion $\partial q(\boldsymbol{\mu}) \subseteq \operatorname{conv} \{ \boldsymbol{g}(\boldsymbol{x}) \mid \boldsymbol{x} \in X(\boldsymbol{\mu}) \}$ follows from (c) and the convexity of $\partial q(\boldsymbol{\mu})$. The opposite inclusion follows by applying the Separation Theorem 3.24.[7]

(e) See Danskin's Theorem in Proposition 6.16(d). ■

The result in (c) is an independent proof of the concavity of q on $\mathbb{R}^m$.

The result (d) is particularly interesting, because by Carathéodory's Theorem 3.8 it says that every subgradient of q at any point $\boldsymbol{\mu}$ is the convex combination of a finite number (in fact, at most $m+1$) of vectors of the form $\boldsymbol{g}(\boldsymbol{x}^s)$ with $\boldsymbol{x}^s \in X(\boldsymbol{\mu})$. Computationally, this has been utilized to devise efficient (proximal) bundle methods for the Lagrangian dual problem as well as to devise methods to recover primal optimal solutions.

Next, we establish the differentiability of the dual function under additional assumptions.

[6] See [Roc70, Theorem 10.1 and its Corollary 10.1.1].
[7] See [BSS93, Theorem 6.3.7] for a detailed proof.

Proposition 6.18 (differentiability of the dual function) *Suppose that, in the problem* (6.4), (6.38) *holds.*

(a) *Let $\boldsymbol{\mu} \in \mathbb{R}^m$. The dual function q is differentiable at $\boldsymbol{\mu}$ if and only if $\{\, \boldsymbol{g}(\boldsymbol{x}) \mid \boldsymbol{x} \in X(\boldsymbol{\mu}) \,\}$ is a singleton set, that is, if the value of the vector of constraint functions is invariant over the set of solutions $X(\boldsymbol{\mu})$ to the Lagrangian subproblem. Then, we have that*

$$\nabla q(\boldsymbol{\mu}) = \boldsymbol{g}(\boldsymbol{x}),$$

for every $\boldsymbol{x} \in X(\boldsymbol{\mu})$.

(b) *The result in* (a) *holds in particular if the Lagrangian subproblem has a unique solution, that is, $X(\boldsymbol{\mu})$ is a singleton set. In particular, this property is satisfied if further X is a convex set, f is strictly convex on X, and g_i $(i = 1, \dots, m)$ are convex, in which case q is even in C^1.*

Proof. (a) The concave function q is differentiable at the point $\boldsymbol{\mu}$ (where it is finite) if and only if its subdifferential $\partial q(\boldsymbol{\mu})$ there is a singleton, cf. Proposition 6.16(c).

(b) Under either one of the assumptions stated, $X(\boldsymbol{\mu})$ is a singleton, whence the result follows from (a). Uniqueness follows from the convexity of the feasible set and strict convexity of the objective function, according to Proposition 4.10. ■

Proposition 6.19 (twice differentiability of the dual function) *Suppose that, in the problem* (6.4), *$X = \mathbb{R}^n$, and f and g_i $(i = 1, \dots, m)$ are convex functions in C^2. Suppose that, at $\boldsymbol{\mu} \in \mathbb{R}^m$, the solution $\boldsymbol{x}$ to the Lagrangian subproblem not only is unique, but also that the partial Hessian of the Lagrangian is positive definite at the pair $(\boldsymbol{x}, \boldsymbol{\mu})$, that is,*

$$\nabla^2_{\boldsymbol{xx}} L(\boldsymbol{x}, \boldsymbol{\mu}) \text{ is positive definite.}$$

Then, the dual function q is twice differentiable at $\boldsymbol{\mu}$, with

$$\nabla^2 q(\boldsymbol{\mu}) = -\boldsymbol{g}(\boldsymbol{x})^{\mathrm{T}} [\nabla^2_{\boldsymbol{xx}} L(\boldsymbol{x}, \boldsymbol{\mu})]^{-1} \boldsymbol{g}(\boldsymbol{x}).$$

Proof. The result follows from the Implicit Function Theorem, which is stated in Chapter 2, applied to the Lagrangian subproblem.[8] ■

6.5 Subgradient optimization methods

We begin by establishing the convergence of classic subgradient optimization methods as applied to a general convex optimization problem.

[8] See [Ber99, Pages 596–598] for a detailed analysis.

6.5.1 Convex problems

Consider the convex optimization problem to

$$\underset{\boldsymbol{x}}{\text{minimize}} \ f(\boldsymbol{x}), \tag{6.40a}$$

$$\text{subject to } \boldsymbol{x} \in X, \tag{6.40b}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and the set $X \subseteq \mathbb{R}^n$ is nonempty, closed and convex.

The subgradient projection algorithm is as follows: select $\boldsymbol{x}_0 \in X$, and for $k = 0, 1, \ldots$ generate

$$\boldsymbol{g}_k \in \partial f(\boldsymbol{x}_k), \tag{6.41}$$

$$\boldsymbol{x}_{k+1} = \operatorname{Proj}_X (\boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k), \tag{6.42}$$

where the sequence $\{\alpha_k\}$ is generated from one of the following three rules.

The first rule is termed the *divergent series* step length rule, and requires that

$$\alpha_k > 0, \quad k = 0, 1, \ldots; \qquad \lim_{k \to \infty} \alpha_k = 0; \qquad \sum_{k=0}^{\infty} \alpha_k = +\infty. \tag{6.43}$$

The second rule adds to the requirements in (6.43) the square-summable restriction

$$\sum_{k=0}^{\infty} \alpha_k^2 < +\infty. \tag{6.44}$$

The conditions in (6.43) allow for convergence to *any* point from *any* starting point, since the total step is infinite, but convergence is therefore also quite slow; the additional condition in (6.44) means that the fastest among these sequences are selected. An instance of the step length formulas which satisfies both (6.43) and (6.44) is the following:

$$\alpha_k = \beta/(k+1), \qquad k = 0, 1, \ldots,$$

where $\beta > 0$.

The third step length rule is

$$\sigma \le \alpha_k \le 2[f(\boldsymbol{x}_k) - f^*]/\|\boldsymbol{g}_k\|^2 - \sigma, \tag{6.45}$$

where f^* is the optimal value of (6.40). We refer to this step length formula as the *Polyak step*, after the Russian mathematician Boris Polyak who invented the subgradient method in the 1960s together with Ermol'ev and Shor.

How is convergence established for subgradient optimization methods? As shall be demonstrated in Chapters 11 and 12 convergence of algorithms for problems with a *differentiable* objective function is typically based on generating descent directions, and step length rules that result in the sequence $\{\boldsymbol{x}_k\}$ of iterates being strictly descending in the value of f. For the non-differentiable problem at hand, generating descent directions is a difficult task, since it is not true that the negative of an arbitrarily chosen subgradient of f at a non-optimal vector $\boldsymbol{x}$ defines a descent direction.

In *bundle methods* one gathers information from more than one subgradient (hence the term *bundle*) around a current iteration point so that a descent direction can be generated, followed by an inexact line search. We concentrate here on the simpler methodology of subgradient optimization methods, in which we apply the formula (6.52) where the step length α_k is chosen based on very simple rules.

We establish below that if the step length is small enough, an iteration of the subgradient projection method leads to a vector that is closer to the set of optimal solutions. This technical result also motivates the construction of the Polyak step length rule, and hence shows that convergence of subgradient methods are based on the reduction of the Euclidean distance to the optimal solutions rather than on the objective function f.

Proposition 6.20 (decreasing distance to the optimal set) *Suppose that* $\boldsymbol{x}_k \in X$ *is not optimal in* (6.40), *and that* $\boldsymbol{x}_{k+1}$ *is given by* (6.42) *for some step length* $\alpha_k > 0$.

Then, for every optimal solution $\boldsymbol{x}^*$ *in* (6.40),

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| < \|\boldsymbol{x}_k - \boldsymbol{x}^*\|$$

holds for every step length α_k *in the interval*

$$\alpha_k \in (0, 2[f(\boldsymbol{x}_k) - f^*]/\|\boldsymbol{g}_k\|^2). \tag{6.46}$$

Proof. We have that

$$\begin{aligned}
\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 &= \|\mathrm{Proj}_X(\boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k) - \boldsymbol{x}^*\|^2 \\
&= \|\mathrm{Proj}_X(\boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k) - \mathrm{Proj}_X(\boldsymbol{x}^*)\|^2 \\
&\le \|\boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k - \boldsymbol{x}^*\|^2 \\
&= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - 2\alpha_k(\boldsymbol{x}_k - \boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{g}_k + \alpha_k^2\|\boldsymbol{g}_k\|^2 \\
&\le \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - 2\alpha_k[f(\boldsymbol{x}_k) - f^*] + \alpha_k^2\|\boldsymbol{g}_k\|^2 \\
&< \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2,
\end{aligned}$$

where we have utilized the property that the Euclidean projection is non-expansive (Theorem 4.31), the subgradient inequality (6.34) for convex functions, and the bounds on α_k given by (6.46). ∎

Our first convergence result is based on the divergent series step length formula (6.43), and establishes convergence to the optimal solution set X^* under an assumption on its boundedness. With the other two step length formulas, this condition will be possible to remove.

Recall the definition (3.10) of the minimum distance from a vector to a closed and convex set; our interest is in the distance from an arbitrary vector $\boldsymbol{x} \in \mathbb{R}^n$ to the solution set X^*:

$$\operatorname{dist}_{X^*}(\boldsymbol{x}) := \underset{\boldsymbol{y} \in X^*}{\operatorname{minimum}} \|\boldsymbol{y} - \boldsymbol{x}\|.$$

Theorem 6.21 (convergence of subgradient optimization methods, I) *Let $\{\boldsymbol{x}_k\}$ be generated by the method (6.42), (6.43). If X^* is bounded and the sequence $\{\boldsymbol{g}_k\}$ is bounded, then $\{f(\boldsymbol{x}_k)\} \to f^*$ and $\{\operatorname{dist}_{X^*}(\boldsymbol{x}_k)\} \to 0$ holds.*

Proof. We show that the iterates will eventually belong to an arbitrarily small neighbourhood of the set of optimal solutions to (6.40).

Let $\delta > 0$ and $B^\delta = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| \leq \delta \,\}$. Since f is convex, X is nonempty, closed and convex, and X^* is bounded, it follows from [Roc70, Theorem 27.2], applied to the lower semi-continuous, proper[9] and convex function $f + \chi_X$)[10] that there exists an $\varepsilon = \varepsilon(\delta) > 0$ such that the level set $\{\, \boldsymbol{x} \in X \mid f(\boldsymbol{x}) \leq f^* + \varepsilon \,\} \subseteq X^* + B^{\delta/2}$; this level set is denoted by X^ε. Moreover, since for all k, $\|\boldsymbol{g}_k\| \leq \sup_s\{\|\boldsymbol{g}_s\|\} < \infty$, and $\{\alpha_k\} \to 0$, there exists an $N(\delta)$ such that $\alpha_k\|\boldsymbol{g}_k\|^2 \leq \varepsilon$ and $\alpha_k\|\boldsymbol{g}_k\| \leq \delta/2$ for all $k \geq N(\delta)$.

The sequel of the proof is based on induction and is organized as follows. In the first part, we show that there exists a finite $k(\delta) \geq N(\delta)$ such that $\boldsymbol{x}_{k(\delta)} \in X^* + B^\delta$. In the second part, we establish that if $\boldsymbol{x}_k$ belongs to $X^* + B^\delta$ for some $k \geq N(\delta)$ then so does $\boldsymbol{x}_{k+1}$, by showing that either $\operatorname{dist}_{X^*}(\boldsymbol{x}_{k+1}) < \operatorname{dist}_{X^*}(\boldsymbol{x}_k)$ holds, or $\boldsymbol{x}_k \in X^\varepsilon$ so that $\boldsymbol{x}_{k+1} \in X^* + B^\delta$ since the step taken is not longer than $\delta/2$.

[9] A *proper* function is a function which is finite at least at some vector and nowhere attains the value $-\infty$. See also Section 1.4.

[10] For any set $S \subset \mathbb{R}^n$ the function χ_S is the *indicator function* of the set S, that is, $\chi_S(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in S$; and $\chi_S(\boldsymbol{x}) = +\infty$ if $\boldsymbol{x} \notin S$.

Let $\boldsymbol{x}^* \in X^*$ be arbitrary. In every iteration k we then have

$$
\begin{aligned}
\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|^2 &= \|\boldsymbol{x}^* - \operatorname{Proj}_X(\boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k)\|^2 && (6.47\text{a})\\
&\le \|\boldsymbol{x}^* - \boldsymbol{x}_k + \alpha_k \boldsymbol{g}_k\|^2 && (6.47\text{b})\\
&= \|\boldsymbol{x}^* - \boldsymbol{x}_k\|^2 + \alpha_k \left(2\boldsymbol{g}_k^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_k) + \alpha_k \|\boldsymbol{g}_k\|^2\right), && (6.47\text{c})
\end{aligned}
$$

where the inequality follows from the projection property. Now, suppose that

$$2\,\boldsymbol{g}_s^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_s) + \alpha_s \|\boldsymbol{g}_s\|^2 < -\varepsilon \tag{6.48}$$

for all $s \ge N(\delta)$. Then, using (6.47) repeatedly, we obtain that for any $k \ge N(\delta)$,

$$\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|^2 < \left\|\boldsymbol{x}^* - \boldsymbol{x}_{N(\delta)}\right\|^2 - \varepsilon \sum_{s=N(\delta)}^{k} \alpha_s,$$

and from (6.52) it follows that the right-hand side of this inequality tends to minus infinity as $k \to \infty$, which clearly is impossible. Therefore,

$$2\,\boldsymbol{g}_k^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_k) + \alpha_k \|\boldsymbol{g}_k\|^2 \ge -\varepsilon \tag{6.49}$$

for at least one $k \ge N(\delta)$, say $k = k(\delta)$. From the definition of $N(\delta)$, it follows that $\boldsymbol{g}_{k(\delta)}^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_{k(\delta)}) \ge -\varepsilon$. From the definition of a subgradient (cf. Definition 6.15) we have that $f(\boldsymbol{x}^*) - f(\boldsymbol{x}_{k(\delta)}) \ge \boldsymbol{g}_{t(\delta)}^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_{k(\delta)})$, since $\boldsymbol{x}^*, \boldsymbol{x}_{k(\delta)} \in X$. Hence, $f(\boldsymbol{x}_{k(\delta)}) \le f^* + \varepsilon$, that is, $\boldsymbol{x}_{k(\delta)} \in X^\varepsilon \subseteq X^* + B^{\delta/2} \subset X^* + B^\delta$.

Now, suppose that $\boldsymbol{x}_k \in X^* + B^\delta$ for some $k \ge N(\delta)$. If (6.48) holds, then, using (6.47), we have that $\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\| < \|\boldsymbol{x}^* - \boldsymbol{x}_k\|$ for any $\boldsymbol{x}^* \in \boldsymbol{x}^*$. Hence,

$$
\begin{aligned}
\operatorname{dist}_{X^*}(\boldsymbol{x}_{k+1}) \le \|\operatorname{Proj}_{X^*}(\boldsymbol{x}_k) - \boldsymbol{x}_{k+1}\| &< \|\operatorname{Proj}_{X^*}(\boldsymbol{x}_k) - \boldsymbol{x}_k\|\\
&= \operatorname{dist}_{X^*}(\boldsymbol{x}_k) \le \delta.
\end{aligned}
$$

Thus, $\boldsymbol{x}_{k+1} \in X^* + B^\delta$. Otherwise, (6.49) must hold and, using the same arguments as above, we obtain that $f(\boldsymbol{x}_k) \le f^* + \varepsilon$, i.e., $\boldsymbol{x}_k \in X^\varepsilon \subseteq \boldsymbol{x}^* + B^{\delta/2}$. As

$$
\begin{aligned}
\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| = \|\operatorname{Proj}_X(\boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k) - \boldsymbol{x}_k\| &\le \|\boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k - \boldsymbol{x}_k\|\\
&= \alpha_k \|\boldsymbol{g}_k\| \le \frac{\delta}{2}
\end{aligned}
$$

whenever $k \ge N(\delta)$, it follows that $\boldsymbol{x}_{k+1} \in X^* + B^{\delta/2} + B^{\delta/2} = X^* + B^\delta$.

By induction with respect to $k \geq k(\delta)$, it follows that $\boldsymbol{x}_k \in X^* + B^\delta$ for all $k \geq k(\delta)$. Since this holds for arbitrarily small values of $\delta > 0$ and f is continuous, the theorem follows. ∎

We next introduce the additional requirement (6.44); the resulting algorithm's convergence behaviour is now much more favourable, and the proof is at the same time less technical.

Theorem 6.22 (convergence of subgradient optimization methods, II) *Let $\{\boldsymbol{x}_k\}$ be generated by the method* (6.42), (6.43), (6.44). *If X^* is nonempty and the sequence $\{\boldsymbol{g}_k\}$ is bounded, then $\{f(\boldsymbol{x}_k)\} \to f^*$ and $\{\boldsymbol{x}_k\} \to \boldsymbol{x}^* \in X^*$ holds.*

Proof. Let $\boldsymbol{x}^* \in X^*$ and $k \geq 1$. Repeated application of (6.47) yields that

$$\|\boldsymbol{x}^* - \boldsymbol{x}_k\|^2 \leq \|\boldsymbol{x}^* - \boldsymbol{x}_0\|^2 + 2\sum_{s=0}^{k-1} \alpha_s \boldsymbol{g}_s^{\mathrm{T}} (\boldsymbol{x}^* - \boldsymbol{x}_s) + \sum_{s=0}^{k-1} \alpha_s^2 \|\boldsymbol{g}_s\|^2 . \tag{6.50}$$

Since $\boldsymbol{x}^* \in X^*$ and $\boldsymbol{g}_s \in \partial f(\boldsymbol{x}_s)$ for all $s \geq 0$ we obtain that

$$f(\boldsymbol{x}_s) \geq f^* \geq f(\boldsymbol{x}_s) + \boldsymbol{g}_s^{\mathrm{T}} (\boldsymbol{x}^* - \boldsymbol{x}_s), \qquad s \geq 0, \tag{6.51}$$

and hence that $\boldsymbol{g}_s^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_s) \leq 0$ for all $s \geq 0$. Define $c := \sup_k\{\|\boldsymbol{g}_k\|\}$ and $p = \sum_{k=0}^{\infty} \alpha_k^2$, so that $\|\boldsymbol{g}_s\| \leq c$ for any $s \geq 0$ and $\sum_{s=0}^{k-1} \alpha_s^2 < p$. From (6.50) we then conclude that $\|\boldsymbol{x}^* - \boldsymbol{x}_k\|^2 < \|\boldsymbol{x}^* - \boldsymbol{x}_0\|^2 + pc^2$ for any $k \geq 1$, and thus that the sequence $\{\boldsymbol{x}_k\}$ is bounded.

Assume now that there is no subsequence $\{\boldsymbol{x}^{k_i}\}$ of $\{\boldsymbol{x}_k\}$ with $\{\boldsymbol{g}_{k_i}^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_{k_i})\} \to 0$. Then there must exist an $\varepsilon > 0$ with $\boldsymbol{g}_s^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_s) \leq -\varepsilon$ for all sufficiently large values of s. From (6.50) and the conditions on the step lengths it follows that $\{\|\boldsymbol{x}^* - \boldsymbol{x}_s\|\} \to -\infty$, which clearly is impossible. The sequence $\{\boldsymbol{x}_k\}$ must therefore contain a subsequence $\{\boldsymbol{x}_{k_i}\}$ such that $\{\boldsymbol{g}_{k_i}^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_{k_i})\} \to 0$. From (6.51) it follows that $\{f(\boldsymbol{x}_{k_i})\} \to f^*$. The boundedness of $\{\boldsymbol{x}_k\}$ implies the existence of an accumulation point of the subsequence $\{\boldsymbol{x}_{k_i}\}$, say $\boldsymbol{x}^\infty$. From the continuity of f it follows that $\boldsymbol{x}^\infty \in X^*$.

To show that $\boldsymbol{x}^\infty$ is the only accumulation point of $\{\boldsymbol{x}_k\}$, let $\delta > 0$ and find an $M(\delta)$ such that $\|\boldsymbol{x}^\infty - \boldsymbol{x}_{M(\delta)}\|^2 \leq \delta/2$ and $\sum_{s=M(\delta)}^{\infty} \alpha_s^2 \leq \delta/(2c^2)$. Consider any $k > M(\delta)$. Analogously to the derivation of (6.50), and using (6.51), we then obtain that

$$\|\boldsymbol{x}^\infty - \boldsymbol{x}_k\|^2 \leq \left\|\boldsymbol{x}^\infty - \boldsymbol{x}_{M(\delta)}\right\|^2 + \sum_{s=M(\delta)}^{k-1} \alpha_s^2 \|\boldsymbol{g}_s\|^2 < \frac{\delta}{2} + \frac{\delta}{2c^2} c^2 = \delta.$$

Since this holds for arbitrarily small values of $\delta > 0$, we are done. ∎

We note that it is possible to remove the boundedness condition on $\{\boldsymbol{g}_k\}$ at the simple price of scaling the length of the search direction; we preferred to keep the simplicity of the algorithms, however. Note further that this condition is immediately ensured to be fulfilled whenever we know before-hand that the sequence $\{\boldsymbol{x}_k\}$ is bounded, such as is the case when X itself is bounded.

Convergence of this type of method is quite slow. In fact, the step length rule (6.43) *prevents* fast convergence to be possible. We say that the sequence $\{\boldsymbol{x}_k\}$ converges to $\boldsymbol{x}^*$ with a *geometric* rate if there exists $M > 0$ and $q \in (0, 1)$ with

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq Mq^k, \qquad k = 0, 1, \dots .$$

Suppose that this speed is possible together with the step length rule (6.43). For all k, the above yields

$$\alpha_k = \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| \leq \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| + \|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq M(q+1)q^k;$$

summing these inequalities up implies

$$\sum_{k=0}^{\infty} \alpha_k \leq M(q+1)/(1-q),$$

which contradicts the divergent series condition (6.43).

We finally present the convergence properties of the subgradient projection method using the Polyak step; it is even stronger than the result of Theorem 6.22.

Theorem 6.23 (convergence of subgradient optimization methods, III) *Let $\{\boldsymbol{x}_k\}$ be generated by the method* (6.42), (6.45). *If X^* is nonempty then $\{f(\boldsymbol{x}_k)\} \to f^*$ and $\{\boldsymbol{x}_k\} \to \boldsymbol{x}^* \in X^*$ holds.*

Proof. From Proposition 6.20 follows that the sequence $\{\{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|\}$ is strictly decreasing for every $\boldsymbol{x}^* \in X^*$, and therefore has a limit. By construction of the step length, in which the step lengths are bounded away from zero and $2[f(\boldsymbol{x}_k) - f^*]/\|\boldsymbol{g}_k\|^2$, it follows from the proof of Proposition 6.20 that $\{[f(\boldsymbol{x}_k) - f^*]/\|\boldsymbol{g}_k\|^2\} \to 0$ must hold. Since $\{\boldsymbol{g}_k\}$ must be bounded due to the boundedness of $\{\boldsymbol{x}_k\}$, we have that $\{f(\boldsymbol{x}_k)\} \to f^*$. Further, $\{\boldsymbol{x}_k\}$ is bounded, and due to the continuity properties of f every limit point must then belong to X^*.

It remains to show that there can be only one limit point. This property follows from the monotone decrease of the distance $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|$.

In detail, the proof is as follows. Suppose two subsequences of $\{\boldsymbol{x}_k\}$ exist, such that they converge to two different vectors in X^*:

$$\{\boldsymbol{x}_{m_i}\} \to \boldsymbol{x}_1^*; \qquad \{\boldsymbol{x}_{l_i}\} \to \boldsymbol{x}_2^*; \qquad \boldsymbol{x}_1^* \neq \boldsymbol{x}_2^*.$$

We must then have $\{\|\boldsymbol{x}_{l_i} - \boldsymbol{x}_1^*\|\} \to \rho_1 > 0$, while $\{\|\boldsymbol{x}_{m_i} - \boldsymbol{x}_2^*\|\} \to \rho_2 > 0$. Since the two vectors are optimal and the distance to X^* is descending, $\{\|\boldsymbol{x}_k - \boldsymbol{x}_1^*\|\} \to \rho_1$, while $\{\|\boldsymbol{x}_k - \boldsymbol{x}_2^*\|\} \to \rho_2$ holds; in particular, $\{\|\boldsymbol{x}_{m_i} - \boldsymbol{x}_1^*\|\} \to \rho_1$, while $\{\|\boldsymbol{x}_{l_i} - \boldsymbol{x}_2^*\|\} \to \rho_2$. For clarity, assume that $\rho_1 \leq \rho_2$ holds. Then,

$$\|\boldsymbol{x}_{m_i} - \boldsymbol{x}_1^*\|^2 = \|\boldsymbol{x}_{m_i} - \boldsymbol{x}_2^* + \boldsymbol{x}_2^* - \boldsymbol{x}_1^*\|^2 \to \rho_2^* + \|\boldsymbol{x}_2^* - \boldsymbol{x}_1^*\| > \rho_1^2,$$

which is impossible since $\|\boldsymbol{x}_{m_i} - \boldsymbol{x}_1^*\| \to \rho_1$. ∎

Contrary to the slow convergence of the subgradient projection algorithms that rely on the divergent series step length rule, those based on the Polyak step length (6.45) are geometrically convergent under additional assumptions on the function f. For example, geometric convergence follows rather easily from the condition that f has a set of *weak sharp minima*: there exists $m \geq 0$ such that

$$f(\boldsymbol{x}) - f^* \geq m \operatorname{dist}_{X^*}(\boldsymbol{x}), \qquad \boldsymbol{x} \in X.$$

(It can be shown that this condition holds, for example, for every LP problem which has a bounded optimal solution.) The argument is that there exists a large enough $L > 0$ (due to the boundedness of $\{\|\boldsymbol{g}_k\|\}$) such that

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 \leq \underbrace{\left(1 - \frac{\sigma(2-\sigma)m}{L^2}\right)}_{=q^2<1} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2.$$

6.5.2 Application to the Lagrangian dual problem

We remind ourselves that the Lagrangian dual problem is a concave maximization problem, and that the appearance of the dual function is similar to that of the following example:

Let $h(x) := \text{minimum}\,\{h_1(x), h_2(x)\}$, where $h_1(x) := 4 - |x|$ and $h_2(x) := 4 - (x-2)^2$. Then,

$$h(x) = \begin{cases} 4 - x, & 1 \leq x \leq 4, \\ 4 - (x-2)^2, & x \leq 1,\ x \geq 4; \end{cases}$$

cf. Figure 6.3.

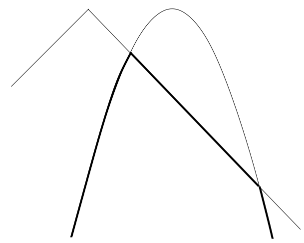


Figure 6.3: A min-function with two pieces.

The function h is non-differentiable at $x = 1$ and $x = 4$, since its graph has non-unique supporting hyperplanes there:

$$\partial h(x) = \begin{cases} \{-1\}, & 1 < x < 4 \\ \{4 - 2x\}, & x < 1, \ x > 4 \\ [-1, 2]\,, & x = 1 \\ [-4, -1]\,, & x = 4; \end{cases}$$

the subdifferential is here either a singleton (at differentiable points) or an interval (at non-differentiable points).

Note that the subdifferential includes zero at $x^* = 1$, whence it defines the (unique) maximum.

Now, let $\boldsymbol{g} \in \partial q(\bar{\boldsymbol{\mu}})$, and let U^* be the set of optimal solutions to (6.9). Then,

$$U^* \subseteq \{\, \boldsymbol{\mu} \in \mathbb{R}^m \mid \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) \geq 0 \,\}.$$

In other words, $\boldsymbol{g}$ defines a half-space that contains the set of optimal solutions. We therefore know that if the step length is small enough we get closer to the set of optimal solutions. Consider Figure 6.4 however: the subgradient depicted is not an ascent direction! As we saw in the previous section, convergence must be based on other arguments, like the decreasing distance to U^* alluded to above, and in the previous section.

We consider the Lagrangian dual problem (6.9). We suppose, as in the previous section, that X is compact so that the infimum in (6.8) is attained for every $\boldsymbol{\mu} \geq \mathbf{0}^m$ (which is the set over which we wish to maximize q) and q is real-valued over $\mathbb{R}^m_+$.

In the case of our special concave maximization problem, the iteration has the form

$$\begin{aligned} \boldsymbol{\mu}_{k+1} &= \mathrm{Proj}_{\mathbb{R}^m_+}[\boldsymbol{\mu}_k + \alpha_k \boldsymbol{g}_k] && \text{(6.52a)} \\ &= [\boldsymbol{\mu}_k + \alpha_k \boldsymbol{g}_k]_+ && \text{(6.52b)} \\ &= (\text{maximum}\,\{0, (\boldsymbol{\mu}_k)_i + \alpha_k (\boldsymbol{g}_k)_i\})_{i=1}^m, && \text{(6.52c)} \end{aligned}$$

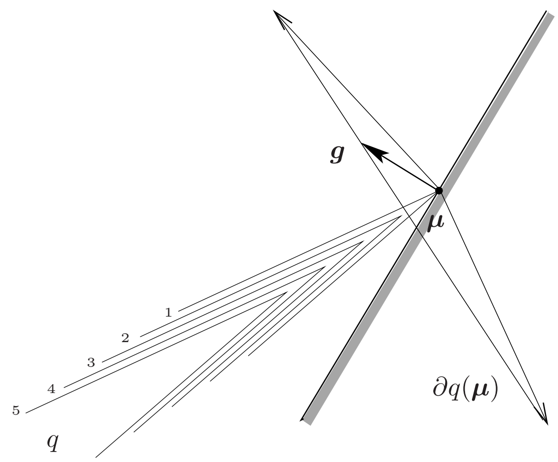


Figure 6.4: The half-space defined by the subgradient $\boldsymbol{g}$ of q at $\boldsymbol{\mu}$. Note that the subgradient is not an ascent direction.

where $\boldsymbol{g}_k \in \partial q(\boldsymbol{\mu}_k)$ is arbitrarily chosen; we would typically use $\boldsymbol{g}_k = \boldsymbol{g}(\boldsymbol{x}_k)$, where $\boldsymbol{x}_k \in \arg\operatorname{minimum}_{\boldsymbol{x}\in X} L(\boldsymbol{x}, \boldsymbol{\mu}_k)$. The projection operation onto the first orthant is, as we can see, very simple.

Replacing the Polyak step (6.45) with the corresponding dual form

$$\sigma \leq \alpha_k \leq 2[q^* - q(\boldsymbol{\mu}_k)]/\|\boldsymbol{g}_k\|^2 - \sigma, \qquad k = 1, 2, \dots, \tag{6.53}$$

convergence will now be a simple consequence of the above theorems.

The conditions (6.38) and that the feasible set of (6.4) is nonempty ensure that the problem (6.4) has an optimal solution; in particular, (6.5) then holds. Further, if we introduce the Slater condition (6.15), we are ensured that there is no duality gap, and that the dual problem (6.9) has a compact set U^* of optimal solutions. Under these assumptions, we have the following results for subgradient optimization methods.

Theorem 6.24 (convergence of subgradient optimization methods) *Suppose that the problem* (6.4) *is feasible, and that* (6.38) *and* (6.15) *hold.*

(a) *Let* $\{\boldsymbol{\mu}_k\}$ *be generated by the method* (6.52)*, under the divergent step length rule* (6.43)*. Then,* $\{q(\boldsymbol{\mu}_k)\} \to q^*$*, and* $\{\operatorname{dist}_{U^*}(\boldsymbol{\mu}_k)\} \to 0$.

(b) *Let* $\{\boldsymbol{\mu}_k\}$ *be generated by the method* (6.52)*, under the divergent step length rule* (6.43), (6.44)*. Then,* $\{\boldsymbol{\mu}_k\}$ *converges to an optimal solution to* (6.9).

(c) *Let* $\{\boldsymbol{\mu}_k\}$ *be generated by the method* (6.52)*, under the Polyak step length rule* (6.53)*, where* σ *is a small positive number. Then,* $\{\boldsymbol{\mu}_k\}$

converges to an optimal solution to (6.9)*.*

Proof. The results follow from Theorems 6.21, 6.22, and 6.23. Note that in the first two cases, boundedness conditions were assumed for X^* and the sequence of subgradients. The corresponding conditions for the Lagrangian dual problem are fulfilled under the CQs imposed, since they imply that the search for an optimal solution is done over a compact set; cf. Theorem 6.8(a) and its proof. ∎

6.6 *Obtaining a primal solution

It remains for us to show how an optimal dual solution $\boldsymbol{\mu}^*$ can be *translated* into an optimal primal solution $\boldsymbol{x}^*$. Obviously, convexity and strong duality will be needed in general, if we are to be able to utilize the primal–dual optimality characterization in Theorem 6.6. It turns out that the generation of a primal optimum is automatic if q is differentiable at $\boldsymbol{\mu}^*$, something which we can refer to as the *Lagrange multiplier method*. Unfortunately, in many cases, such as for most non-strictly convex optimization problems (like linear programming), this will not be the case, and then the translation work then becomes rather more complex.

We start with the ideal case.

6.6.1 Differentiability at the optimal solution

The following results summarize the optimality conditions for the Lagrangian dual problem (6.9), and their consequences for the availability of a primal optimal solution in the absence of a duality gap.

Proposition 6.25 (optimality conditions for the dual problem) *Suppose that, in the problem* (6.4)*, the condition* (6.38) *holds. Suppose further that the Lagrangian dual problem has an optimal solution,* $\boldsymbol{\mu}^*$*.*

(a) *The dual optimal solution is characterized by the inclusion*

$$\mathbf{0}^m \in -\partial q(\boldsymbol{\mu}^*) + N_{\mathbb{R}^m_+}(\boldsymbol{\mu}^*). \tag{6.54}$$

In other words, there then exists $\boldsymbol{\gamma}^* \in \partial q(\boldsymbol{\mu}^*)$*—an optimality-characterizing subgradient of* q *at* $\boldsymbol{\mu}^*$*—such that*

$$\mathbf{0}^m \leq \boldsymbol{\mu}^* \perp \boldsymbol{\gamma}^* \leq \mathbf{0}^m. \tag{6.55}$$

There exist a finite set of solutions $\boldsymbol{x}^i \in X(\boldsymbol{\mu}^*)$ $(i = 1, \dots, k)$ *where* $k \leq m + 1$ *such that*

$$\boldsymbol{\gamma}^* = \sum_{i=1}^{k} \alpha_i \boldsymbol{g}(\boldsymbol{x}^i); \qquad \sum_{i=1}^{k} \alpha_i = 1; \qquad \alpha_i \geq 0, \; i = 1, \dots, k. \tag{6.56}$$

Hence, we have that

$$\sum_{i=1}^{k} \alpha_i \mu_i^* g_i(\boldsymbol{x}^i) = 0, \qquad j = 1, \dots, m. \tag{6.57}$$

(b) *If there is a duality gap, then* q *is non-differentiable at* $\boldsymbol{\mu}^*$*.*

(c) *If* q *is differentiable at* $\boldsymbol{\mu}^*$*, then there is no duality gap. Further, any vector in* $X(\boldsymbol{\mu}^*)$ *then solves the primal problem* (6.4).

Proof. (a) The first result is a direct statement of the optimality conditions of the convex and sub-differentiable program (6.9); the complementarity conditions in (6.55) are an equivalent statement of the inclusion in (6.54).

The second result is an application of Carathéodory's Theorem 3.8 to the compact and convex set $\partial q(\boldsymbol{\mu}^*)$.

(b) The result is immediately established, once (c) is, since they are equivalent.

(c) Let $\bar{\boldsymbol{x}}$ be any vector in $X(\boldsymbol{\mu}^*)$ for which $\nabla q(\boldsymbol{\mu}^*) = \boldsymbol{g}(\bar{\boldsymbol{x}})$ holds, cf. Proposition 6.18(a). We obtain from (6.55) that

$$\mathbf{0}^m \leq \boldsymbol{\mu}^* \perp \boldsymbol{g}(\bar{\boldsymbol{x}}) \leq \mathbf{0}^m.$$

Hence, the pair $(\boldsymbol{\mu}, \bar{\boldsymbol{x}})$ fulfills all the conditions stated in (6.11), so that, by Theorem 6.6, $\bar{\boldsymbol{x}}$ is an optimal solution to (6.4). ∎

Many interesting problems do not comply with the conditions in (c); for example, linear programming is one where the Lagrangian dual problem often is non-differentiable at every dual optimal solution.[11] This is sometimes called the *non-coordinability phenomenon* (cf. [Las70, DiJ79]). It was in order to cope with this phenomenon that Dantzig–Wolfe decomposition ([DaW60, Las70]) and other column generation algorithms, Benders decomposition ([Ben62, Las70]) and generalized linear programming were developed; noticing that the convex combination of a finite number of candidate primal solutions are sufficient to verify an optimal primal–dual solution [cf. (6.57)], methodologies were developed

[11] In other words, even if a Lagrange multiplier vector is known, the Lagrangian subproblem may not identify a primal optimal solution.

to generate those vectors algorithmically. See also [LPS99] for overviews on the subject of generating primal optimal solutions from dual optimal ones, and [BSS93, Theorem 6.5.2] for an LP procedure that provides primal feasible solutions for convex programs.

Note that the equation (6.57) in (a) reduces to the complementarity condition that $\mu_i^* g_i(\bar{\boldsymbol{x}}) = 0$ holds, for the *averaged* solution, $\bar{\boldsymbol{x}} := \sum_{i=1}^k \alpha_i \boldsymbol{x}^i$, whenever all the functions g_i are affine.

6.6.2 Everett's Theorem

The next result shows that the solution to the Lagrangian subproblem solves a perturbed version of the original problem. We state the result for the general problem to find

$$f^* := \operatorname*{infimum}_{\boldsymbol{x}} f(\boldsymbol{x}), \tag{6.58a}$$
$$\text{subject to } \boldsymbol{x} \in X, \tag{6.58b}$$
$$g_i(\boldsymbol{x}) \leq 0, \qquad i = 1, \dots, m, \tag{6.58c}$$
$$h_j(\boldsymbol{x}) = 0, \qquad j = 1, \dots, \ell, \tag{6.58d}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$ ($i = 1, 2, \dots, m$), and $h_j : \mathbb{R}^n \to \mathbb{R}$ ($j = 1, 2, \dots, \ell$) are given functions, and $X \subseteq \mathbb{R}^n$.

Theorem 6.26 (Everett's Theorem) *Let $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathbb{R}^m_+ \times \mathbb{R}^\ell$. Consider the Lagrangian subproblem to*

$$\operatorname*{minimize}_{\boldsymbol{x} \in X} \left\{ f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{h}(\boldsymbol{x}) \right\}. \tag{6.59}$$

Suppose that $\bar{\boldsymbol{x}}$ is an optimal solution to this problem, and let $\mathcal{I}(\boldsymbol{\mu}) \subseteq \{1, \dots, m\}$ denote the set of indices i for which $\mu_i > 0$.

(a) *$\bar{\boldsymbol{x}}$ is an optimal solution to the perturbed primal problem to*

$$\operatorname*{minimize}_{\boldsymbol{x}} f(\boldsymbol{x}), \tag{6.60a}$$
$$\text{subject to } \boldsymbol{x} \in X, \tag{6.60b}$$
$$g_i(\boldsymbol{x}) \leq g_i(\bar{\boldsymbol{x}}), \qquad i \in \mathcal{I}(\bar{\boldsymbol{x}}), \tag{6.60c}$$
$$h_j(\boldsymbol{x}) = h_j(\bar{\boldsymbol{x}}), \qquad j = 1, \dots, \ell. \tag{6.60d}$$

(b) *If $\bar{\boldsymbol{x}}$ is feasible in* (6.58) *and $\boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\bar{\boldsymbol{x}}) = 0$ holds, then $\bar{\boldsymbol{x}}$ solves* (6.58). *Moreover, the pair $(\boldsymbol{\mu}, \boldsymbol{\lambda})$ then solves the Lagrangian dual problem.*

Proof. (a) The proof proceeds by showing that the triple $(\bar{\boldsymbol{x}}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ is a saddle point of the function $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \mapsto f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}}[\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\bar{\boldsymbol{x}})] + \boldsymbol{\lambda}^{\mathrm{T}}[\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{h}(\bar{\boldsymbol{x}})]$ over $X \times \mathbb{R}^m_+ \times \mathbb{R}^\ell$.

Let $\boldsymbol{x}$ satisfy the constraints (6.60b)–(6.60d). Since we have that $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{h}(\bar{\boldsymbol{x}})$ and $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{g}(\bar{\boldsymbol{x}})$, the optimality of $\bar{\boldsymbol{x}}$ in (6.59) yields that

$$\begin{aligned} f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{g}(\bar{\boldsymbol{x}}) + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{h}(\bar{\boldsymbol{x}}) &\geq f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{h}(\boldsymbol{x}) \\ &\geq f(\bar{\boldsymbol{x}}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{g}(\bar{\boldsymbol{x}}) + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{h}(\bar{\boldsymbol{x}}), \end{aligned}$$

which shows that $f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}})$. We are done.

(b) $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{g}(\bar{\boldsymbol{x}}) = 0$ implies that $g_i(\bar{\boldsymbol{x}}) = 0$ for $i \in \mathcal{I}(\boldsymbol{\mu})$; from (a) $\bar{\boldsymbol{x}}$ solves the problem to

$$\underset{\boldsymbol{x}}{\operatorname{minimize}}\ f(\boldsymbol{x}), \tag{6.61a}$$

$$\text{subject to } \boldsymbol{x} \in X, \tag{6.61b}$$

$$g_i(\boldsymbol{x}) \leq 0, \qquad i \in \mathcal{I}(\bar{\boldsymbol{x}}), \tag{6.61c}$$

$$h_j(\boldsymbol{x}) = 0, \qquad j = 1, \dots, \ell. \tag{6.61d}$$

In particular, then, since the feasible set of (6.58) is contained in that of (6.61) and $\bar{\boldsymbol{x}}$ is feasible in the former, $\bar{\boldsymbol{x}}$ must also solve (6.58). That the pair $(\boldsymbol{\mu}, \boldsymbol{\lambda})$ solves the Lagrangian dual problem follows by the equality between the primal and dual objective functions at $(\bar{\boldsymbol{x}}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ and weak duality. ■

The result is taken from Everett [Eve63]. One important consequence of the result is that if the right-hand side perturbations $g_i(\bar{\boldsymbol{x}})$ and $h_i(\bar{\boldsymbol{x}})$ all are close to zero, the vector $\bar{\boldsymbol{x}}$ being near-feasible might mean that it is in fact acceptable as an approximate solution to the original problem. (This interpretation hinges on the dualized constraints being *soft* constraints, in the sense that a small violation is acceptable. See Section 1.7 for an introduction to the topic of soft constraints.)

6.7 *Sensitivity analysis

6.7.1 Analysis for convex problems

Consider the inequality constrained convex program (6.4), where $f : \mathbb{R}^n \to \mathbb{R}$ and g_i $(i = 1, \dots, m)$ are convex functions and $X \subseteq \mathbb{R}^n$ is a convex set. Suppose that the problem (6.4) is feasible, that the compactness condition (6.38) and Slater condition (6.15) hold. This is the classic case where there exist multipliers $\boldsymbol{\mu}^*$, according to Theorem 6.8, and strong duality holds.

For certain types of problems where there is no duality gap and where there exist primal–dual optimal solutions, we have access to a beautiful theory of *sensitivity analysis*. The classic meaning of the term is the answer to the following question: what is the rate of change in f^* when a constraint right-hand side changes? This question answers important practical questions, like the following in manufacturing:

- If we buy one unit of additional resource at a given price, or if the demand of a product that we sell increases by a certain amount, then how much additional profit do we make?

We will here provide a basic result which states when this sensitivity analysis of the optimal objective value can be performed for the problem (6.4), and establish that the answer is determined precisely by the value of the Lagrange multiplier vector $\boldsymbol{\mu}^*$, provided that it is unique.

Definition 6.27 (perturbation function) *Consider the function* $p : \mathbb{R}^m \to \mathbb{R} \cup \{\pm\infty\}$ *defined by*

$$p(\boldsymbol{u}) := \operatorname*{infimum}_{\boldsymbol{x}} \ f(\boldsymbol{x}), \tag{6.62a}$$

$$\text{subject to } \boldsymbol{x} \in X, \tag{6.62b}$$

$$g_i(\boldsymbol{x}) \leq u_i, \quad i = 1, \dots, m, \qquad \boldsymbol{u} \in \mathbb{R}^m; \tag{6.62c}$$

it is called the perturbation function, *or* primal function, *associated with the problem* (6.4). *Its effective domain is the set* $P := \{\, \boldsymbol{u} \in \mathbb{R}^m \mid p(\boldsymbol{u}) < +\infty \,\}$. ∎

Under the above convexity conditions, we can establish that p is a convex function. Indeed, it holds that for any value of the Lagrange multiplier vector $\boldsymbol{\mu}^*$ for the problem (6.4) that

$$\begin{aligned}
q(\boldsymbol{\mu}^*) &= \operatorname*{infimum}_{\boldsymbol{x} \in X} \{f(\boldsymbol{x}) + (\boldsymbol{\mu}^*)^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x})\} \\
&= \operatorname*{infimum}_{\{\,(\boldsymbol{u},\boldsymbol{x}) \in P \times X \mid \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{u}\,\}} \{f(\boldsymbol{x}) + (\boldsymbol{\mu}^*)^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x})\} \\
&= \operatorname*{infimum}_{\{\,(\boldsymbol{u},\boldsymbol{x}) \in P \times X \mid \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{u}\,\}} \{f(\boldsymbol{x}) + (\boldsymbol{\mu}^*)^{\mathrm{T}} \boldsymbol{u}\} \\
&= \operatorname*{infimum}_{\boldsymbol{u} \in P} \ \operatorname*{infimum}_{\{\boldsymbol{x} \in X \mid \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{u}\,\}} \{f(\boldsymbol{x}) + (\boldsymbol{\mu}^*)^{\mathrm{T}} \boldsymbol{u}\}.
\end{aligned}$$

Since $\boldsymbol{\mu}^*$ is assumed to be a Lagrange multiplier, we have that $q(\boldsymbol{\mu}^*) = f^* = p(\mathbf{0}^m)$. By the definition of infimum, then, we have that

$$p(\mathbf{0}^m) \leq p(\boldsymbol{u}) + (\boldsymbol{\mu}^*)^{\mathrm{T}} \boldsymbol{u}, \qquad \boldsymbol{u} \in \mathbb{R}^m,$$

that is, $-\boldsymbol{\mu}^*$ (notice the sign!) is a subgradient of p at $\boldsymbol{u} = \mathbf{0}^m$ (see Definition 6.15). Moreover, by the result in Proposition 6.16(c), p is differentiable at $\mathbf{0}^m$ if and only if p is finite in a neighbourhood of $\mathbf{0}^m$ and $\boldsymbol{\mu}^*$ is a *unique* Lagrange multiplier vector, that is, the Lagrangian dual problem (6.9) has a unique optimal solution. We have therefore proved the following result:

Proposition 6.28 (a sensitivity analysis result) *Suppose that in the inequality constrained problem* (6.4), $f : \mathbb{R}^n \to \mathbb{R}$ *and* $g_i : \mathbb{R}^n \to \mathbb{R}$ $(i = 1, \dots, m)$ *are convex functions and* $X \subseteq \mathbb{R}^n$ *is a convex set. Suppose further that the problem* (6.4) *is feasible, and that the compactness assumption* (6.38) *and Slater condition* (6.15) *hold. Suppose, finally, that the perturbed problem defined in* (6.62) *has an optimal solution in a neighbourhood of* $\boldsymbol{u} = \mathbf{0}^m$, *and that on the set of primal–dual optimal solutions to* (6.4)–(6.9), *the Lagrangian dual optimal solution* $\boldsymbol{\mu}^*$ *is unique. Then, the perturbation function* p *is differentiable at* $\boldsymbol{u} = \mathbf{0}^m$, *and*

$$\nabla p(\mathbf{0}^m) = -\boldsymbol{\mu}^*$$

holds. ∎

It is intuitive that the sign of $\nabla p(\mathbf{0}^m)$ should be non-positive; if a right-hand side of the (less-than) inequality constraints in (6.4) increases, then the feasible set becomes larger. [This means that we might be able to find feasible vectors $\boldsymbol{x}$ in the new problem with $f(\boldsymbol{x}) < f^*$, where f^* is the optimal value of the minimization problem (6.4).]

The result specializes immediately to linear programming problems, which is the problem type where this type of analysis is most often utilized. The proof of differentiability of the perturbation function at zero for that special case can however be done much more simply. (See Section 10.3.1.)

6.7.2 Analysis for differentiable problems

There exist local versions of the analysis valid also for non-convex problems, where we are interested in the effect of a problem perturbation on a KKT point. A special such analysis was recently performed by Bertsekas [Ber04], in which he shows that even when the problem is non-convex and the set of Lagrange multipliers are not unique, a sensitivity analysis is available as long as data is differentiable. Suppose then that in the problem (6.4) the functions f and g_i, $i = 1, \dots, m$ are in C^1 and that X is nonempty. We generalize the concept of a *Lagrange multiplier* to here mean that it is a vector $\boldsymbol{\mu}^*$ associated with a *local* minimum $\boldsymbol{x}^*$ such that

$$\left(\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(\boldsymbol{x}^*)\right)^{\mathrm{T}} \boldsymbol{p} \geq 0, \qquad \boldsymbol{p} \in T_X(\boldsymbol{x}^*), \tag{6.63a}$$

$$\mu_i^* \geq 0, \qquad i = 1, \dots, m, \tag{6.63b}$$

$$\mu_i^* = 0, \qquad i \notin \mathcal{I}(\boldsymbol{x}^*), \tag{6.63c}$$

where we note that $T_X(\boldsymbol{x}^*)$ is the tangent cone to X at $\boldsymbol{x}^*$ (cf. Definition 5.2). Notice that under an appropriate CQ this is equivalent to the KKT conditions, in which case we are simply requiring here that $\boldsymbol{x}^*$ is a local minimum.

In the below result we utilize the notation

$$g_i^+(\boldsymbol{x}) := \text{maximum}\,\{0, g_i(\boldsymbol{x})\}, \qquad i = 1, \dots, m,$$

and let $\boldsymbol{g}^+(\boldsymbol{x})$ be the m-vector of elements $g_i^+(\boldsymbol{x})$, $i = 1, \dots, m$.

Theorem 6.29 (sensitivity from the minimum norm multiplier) *Suppose that $\boldsymbol{x}^*$ is a local minimum in the problem* (6.4), *and that the set of Lagrange multipliers is nonempty. Let $\boldsymbol{\mu}^*$ denote the Lagrange multiplier of minimum Euclidean norm. Then, for every sequence $\{\boldsymbol{x}_k\} \subset X$ of infeasible vectors such that $\{\boldsymbol{x}_k\} \to \boldsymbol{x}^*$ we have that*

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k) \leq \|\boldsymbol{\mu}^*\| \cdot \|\boldsymbol{g}^+(\boldsymbol{x}_k)\| + o(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|). \tag{6.64}$$

Furthermore, if $\boldsymbol{\mu}^* \neq \mathbf{0}^m$ and $T_X(\boldsymbol{x}^*)$ is convex, the above inequality is sharp in the sense that there exists a sequence of infeasible vectors $\{\boldsymbol{x}_k\} \subset X$ such that

$$\lim_{k\to\infty} \frac{f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k)}{\|\boldsymbol{g}^+(\boldsymbol{x}_k)\|} = \|\boldsymbol{\mu}^*\|,$$

and for this sequence

$$\lim_{k\to\infty} \frac{g_i^+(\boldsymbol{x}_k)}{\|\boldsymbol{g}^+(\boldsymbol{x}_k)\|} = \frac{\mu_i^*}{\|\boldsymbol{\mu}^*\|}, \qquad i = 1, \dots, m,$$

holds. ∎

Theorem 6.29 establishes the optimal rate of cost improvement with respect to infeasible constraint perturbations (in effect, those that imply an increase in the feasible set).

We finally remark that under stronger conditions still, even the optimal solution $\boldsymbol{x}^*$ is differentiable. Such a result is reminiscent to the

Implicit Function Theorem, which however only covers equality systems. If we are to study the sensitivity of $\boldsymbol{x}^*$ to changes in the right-hand sides of inequality constraints as well, then the analysis becomes complicated due to the fact that we must be able to predict if some active constraints may become inactive in the process. In some circumstances, different directions of change in the right-hand sides may cause different subsets of the active constraints $\mathcal{I}(\boldsymbol{x}^*)$ at $\boldsymbol{x}^*$ to become inactive, and this would most probably then be a non-differentiable point. A sufficient (but not necessary at least in the case of linear constraints) condition when this cannot happen is when $\boldsymbol{x}^*$ is *strictly complementary*, that is, when there exists a multiplier vector $\boldsymbol{\mu}^*$ where $\mu_i^* > 0$ for every $i \in \mathcal{I}(\boldsymbol{x}^*)$.

6.8 Notes and further reading

Lagrangian duality has been developed in many sources, including early developments by Arrow, Hurwicz, and Uzawa [AHU58], Everett [Eve63], and Falk [Fal67], and later on by Rockafellar [Roc70]. Our development follows to a large extent that of portions of the text books by Bertsekas [Ber99], Bazaraa et al. [BSS93], and Rockafellar [Roc70].

The Relaxation Theorem 6.1 can almost be considered to be folklore, and can be found in a slightly different form in [Wol98, Proposition 2.3].

The traveling salesman problem is an essential model problem in combinatorial optimization. Excellent introductions to the field can be found in [Law76, PaS82, NeW88, Wol98, Sch03]. It was the work in [HWC74, Geo74, Fis81, Fis85], among others, in the 1970s and 1980s on the traveling salesman problem and its relatives that made Lagrangian relaxation and subgradient optimization popular, and it remains most popular within the combinatorial optimization field.

The differentiability properties of convex functions were developed largely by Rockafellar [Roc70], whose text we mostly follow.

Subgradient methods were developed in the Soviet Union in the 1960s, predominantly by Ermol'ev, Polyak, and Shor. Text book treatments of subgradient methods are found, for example, in [Sho85, HiL93, Ber99]. Theorem 6.21 is essentially due to Ermol'ev [Erm66]; the proof stems from [LPS96]. Theorem 6.22 is due to Shepilov [She76]; finally, Theorem 6.23 is due to Polyak [Pol69].

Everett's Theorem is due to Everett [Eve63].

Theorem 6.29 stems from [Ber04, Proposition 1.1].

6.9 Exercises

Exercise 6.1 (numerical example of Lagrangian relaxation) Consider the convex problem to

$$\begin{aligned}
\text{minimize} \quad & \frac{1}{x_1} + \frac{4}{x_2}, \\
\text{subject to} \quad & x_1 + x_2 \leq 4, \\
& x_1, x_2 \geq 0.
\end{aligned}$$

(a) Lagrangian relax the first constraint, and write down the resulting implicit dual objective function and the dual problem. Motivate why the relaxed problem always has a unique optimum, whence the dual objective function is everywhere differentiable.

(b) Solve the implicit Lagrangian dual problem by utilizing that the gradient to a differentiable dual objective function can be expressed by using the functions that are involved in the relaxed constraints and the unique solution to the relaxed problem.

(c) Write down an explicit Lagrangian dual problem, that is, a dual problem only in terms of the Lagrange multipliers. Solve it, and confirm the results in (b).

(d) Find the original problem's optimal solution.

(e) Show that strong duality holds. Why does it? ■

Exercise 6.2 (global optimality conditions) Consider the problem to

$$\begin{aligned}
\text{minimize} \quad & f(\boldsymbol{x}) := x_1 + 2x_2^2 + 3x_3^3, \\
\text{subject to} \quad & x_1 + 2x_2 + x_3 \leq 3, \\
& 2x_1^2 + x_2 \geq 2, \\
& 2x_1 + x_3 = 2, \\
& x_j \geq 0, \qquad j = 1, 2, 3.
\end{aligned}$$

(a) Formulate the Lagrangian dual problem that results from Lagrangian relaxing all but the sign constraints.

(b) State the global primal–dual optimality conditions. ■

Exercise 6.3 (Lagrangian relaxation) Consider the problem to

$$\begin{aligned}
\text{minimize} \quad & f(\boldsymbol{x}) := x_1^2 + 2x_2^2, \\
\text{subject to} \quad & x_1 + x_2 \geq 2, \\
& x_1^2 + x_2^2 \leq 5.
\end{aligned}$$

Find an optimal solution through Lagrangian duality. ■

Exercise 6.4 (Lagrangian relaxation) In many circumstances it is of interest to calculate the Euclidean projection of a vector onto a subspace. Especially, consider the problem to find the Euclidean projection of the vector $\boldsymbol{y} \in \mathbb{R}^n$ onto the null space of the matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, that is, to find an $\boldsymbol{x} \in \mathbb{R}^n$ that solves the problem to

$$\begin{aligned} \text{minimize } & f(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2, \\ \text{subject to } & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}^m, \end{aligned}$$

where $\boldsymbol{A}$ is such that $\operatorname{rank} \boldsymbol{A} = m$.

The solution to this problem is classic: the projection is given explicitly by

$$\boldsymbol{x}^* = \boldsymbol{y} - \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{A}\boldsymbol{y}.$$

If we let $\boldsymbol{P} := \boldsymbol{I}^n - \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{A}$, where $\boldsymbol{I}^n \in \mathbb{R}^{n \times n}$ is the unit matrix, be the *projection matrix*, the formula is simply $\boldsymbol{x}^* = \boldsymbol{P}\boldsymbol{y}$.

Your task is to derive this formula by utilizing Lagrangian duality. Motivate every step made by showing that the necessary properties are fulfilled.

[*Note:* This exercise is similar to that in Example 5.50, but utilizes Lagrangian duality rather than the KKT conditions to derive the projection formula.] ■

Exercise 6.5 (Lagrangian relaxation, exam 040823) Consider the following optimization (linear) problem:

$$\begin{aligned} \text{minimize } & f(x, y) = x - 0.5y, \\ \text{subject to } & -x + y \leq -1, \\ & -2x + y \leq -2, \\ & (x, y) \in \mathbb{R}^2_+. \end{aligned} \tag{6.65}$$

(a) Show that the problem satisfies Slater's constraint qualification. Derive the Lagrangian dual problem corresponding to the Lagrangian relaxation of the two linear inequality constraints, and show that its set of optimal solutions is convex and bounded.

(b) Calculate the set of subgradients of the Lagrangian dual function at the dual points $(1/4, 1/3)^{\mathrm{T}}$ and $(1, 0)^{\mathrm{T}}$. ■

Exercise 6.6 (Lagrangian relaxation) Provide an explicit form of the Lagrangian dual problem for the problem to

$$\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} \ln x_{ij} \\
\text{subject to} \quad & \sum_{i=1}^{m} x_{ij} = b_j, \qquad j = 1, \dots, n, \\
& \sum_{j=1}^{n} x_{ij} = a_i, \qquad i = 1, \dots, m, \\
& \quad x_{ij} \geq 0, \qquad i = 1, \dots, m, \quad j = 1, \dots, n,
\end{aligned}$$

where $a_i > 0$, $b_j > 0$ for all i, j, and where the linear equalities are Lagrangian relaxed. ■

Exercise 6.7 (Lagrangian relaxation) Given is the problem to

$$\begin{aligned}
\underset{\boldsymbol{x}}{\text{minimize}} \quad & f(\boldsymbol{x}) = 2x_1^2 + x_2^2 + x_1 - 3x_2, && \text{(6.66a)} \\
\text{subject to} \quad & x_1^2 + x_2 \geq 8, && \text{(6.66b)} \\
& x_1 \in [1, 3], && \text{(6.66c)} \\
& x_2 \in [2, 5]. && \text{(6.66d)}
\end{aligned}$$

Lagrangian relax the constraint (6.66b) with a multiplier μ. Formulate the Lagrangian dual problem and calculate the dual function's value at $\mu = 1$, $\mu = 2$, and $\mu = 3$. Within which interval lies the optimal value f^*? Also, draw the dual function. ■

Exercise 6.8 (Lagrangian duality for integer problems) Consider the primal problem to

$$\begin{aligned}
\text{minimize} \quad & f(\boldsymbol{x}), \\
\text{subject to} \quad & \boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}^m, \\
& \boldsymbol{x} \in X,
\end{aligned}$$

where $X \subseteq \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$, and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$. If the restrictions $\boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}^m$ are complicating side constraints which are Lagrangian relaxed, we obtain the Lagrangian dual problem to

$$\underset{\boldsymbol{\mu} \geq \mathbf{0}^m}{\text{maximize}}\ q(\boldsymbol{\mu}),$$

where

$$q(\boldsymbol{\mu}) := \underset{\boldsymbol{x} \in X}{\text{minimum}}\ \{f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x})\}, \qquad \boldsymbol{\mu} \in \mathbb{R}^m.$$

(a) Suppose that the set X is finite (for example, consisting of a finite number of integer vectors). Denote the elements of X by $\boldsymbol{x}^p$, $p = 1, \dots, P$. Show that the dual objective function is piece-wise linear.

How many linear segments can it have, at most? Why is it *not always* built up by that many segments?

[*Note:* This property holds regardless of any properties of f and $\boldsymbol{g}$.]

(b) Illustrate the result in (a) on the linear 0/1 problem to find

$$\begin{array}{rlrcrcrcrcl} z^* = & \text{maximum } z = & 5x_1 & + & 8x_2 & + & 7x_3 & + & 9x_4, & & \\ & \text{subject to} & 3x_1 & + & 2x_2 & + & 2x_3 & + & 4x_4 & \leq & 5, \\ & & 2x_1 & + & x_2 & + & 2x_3 & + & x_4 & = & 3, \\ & & x_1 & , & x_2 & , & x_3 & , & x_4 & = & 0/1, \end{array}$$

where the first constraint is considered complicating.

(c) Suppose that the function f and all components of $\boldsymbol{g}$ are linear, and that the set X is a polytope (that is, a bounded polyhedron). Show that the dual objective function is also in this case piece-wise linear. How many linear pieces can it be built from, at most? ■

Exercise 6.9 (Lagrangian relaxation) Consider the problem to

$$\begin{array}{lrcrcll} \text{minimize } z = & 2x_1 & + & x_2, & & & \\ \text{subject to} & x_1 & + & x_2 & \geq & 5, & \\ & x_1 & & & \leq & 4, & \\ & & & x_2 & \leq & 4, & \\ & x_1 & , & x_2 & \geq & 0, & \text{integer.} \end{array}$$

Lagrangian relax the all-embracing constraint. Describe the Lagrangian function and the dual problem. Calculate the Lagrangian dual function at these four points: $\mu = 0, 1, 2, 3$. Give the best lower and upper bounds on the optimal value of the original problem that you have found. ■

Exercise 6.10 (surrogate relaxation) Consider an optimization problem of the form

$$\begin{aligned} &\text{minimize } f(\boldsymbol{x}), \\ &\text{subjectto } g_i(\boldsymbol{x}) \leq 0, \qquad i = 1, \dots, m, \qquad (P) \\ &\qquad\qquad\quad \boldsymbol{x} \in X, \end{aligned}$$

where the functions $f, g_i : \mathbb{R}^n \to \mathbb{R}$ are continuous and the set $X \subset \mathbb{R}^n$ is closed and bounded. The problem is assumed to have an optimal solution, $\boldsymbol{x}^*$. Introduce parameters $\mu_i \geq 0$, $i = 1, \dots, m$, and define

$$\begin{aligned} s(\boldsymbol{\mu}) := \text{minimum } & f(\boldsymbol{x}), \\ \text{subject to } & \sum_{i=1}^{m} \mu_i g_i(\boldsymbol{x}) \leq 0, \qquad (S) \\ & \boldsymbol{x} \in X. \end{aligned}$$

This problem therefore has exactly one explicit constraint.

(a) [weak duality] Show that $\boldsymbol{x}^*$ is a feasible solution to the problem (S) and that $s(\boldsymbol{\mu}) \leq z^*$ therefore always holds, that is, the problem (S) is a *relaxation* of the original one. Motivate also why $\text{maximum}_{\boldsymbol{\mu} \geq \mathbf{0}^m} s(\boldsymbol{\mu}) \leq z^*$ must hold. Explain the potential usefulness of this result!

(b) [example] Consider the linear 0/1 problem

$$\begin{array}{rlrl} z^* = \text{maximum } z = & 5x_1 + 8x_2 + 7x_3 + 9x_4, & & \\ \text{subject to} & 3x_1 + 2x_2 + 3x_3 + 3x_4 \leq & 6, & (1) \\ & 2x_1 + 3x_2 + 3x_3 + 4x_4 \leq & 5, & (2) \\ & 2x_1 + x_2 + 2x_3 + x_4 = & 3, & \\ & x_1\ , \quad x_2\ , \quad x_3\ , \quad x_4 = & 0/1. & \end{array}$$

Surrogate relax the constraints (1) and (2) with multipliers $\mu_1, \mu_2 \geq 0$ and formulate the problem (S). Let $\bar{\boldsymbol{\mu}} = (1, 2)^{\mathrm{T}}$. Calculate $s(\bar{\boldsymbol{\mu}})$.

Consider again the original problem and Lagrangian relax the constraints (1) and (2) with multipliers $\mu_1, \mu_2 \geq 0$. Calculate the Lagrangian dual objective value at $\boldsymbol{\mu} = \bar{\boldsymbol{\mu}}$.

Compare the two results!

(c) [comparison with Lagrangian duality] Let $\boldsymbol{\mu} \geq \mathbf{0}^m$ and

$$q(\boldsymbol{\mu}) := \underset{\boldsymbol{x} \in X}{\text{minimum}}\ f(\boldsymbol{x}) + \sum_{i=1}^{m} \mu_i g_i(\boldsymbol{x}).$$

Show that $q(\boldsymbol{\mu}) \leq s(\boldsymbol{\mu})$, and that

$$\underset{\boldsymbol{u} \geq \mathbf{0}^m}{\text{maximum}}\, q(\boldsymbol{\mu}) \leq \underset{\boldsymbol{u} \geq \mathbf{0}^m}{\text{maximum}}\, s(\boldsymbol{\mu}) \leq z^*$$

holds. ■

Part IV

Linear Optimization

Linear programming: An introduction

VII

Linear programming (LP) models, that is, the collection of optimization models with linear objective functions and polyhedral feasible regions, are very useful in practice. The reason for this is that many real world problems can be described by LP models (even if several approximations must be made first) and, perhaps more importantly, there exist efficient algorithms for solving linear programs; the most famous of them is the Simplex method, which will be presented in Chapter 9. Often, LP models deal with situations where a number of resources (materials, machines, people, land, etcetera) are available and are to be combined to yield several products.

To introduce the concept of linear programming we use a (oversimplified) manufacturing problem. In Section 7.1 we describe the problem. From the problem description we develop an LP model in Section 7.2. It turns out that the LP model only contains two variables. Hence it is possible to solve the problem graphically, which is done in Section 7.3. In Section 7.4 we discuss what happens if the data of the problem is modified, namely, we see how the optimal solution changes if the supply of raw-material or the prices of the products are modified. Finally, in Section 7.5 we develop what we call the linear programming *dual* problem to the manufacturing problem.

7.1 The manufacturing problem

A manufacturer produces two pieces of furniture: tables and chairs. The production of the furniture requires the use of two different pieces of raw-material, large and small pieces. One table is assembled by putting together two pieces of each, while one chair is assembled from one of the larger pieces and two of the smaller pieces (see Figure 7.1).

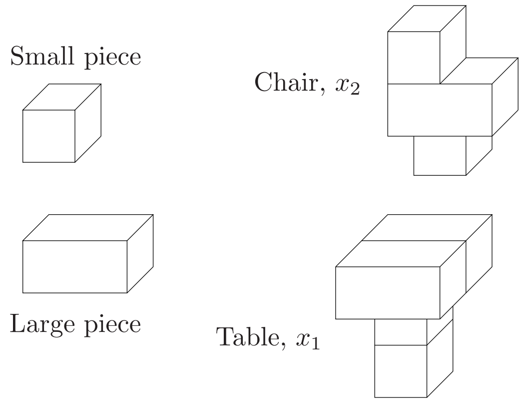


Figure 7.1: Illustration of the manufacturing problem.

When determining the optimal production plan, the manufacturer must take into account that only 6 large and 8 small pieces are available. One table is sold for 1600 SEK, while the chair sells for 1000 SEK. Under the assumption that all items produced can be sold, and that the raw-material has already been paid for, the problem is to determine the production plan that maximizes the total income, within the limited resources.

7.2 A linear programming model

In order to develop a linear programming model for the manufacturing problem we introduce the following variables:

$$\begin{aligned} x_1 &= \text{number of tables manufactured and sold,} \\ x_2 &= \text{number of chairs manufactured and sold,} \\ z &= \text{total income.} \end{aligned}$$

The variable z is, strictly speaking, not a variable, but will be defined by the variables x_1 and x_2.

The income from each product is given by the price of the product multiplied by the number of products sold. Hence the total income becomes

$$z = 1600x_1 + 1000x_2. \tag{7.1}$$

Given that we produce x_1 tables and x_2 chairs the required number of large pieces is $2x_1 + x_2$ and the required number of small peaces is

$2x_1 + 2x_2$. But only 6 large pieces and 8 small pieces are available, so we must have that

$$2x_1 + x_2 \leq 6, \tag{7.2}$$
$$2x_1 + 2x_2 \leq 8. \tag{7.3}$$

Further, it is impossible to produce a negative number of chairs or tables, which gives that

$$x_1,\, x_2 \geq 0. \tag{7.4}$$

(Also, the number of chairs and tables produced must be integers, but we will not take that into account here.)

Now the objective is to maximize the total income, so if we put the income function (7.1) together with the constraints (7.2)–(7.4) we get the following linear programming model:

$$\begin{aligned} \text{maximize} \quad z = \; & 1600x_1 + 1000x_2 \\ \text{subject to} \quad & 2x_1 + x_2 \leq 6, \\ & 2x_1 + 2x_2 \leq 8, \\ & x_1, \quad x_2 \geq 0. \end{aligned} \tag{7.5}$$

7.3 Graphical solution

The feasible region of the linear programming formulation (7.5) is graphed in Figure 7.2. The figure also includes lines corresponding to various values of the cost function. For example, the line $z = 0 = 1600x_1 + 1000x_2$ passes through the origin, and the line $z = 2600 = 1600x_1 + 1000x_2$ passes through the point $(1, 1)^{\mathrm{T}}$. We see that the value of the cost function increases as these lines move upward in the value of z, and it follows that the optimal solution is $\boldsymbol{x}^* = (2, 2)^{\mathrm{T}}$ and $z^* = 5200$. Observe that the optimal solution is an extreme point, which is in accordance with Remark 4.12. This fact will be very important in the development of the Simplex method in Chapter 9, and established in Theorem 8.10.

7.4 Sensitivity analysis

In this section we investigate how the optimal solution changes if the data of the problem is changed. We consider three different changes (made independent of each other), namely

1. an increase in the number of large pieces available;

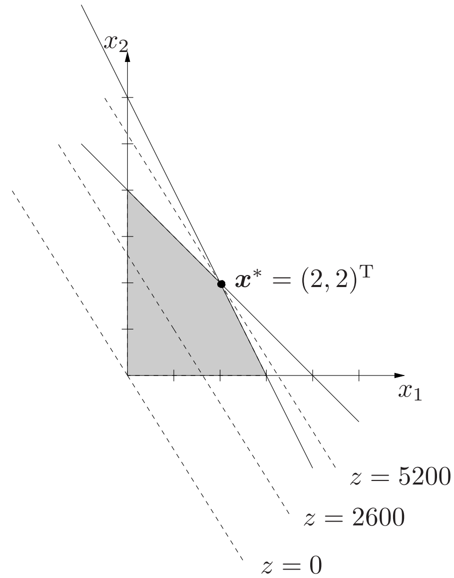


Figure 7.2: Graphical solution of the manufacturing problem.

2. an increase in the number of small pieces available; and
3. a decrease in the price of the tables.

7.4.1 An increase in the number of large pieces available

Assume that the number of large pieces available increases from 6 to 7. Then the linear program becomes

$$
\begin{array}{llrrl}
\text{maximize} & z = & 1600x_1 & +1000x_2 & \\
\text{subject to} & & 2x_1 & +x_2 & \leq 7, \\
& & 2x_1 & +2x_2 & \leq 8, \\
& & x_1, & x_2 & \geq 0.
\end{array}
$$

The feasible region is shown in Figure 7.3.

We see that the optimal solution becomes $(3,1)^{\mathrm{T}}$ and $z^* = 5800$, which means that an additional large piece increases the income by $5800 - 5200 = 600$. Hence the *shadow price* of the large pieces is 600. The figure also illustrates what happens if the number of large pieces is 8. Then the optimal solution becomes $(4,0)^{\mathrm{T}}$ and $z^* = 6400$. But what

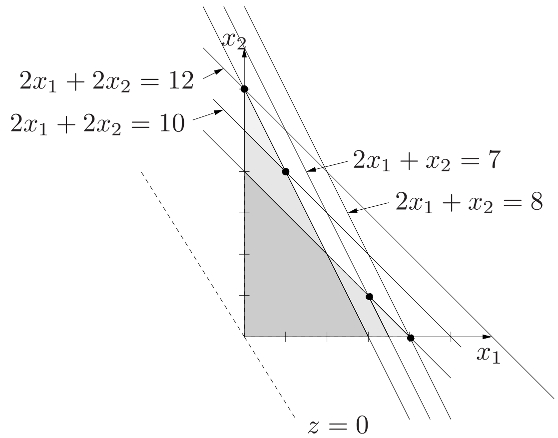


Figure 7.3: An increase in the number of large and small pieces available.

happens if we increase the number of large pieces further? From the figure it follows that the optimal solution will not change (since $x_2 \geq 0$ must apply), so an increase larger than 2 in the number of large pieces gives no further income.

7.4.2 An increase in the number of small pieces available

Starting from the original setup, in the same manner as for the large pieces it follows from Figure 7.3 that two additional small pieces give the new optimal solution $\boldsymbol{x}^* = (1, 4)^{\mathrm{T}}$ and $z^* = 5600$, so the income per additional small piece is $(5600 - 5200)/2 = 200$. Hence the shadow price of the small pieces is 200. However, no more than 4 small pieces are worth this price, since $x_1 \geq 0$ must apply.

7.4.3 A decrease in the price of the tables

Now assume that the price of tables is decreased from 1600 to 800. The new linear program becomes

$$
\begin{array}{llrr}
\text{maximize} & z = & 800x_1 & +1000x_2 \\
\text{subject to} & & 2x_1 & +x_2 \leq 6, \\
& & 2x_1 & +2x_2 \leq 8, \\
& & x_1, & x_2 \geq 0.
\end{array}
$$

This new situation is illustrated in Figure 7.4, from which we see that

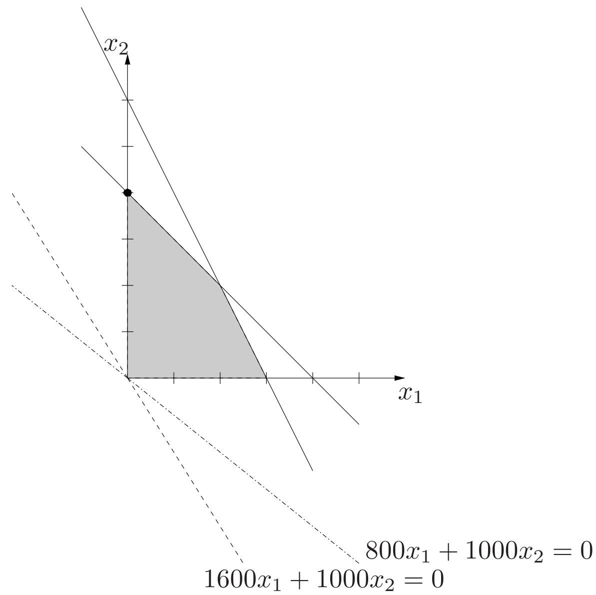


Figure 7.4: A decrease in the price of the tables.

the optimal solution is $(0, 4)^{\mathrm{T}}$, that is, we will not produce any tables. This is natural, since it takes the same number of small pieces to produce a table and a chair, but the table requires one more large piece, and in addition the price of a table is lower than that of a chair.

7.5 The dual of the manufacturing problem

7.5.1 A competitor

Suppose that another manufacturer (let us call them Billy) produce book shelves whose raw material is identical to those used for the table and chairs, that is, the small and large pieces. Billy wish to expand their production, and are interested in acquiring the resources that "our" factory sits on. Let us ask ourselves two questions, which (as we shall see) have identical answers: (1) what is the lowest bid (price) for the total capacity at which we are willing to sell?; (2) what is the highest bid (price) that Billy are prepared to offer for the resources? The answer to those two questions is a measure of the wealth of the company in terms of their resources.

7.5.2 A dual problem

To study the problem, we introduce the variables

$y_1 =$ the price which Billy offers for each large piece,
$y_2 =$ the price which Billy offers for each small piece,
$w =$ the total bid which Billy offers.

In order to accept to sell our resources, it is reasonable to require that the price offered is at least as high as the value that the resource represents in our optimal production plan, as otherwise we would earn more by using the resource ourselves. Consider, for example, the net income on a table sold. It is 1600 SEK, and for that we use two large and two small pieces. The bid would therefore clearly be too low unless $2y_1 + 2y_2 \geq 1600$. The corresponding requirement for the chairs is that $y_1 + 2y_2 \geq 1000$.

Billy is interested in minimizing the total bid, under the condition that the offer is accepted. Observing that y_1 and y_2 are prices and therefore non-negative, we have the following mathematical model for Billy's problem:

$$\begin{aligned} \text{minimize} \quad w = \ & 6y_1 + 8y_2 \\ \text{subject to} \quad & 2y_1 + 2y_2 \geq 1600, \\ & y_1 + 2y_2 \geq 1000, \\ & y_1, \quad y_2 \geq 0. \end{aligned} \tag{7.6}$$

This is usually called the *dual problem* of our production planning problem (which would then be the *primal problem*).

The optimal solution to this problem is $\boldsymbol{y}^* = (600, 200)^{\mathrm{T}}$. The total offer is $w^* = 5200$.

Remark 7.1 (the linear programming dual) Observe that the dual problem (7.6) is in accordance with the Lagrangian duality theory in Section 6.2.4. The linear programming dual will be discussed further in Chapter 10. ■

7.5.3 Interpretations of the dual optimal solution

From the above we see that the dual optimal solution is identical to the shadow prices for the resource (capacity) constraints. (This is indeed a general conclusion in linear programming.) To motivate that this is reasonable in our setting, we may consider Billy as a fictitious competitor only, which we use together with the dual problem to measure the value of our resources. This (fictitious) measure can be used to create internal prices in a company in order to utilize limited resources as efficiently as possible, especially if the resource is common to several independent sub-units. The price that the dual optimal solution provides will then be a price directive for the sub-units, that will make them utilize the scarce resource in a manner which is optimal for the overall goal.

We note that the optimal value of the production ($z^* = 5200$) agrees with the total value $w^* = 5200$ of the resources in our company. (This is also a general result in linear programming; see the Strong Duality Theorem 10.6.) Billy will of course not pay more than what the resource is worth, but can at the same time not offer less than the profit that our company can make ourselves, since we would then not agree to sell. It follows immediately that for each feasible production plan $\boldsymbol{x}$ and price $\boldsymbol{y}$, it holds that $z \leq w$, since

$$\begin{aligned} z = 1600x_1 + 1000x_2 &\leq (2y_1 + 2y_2)x_1 + (y_1 + 2y_2)x_2 \\ &= y_1(2x_1 + x_2) + y_2(2x_1 + 2x_2) \leq 6y_1 + 8y_2 = w, \end{aligned}$$

where in the inequalities we utilize all the constraints of the primal and dual problems. (Also this fact is general in linear programming; see the Weak Duality Theorem 10.4.) So, each offer accepted (from our point of view) must necessarily be an upper bound on our own possible profit, and this upper bound is what Billy wish to minimize in the dual problem.

Linear programming models

VIII

We begin this chapter with a presentation of the axioms underlying the use of linear programming models and discuss the modelling process. Then, in Section 8.2, the geometry of linear programming is studied. It is shown that every linear program can be transformed into the *standard form* which is the form that the Simplex method uses. Further, we introduce the concept of *basic feasible solution* and discuss its connection to extreme points. A version of the Representation Theorem adapted to the standard form is presented, and we show that if there exists an optimal solution to a linear program in standard form, then there exists an optimal solution among the basic feasible solutions. Finally, we define adjacency between extreme points and give an algebraic characterization of adjacency which actually proves that the Simplex method at each iteration step moves from one extreme point to an adjacent one.

8.1 Linear programming modelling

Many real world situations can be modelled as linear programs. However, the applicability of a linear program requires certain axioms to be fulfilled. Hence, often approximations of the real world problem must be made prior to the formulation of a linear program. The axioms underlying the use of linear programming models are:

- proportionality (linearity, e.g., no economies-of-scales, no fixed costs);
- additivity (no substitute-time-effects);
- divisibility (continuity); and
- determinism (no randomness).

George Dantzig presented the linear programming model and the Simplex method for solving it at an econometrics conference in Wisconsin in the late 40s. The economist Hotelling stood up, devastatingly smiling, and stated that "But we all know the world is nonlinear." The young graduate George Dantzig could not respond, but was defended by John von Neumann, who stood up and concluded that "The speaker titled his talk 'linear programming' and carefully stated his axioms. If you have an application that satisfies the axioms, well use it. If it does not, then don't"; he sat down, and Hotelling was silenced. (See Dantzig's account of the early history of linear programming in [LRS91, pp. 19–31].)

Now if the problem considered (perhaps after approximations) fulfills the axioms above, then it can be formulated as a linear program. However, in practical modelling situations we usually do not talk about the axioms; they naturally appear when a linear program is formulated.

To formulate a real world problem as a linear program is an art in itself, and unfortunately there is little theory to help in formulating the problem in this way. The general approach can however be described by two steps:

1. Prepare a list of all the decision variables in the problem. This list must be complete in the sense that if an optimal solution providing the values of each of the variables is obtained, then the decision maker should be able to translate it into an optimum policy that can be implemented.
2. Use the variables from step 1 to formulate all the constraints and the objective function of the problem.

We illustrate the two-step modelling process by an example.

Example 8.1 (the transportation problem) In the transportation problem we have a set of nodes or places called *sources*, which have a commodity available for shipment, and another set of places called *demand centers*, or *sinks*, which require this commodity. The amount of commodity available at each source and the amount required at each demand center are specified, as well as the cost per unit of transporting the commodity from each source to each demand center. The problem is to determine the quantity to be transported from each source to each demand center, so as to meet all the requirements at minimum total shipping cost.

Consider the problem where the commodity is iron ore, the sources are found at mines 1 and 2, where the ore is produced, and the demand centers are three steel plants. The unit costs of shipping ore from each mine to each steel plant are given in Table 8.1.

Table 8.1: Unit cost of shipping ore from mine to steel plant (KSEK per Mton).

Plant	1	2	3
Mine 1	9	16	28
Mine 2	14	29	19

Table 8.2: Amount of ore available at the mines (Mtons).

Mine 1	103
Mine 2	197

Further, the amount of ore available at the mines and the Mtons of ore required at each steel plant are given in the Tables 8.2 and 8.3.

We use the two-step modelling process to formulate a linear programming model.

Step 1: The activities in the transportation model are to ship ore from mine i to steel plant j for $i = 1, 2$ and $j = 1, 2, 3$. It is convenient to represent the variables corresponding to the levels at which these activities are carried out by double subscripted symbols. Hence, for $i = 1, 2$ and $j = 1, 2, 3$, we introduce the following variables:

x_{ij} = amount of ore (in Mtons) shipped from mine i to steel plant j.

We also introduce a variable corresponding to the total cost of the shipping:

$$z = \text{total shipping cost.}$$

Step 2: The transportation problem considered is illustrated in Figure 8.1.

Table 8.3: Ore requirements at the steel plants (Mtons).

Plant 1	71
Plant 2	133
Plant 3	96

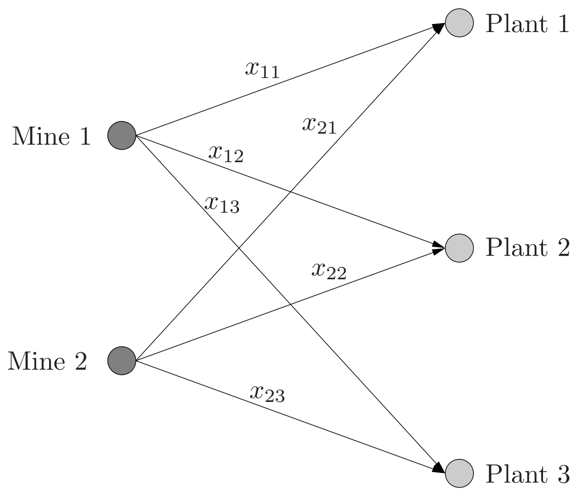


Figure 8.1: Illustration of the transportation problem.

The items in this problem are the ore at various locations. Consider the ore at mine 1. According to Table 8.2 there are only M103 tons of it available, and the amount of ore shipped out of mine 1, which is $x_{11} + x_{12} + x_{13}$, cannot exceed the amount available, leading to the constraint

$$x_{11} + x_{12} + x_{13} \leq 103.$$

Likewise, if we consider ore at mine 2 we get the constraint

$$x_{21} + x_{22} + x_{23} \leq 197.$$

Further, at steel plant 1 according to Table 8.3 there are at least 71 Mtons of ore required, so the amount of ore shipped to steel plant 1 has to be greater than or equal to this amount, leading to the constraint

$$x_{11} + x_{21} \geq 71.$$

In the same manner, for the steel plants 2 and 3 we get

$$x_{12} + x_{22} \geq 133,$$
$$x_{13} + x_{23} \geq 96.$$

Of course it is impossible to ship a negative amount of ore, yielding the constraints

$$x_{ij} \geq 0, \quad i = 1, 2, \quad j = 1, 2, 3.$$

From Table 8.1 it follows that the total cost of the shipping is (in KSEK)

$$z = 9x_{11} + 16x_{12} + 28x_{13} + 14x_{21} + 29x_{22} + 19x_{23}.$$

Finally, since the objective is to minimize the total cost we get the following linear programming model:

$$\begin{array}{llrrrrrrl}
\text{minimize} & z = & 9x_{11} & +16x_{12} & +28x_{13} & +14x_{21} & +29x_{22} & +19x_{23} & \\
\text{subject to} & & x_{11} & +x_{12} & +x_{13} & & & & \leq 103, \\
& & & & & x_{21} & +x_{22} & +x_{23} & \leq 197, \\
& & x_{11} & & & +x_{21} & & & \geq \ \ 71, \\
& & & x_{12} & & & +x_{22} & & \geq 133, \\
& & & & x_{13} & & & +x_{23} & \geq \ \ 96, \\
& & x_{11}, & x_{12}, & x_{13}, & x_{21}, & x_{22}, & x_{23} & \geq \ \ \ 0.
\end{array}$$

Normally, transportation problems consider a large number of sources and demand centers and then it is convenient to use general notation leading to a compact formulation. Assume that we have N sources and M demand centers. For $i = 1, \ldots, N$ and $j = 1, \ldots, M$, introduce the variables

x_{ij} = amount of commodity shipped from source i to demand center j,

and let

$$z = \text{total shipping cost.}$$

Further for $i = 1, \ldots, N$ and $j = 1, \ldots, M$ introduce the shipping costs

c_{ij} = unit cost of shipping commodity from source i to demand center j.

Also, let

s_i = amount of commodity available at source i, $\quad i = 1, \ldots, N$,
d_j = amount of commodity required at demand center j, $\quad j = 1, \ldots, M$.

Now consider source i. The amount of commodity available is given by s_i, which gives the constraint

$$\sum_{j=1}^{M} x_{ij} \leq s_i.$$

Similarly, the amount of commodity required at demand center j is given by d_j, leading to the constraint

$$\sum_{i=1}^{N} x_{ij} \geq d_j.$$

It is impossible to ship a negative amount of commodity, which gives

$$x_{ij} \geq 0, \qquad i = 1, \dots, N, \quad j = 1, \dots, M.$$

Finally, the total cost for shipping is

$$z = \sum_{i=1}^{N} \sum_{j=1}^{M} c_{ij} x_{ij},$$

and we end up with the following linear program:

$$\begin{aligned}
\text{minimize} \quad & z = \sum_{i=1}^{N} \sum_{j=1}^{M} c_{ij} x_{ij} \\
\text{subject to} \quad & \sum_{j=1}^{M} x_{ij} \leq s_i, \quad i = 1, \dots, N, \\
& \sum_{i=1}^{N} x_{ij} \geq d_j, \quad j = 1, \dots, M, \\
& \quad x_{ij} \geq 0, \quad i = 1, \dots, N, \quad j = 1, \dots, M.
\end{aligned}$$

If, for some reason, it is impossible to transport any commodities from a source i to a sink j, then we may either remove this variable altogether from the model, or, more simply, give it the unit price $c_{ij} = +\infty$.

Note, finally, that there exists a feasible solution to the transportation problem if and only if $\sum_{i=1}^{N} s_i \geq \sum_{j=1}^{M} d_j$. ∎

8.2 The geometry of linear programming

In Section 3.2 we studied the class of feasible sets in linear programming, namely the sets of polyhedra; they are sets of the form

$$P = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \,\},$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. In particular we proved the Representation Theorem 3.22 and promised that it would be important in the development of the Simplex method. In this section we revisit this polyhedron. Here, however, we will consider polyhedra of the form

$$P = \{ \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}; \quad \boldsymbol{x} \geq \boldsymbol{0}^n \}, \tag{8.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, and $\boldsymbol{b} \in \mathbb{R}^m$ is such that $\boldsymbol{b} \geq \boldsymbol{0}^m$. The advantage of this form is that the constraints (except for the non-negativity constraints) are *equalities*, which admits pivot operations to to be carried out. The Simplex method uses pivot operations at each iteration step and hence it is necessary that the polyhedron (that is, the feasible region) is represented in the form (8.1). This is, however, not a restriction, as we will see in Section 8.2.1, since every polyhedron can be transformed into this form! We will use the term *standard form* when a polyhedron is represented in the form (8.1). In Section 8.2.2 we introduce the concept of *basic feasible solution* and show that each basic feasible solution corresponds to an extreme point. We also restate the Representation Theorem 3.22 and prove that if there exists an optimal solution to a linear program, then there exists an optimal solution among the extreme points. Finally, in Section 8.2.3, a strong connection between basic feasible solutions and adjacent extreme points is discussed. This connection shows that the Simplex method at each iteration step moves from an extreme point to an adjacent extreme point.

8.2.1 Standard form

A linear programming problem in *standard form* is a problem of the form

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n, \end{aligned}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \geq \boldsymbol{0}^m$. The purpose of this section is to show that every linear program can be transformed into the standard form. In order to do that we must

- express the objective function in minimization form;
- transform all the constraints into equality constraints with non-negative right-hand sides; and
- transform any unrestricted and non-positive variables into non-negative ones.

Objective function

Constant terms in the objective function will not change the set of optimal solutions and can therefore be eliminated. If the objective is to

$$\text{maximize } z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x},$$

then change the objective function so that the objective becomes

$$\text{minimize } \tilde{z} = -z = -\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}.$$

This change does not affect the set of feasible solutions to the problem and the equation

$$[\text{maximum value of } z] = -[\text{minimum value of } \tilde{z}]$$

can be used to get the maximum value of the original objective function.

Inequality constraints and negative right-hand sides

Consider the inequality constraint

$$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} \leq b,$$

where $\boldsymbol{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. By introducing a non-negative *slack variable* s this constraint can be written as

$$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} + s = b, \tag{8.2a}$$
$$s \geq 0, \tag{8.2b}$$

which has the desired form of an equation. Suppose that $b < 0$. By multiplying both sides of (8.2a) by -1 the negativity in the right-hand side is eliminated and we are done. Similarly, a constraint of the form

$$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} \geq b,$$

can be written as

$$\begin{aligned} \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} - s &= b, \\ s &\geq 0. \end{aligned}$$

We call such variables s *surplus variables*.

Remark 8.2 (on the role of slack and surplus variables) Slack and surplus may appear to be only help variables, but they often have a clear interpretation as decision variables. Consider, for example, the model (7.5) of a furniture production problem. The two inequality constraints are associated with the capacity of production stemming from the availability of raw material. Suppose then that we introduce slack variables in these constraints, which leads to the equivalent problem to

$$\begin{aligned}
\underset{(\boldsymbol{x},\boldsymbol{s})}{\text{maximize}} \quad & z = 1600x_1 + 1000x_2, && \text{(8.3a)}\\
\text{subject to} \quad & 2x_1 + x_2 + s_1 = 6, && \text{(8.3b)}\\
& 2x_1 + 2x_2 + s_2 = 8, && \text{(8.3c)}\\
& x_1, \quad x_2, \quad s_1, \quad s_2 \geq 0. && \text{(8.3d)}
\end{aligned}$$

The new variables s_1 and s_2 have the following interpretation: the value of s_i $(i = 1, 2)$ is the level of inventory (or, remaining capacity of raw material of type i) that will be left unused when the production plan (x_1, x_2) has been implemented. This interpretation makes it clear that the values of s_1 and s_2 are clear consequences of our decision-making.

Surplus variables have a corresponding interpretation. In the case of the transportation model of the previous section, a demand constraint $(\sum_{i=1}^{N} x_{ij} \geq d_j,\ j = 1, \dots, M)$ may be fulfilled with equality (in which case the customer gets an amount exactly according to the demand) or it is fulfilled with strict inequality (in which case the customer gets a surplus of the product asked for). ■

Unrestricted and non-positive variables

Consider the linear program

$$\begin{aligned}
\text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} && \text{(8.4)}\\
\text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},\\
& x_2 \leq 0,\\
& x_j \geq 0, \quad j = 3, \dots, n,
\end{aligned}$$

which is assumed to be of standard form except for the unrestricted variable x_1 and the non-positive variable x_2. The x_2-variable can be replaced by the non-negative variable $\tilde{x}_2 = -x_2$. The x_1-variable can be transformed into the difference of two non-negative variables. Namely, introduce the variables $x_1^+ \geq 0$ and $x_1^- \geq 0$ and let $x_1 = x_1^+ - x_1^-$. Then by substituting x_1 with $x_1^+ - x_1^-$ wherever it occurs we have transformed the problem into standard form. The drawback of this method to handle unrestricted variables is that often the resulting problem is numerically unstable. However, there are other techniques to handle unrestricted variables that overcome this problem; one of them is given in Exercise 8.5.

Example 8.3 (standard form) Consider the linear program

$$\begin{array}{lrl} \text{maximize} & z = & 9x_1 - 7x_2 + 3y_1 \\ \text{subject to} & & 3x_1 + x_2 - y_1 \leq 1, \\ & & 4x_1 - x_2 + 2y_1 \geq 3, \\ & & x_1, \quad x_2 \qquad \geq 0. \end{array}$$

In order to transform the objective into the minimization form, let

$$\tilde{z} = -z = -9x_1 + 7x_2 - 3y_1.$$

Further, by introducing the slack variable s_1 and the surplus variable s_2 the constraints can be transformed into an equality form by

$$\begin{array}{rrrrrl} 3x_1 & +x_2 & -y_1 & +s_1 & & = 1, \\ 4x_1 & -x_2 & +2y_1 & & -s_2 & = 3, \\ x_1, & x_2, & & s_1, & s_2 & \geq 0. \end{array}$$

Finally, by introducing the variables y_1^+ and y_1^- we can handle the unrestricted variable y_1 by substituting it by $y_1^+ - y_1^-$ wherever it occurs. We arrive at the standard form to

$$\begin{array}{lrrrrrrrl} \text{minimize} & \tilde{z} = & -9x_1 & +7x_2 & -3y_1^+ & +3y_1^- & & & \\ \text{subject to} & & 3x_1 & +x_2 & -y_1^+ & +y_1^- & +s_1 & & = 1, \\ & & 4x_1 & -x_2 & +2y_1^+ & -2y_1^- & & -s_2 & = 3, \\ & & x_1, & x_2, & y_1^+, & y_1^-, & s_1, & s_2 & \geq 0. \end{array}$$

■

8.2.2 Basic feasible solutions and the Representation Theorem

In this section we introduce the concept of *basic feasible solution* and show the equivalence between extreme points and a basic feasible solutions. From this we can draw the conclusion that if there exists an optimal solution then there exists an optimal solution among the basic feasible solutions. This fact is crucial in the Simplex method which searches for an optimal solution among the basic feasible solutions.

Consider a linear program in standard form,

$$\begin{array}{ll} \text{minimize} & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n, \end{array} \tag{8.5}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with rank $\boldsymbol{A} = m$ (otherwise, we can always delete rows), $n > m$, and $\boldsymbol{b} \in \mathbb{R}^m$ is such that $\boldsymbol{b} \geq \boldsymbol{0}^m$. A point $\tilde{\boldsymbol{x}}$ is a *basic solution* of (8.5) if

1. the equality constraints are satisfied at $\tilde{\boldsymbol{x}}$, that is, $\boldsymbol{A}\tilde{\boldsymbol{x}} = \boldsymbol{b}$; and
2. the columns of $\boldsymbol{A}$ corresponding to the non-zero components of $\tilde{\boldsymbol{x}}$ are linearly independent.

A basic solution that also satisfies the non-negativity constraints is called a *basic feasible solution*, or, in short, a BFS.

Since rank $\boldsymbol{A} = m$, we can solve the system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ by selecting m variables of $\boldsymbol{x}$ corresponding to m linearly independent columns of $\boldsymbol{A}$. Hence, we partition the columns of $\boldsymbol{A}$ in two parts: one with $n - m$ columns of $\boldsymbol{A}$ corresponding to components of $\boldsymbol{x}$ that are set to 0; these are called the *non-basic* variables and are denoted by the subvector $\boldsymbol{x}_N \in \mathbb{R}^{n-m}$. The other represents the *basic variables*, and are denoted by $\boldsymbol{x}_B \in \mathbb{R}^m$. According to this partition,

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_B \\ \boldsymbol{x}_N \end{pmatrix}, \quad \boldsymbol{A} = (\boldsymbol{B}, \boldsymbol{N}),$$

which yields that

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{B}\boldsymbol{x}_B + \boldsymbol{N}\boldsymbol{x}_N = \boldsymbol{b}.$$

Since $\boldsymbol{x}_N = \boldsymbol{0}^{n-m}$ by construction, we get the basic solution

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_B \\ \boldsymbol{x}_N \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}^{-1}\boldsymbol{b} \\ \boldsymbol{0}^{n-m} \end{pmatrix}.$$

Further, if $\boldsymbol{B}^{-1}\boldsymbol{b} \geq \boldsymbol{0}^m$ then $\boldsymbol{x}$ is a basic feasible solution.

Remark 8.4 (degenerate solution) If more than $n - m$ variables are zero at a solution $\boldsymbol{x}$, then the partition is not unique, that is, the solution $\boldsymbol{x}$ corresponds to more than one basic solution. Such a solution is called *degenerate*. ■

Example 8.5 (partitioning) Consider the linear program

$$\begin{array}{lrrrrrrl}
\text{minimize} & z = & 4x_1 & +3x_2 & +7x_3 & -2x_4 & & \\
\text{subject to} & & x_1 & & -x_3 & & & = 3, \\
& & x_1 & -x_2 & & -2x_4 & & = 1, \\
& & 2x_1 & & & +x_4 & +x_5 & = 7, \\
& & x_1, & x_2, & x_3, & x_4, & x_5 & \geq 0.
\end{array}$$

The constraint matrix and the right-hand side vector are given by

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & -2 & 0 \\ 2 & 0 & 0 & 1 & 1 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} 3 \\ 1 \\ 7 \end{pmatrix}.$$

(a) The partition $\boldsymbol{x}_B = (x_2, x_3, x_4)^{\mathrm{T}}$, $\boldsymbol{x}_N = (x_1, x_5)^{\mathrm{T}}$,

$$\boldsymbol{B} = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & -2 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{N} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{pmatrix},$$

corresponds the basic solution

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_B \\ \boldsymbol{x}_N \end{pmatrix} = \begin{pmatrix} x_2 \\ x_3 \\ x_4 \\ x_1 \\ x_5 \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}^{-1}\boldsymbol{b} \\ \boldsymbol{0}^2 \end{pmatrix} = \begin{pmatrix} -15 \\ -3 \\ 7 \\ 0 \\ 0 \end{pmatrix}.$$

This is, however, not a basic *feasible* solution (since x_2 and x_3 are negative).

(b) Now take another partition, $\boldsymbol{x}_B = (x_1, x_2, x_5)^{\mathrm{T}}$, $\boldsymbol{x}_N = (x_3, x_4)^{\mathrm{T}}$,

$$\boldsymbol{B} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & -1 & 0 \\ 2 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{N} = \begin{pmatrix} -1 & 0 \\ 0 & -2 \\ 0 & 1 \end{pmatrix}.$$

This partition corresponds to the basic solution

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_B \\ \boldsymbol{x}_N \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_5 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}^{-1}\boldsymbol{b} \\ \boldsymbol{0}^2 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ 1 \\ 0 \\ 0 \end{pmatrix},$$

which is in fact a basic feasible solution.

(c) Further, the partition $\boldsymbol{x}_B = (x_2, x_4, x_5)^{\mathrm{T}}$, $\boldsymbol{x}_N = (x_1, x_3)^{\mathrm{T}}$,

$$\boldsymbol{B} = \begin{pmatrix} 0 & 0 & 0 \\ -1 & -2 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad \boldsymbol{N} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 2 & 0 \end{pmatrix},$$

does not correspond to a basic solution since the system $\boldsymbol{B}\boldsymbol{x}_B = \boldsymbol{b}$ is infeasible.

(d) Finally, the partition $\boldsymbol{x}_B = (x_1, x_4, x_5)^{\mathrm{T}}$, $\boldsymbol{x}_N = (x_2, x_3)^{\mathrm{T}}$,

$$\boldsymbol{B} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & -2 & 0 \\ 2 & 1 & 1 \end{pmatrix}, \quad \boldsymbol{N} = \begin{pmatrix} 0 & -1 \\ -1 & 0 \\ 0 & 0 \end{pmatrix},$$

corresponds to the basic feasible solution

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_B \\ \boldsymbol{x}_N \end{pmatrix} = \begin{pmatrix} x_1 \\ x_4 \\ x_5 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}^{-1}\boldsymbol{b} \\ \boldsymbol{0}^2 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

which is *degenerate* (since a basic variables, x_5, has value zero). ■

Remark 8.6 (partitioning) The above partitioning technique will be used frequently in what follows and from now on when we say that $\boldsymbol{A} = (\boldsymbol{B}, \boldsymbol{N})$ is a *partition* of $\boldsymbol{A}$ we will always mean that the columns of $\boldsymbol{A}$ and the variables of $\boldsymbol{x}$ have been rearranged so that $\boldsymbol{B}$ corresponds to the basic variables $\boldsymbol{x}_B$ and $\boldsymbol{N}$ to the non-basic variables $\boldsymbol{x}_N$. ■

We are now ready to prove the equivalence between extreme points and basic feasible solutions.

Theorem 8.7 (equivalence between extreme point and BFS) *A point $\boldsymbol{x}$ is an extreme point of the set $\{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b};\ \boldsymbol{x} \geq \boldsymbol{0}^n \,\}$ if and only if it is a basic feasible solution.*

Proof. Let $\boldsymbol{x}$ be a basic feasible solution with the corresponding partition $\boldsymbol{A} = (\boldsymbol{B}, \boldsymbol{N})$, where $\operatorname{rank} \boldsymbol{B} = m$ (such a partition exists since $\operatorname{rank} \boldsymbol{A} = m$). Then the equality subsystem of

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} &= \boldsymbol{b}, \\ \boldsymbol{x} &\geq \boldsymbol{0}^n \end{aligned}$$

is given by

$$\begin{aligned} \boldsymbol{B}\boldsymbol{x}_B + \boldsymbol{N}\boldsymbol{x}_N &= \boldsymbol{b}, \\ \boldsymbol{x}_N &= \boldsymbol{0}^{n-m} \end{aligned}$$

(if some of the basic variables equals zero we get additional rows but these will not affect the proof). Since $\operatorname{rank} \boldsymbol{B} = m$ it follows that

$$\operatorname{rank} \begin{pmatrix} \boldsymbol{B} & \boldsymbol{N} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} = n.$$

The theorem then follows from Theorem 3.17. ■

Remark 8.8 (degenerate extreme point) An extreme point that corresponds to more than one BFS is degenerate. This typically occurs when we have redundant constraints. ∎

We present a reformulation of the Representation Theorem 3.22 that is adapted to the standard form. Consider the polyhedral cone $C = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}^m;\ \boldsymbol{x} \geq \boldsymbol{0}^n \,\}$. From Theorem 3.28 it follows that C is finitely generated, that is, there exist vectors $\boldsymbol{d}^1, \dots, \boldsymbol{d}^r \in \mathbb{R}^n$ such that

$$C = \operatorname{cone}\{\boldsymbol{d}^1, \dots, \boldsymbol{d}^r\} = \left\{ \boldsymbol{x} \in \mathbb{R}^n \;\middle|\; \boldsymbol{x} = \sum_{i=1}^{r} \alpha_i \boldsymbol{d}^i; \quad \alpha_1, \dots, \alpha_r \geq 0 \right\}.$$

There are, of course, infinitely many ways to generate a certain polyhedral cone C. Assume that $C = \operatorname{cone}\{\boldsymbol{d}^1, \dots, \boldsymbol{d}^r\}$. If there exists a vector $\boldsymbol{d}^i \in \{\boldsymbol{d}^1, \dots, \boldsymbol{d}^r\}$ such that

$$\boldsymbol{d}^i \in \operatorname{cone}\left\{ \{\boldsymbol{d}^1, \dots, \boldsymbol{d}^r\} \setminus \{\boldsymbol{d}^i\} \right\},$$

then $\boldsymbol{d}^i$ is not necessarily in the description of C. If we similarly continue to remove vectors from $\{\boldsymbol{d}^1, \dots, \boldsymbol{d}^r\}$ one at a time, we end up with a set generating C such that none of the vectors of the set can be written as a non-negative linear combination of the others. Such a set is naturally called the set of *extreme directions* of C (cf. Definition 3.11 of extreme point).

Theorem 8.9 (Representation Theorem) *Let $P = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b};\ \boldsymbol{x} \geq \boldsymbol{0}^n \}$ and $V = \{\boldsymbol{v}^1, \dots, \boldsymbol{v}^k\}$ be the set of extreme points of P. If and only if P is nonempty, V is nonempty (and finite). Further, let $C = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}^m;\ \boldsymbol{x} \geq \boldsymbol{0}^n \,\}$ and $D = \{\boldsymbol{d}^1, \dots, \boldsymbol{d}^r\}$ be the set of extreme directions of C. If and only if P is unbounded D is nonempty (and finite). Every $\boldsymbol{x} \in P$ can be represented as the sum of a convex combination of the points in V and a non-negative linear combination of the points in D, that is,*

$$\boldsymbol{x} = \sum_{i=1}^{k} \alpha_i \boldsymbol{v}^i + \sum_{j=1}^{r} \beta_j \boldsymbol{d}^j,$$

for some $\alpha_1, \dots, \alpha_k \geq 0$ such that $\sum_{i=1}^{k} \alpha_i = 1$, and $\beta_1, \dots, \beta_r \geq 0$. ∎

We have arrived at the important result that if there exists an optimal solution to a linear program in the standard form then there exists an optimal solution among the basic feasible solutions.

Theorem 8.10 (existence and properties of optimal solutions) *Let the sets P, V and D be defined as in Theorem 8.9 and consider the linear program*

$$\begin{array}{ll} \text{minimize} & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} & \boldsymbol{x} \in P. \end{array}$$

This problem has a finite optimal solution if and only if P is nonempty and z is lower bounded on P, that is, if P is nonempty and $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{d}^j \geq 0$ for all $\boldsymbol{d}^j \in D$.

If the problem has a finite optimal solution, then there exists an optimal solution among the extreme points.

Proof. Let $\boldsymbol{x} \in P$. Then it follows from Theorem 8.9 that

$$\boldsymbol{x} = \sum_{i=1}^{k} \alpha_i \boldsymbol{v}^i + \sum_{j=1}^{r} \beta_j \boldsymbol{d}^j, \tag{8.6}$$

for some $\alpha_1, \dots, \alpha_k \geq 0$ such that $\sum_{i=1}^{k} \alpha_i = 1$, and $\beta_1, \dots, \beta_r \geq 0$. Hence

$$\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} = \sum_{i=1}^{k} \alpha_i \boldsymbol{c}^{\mathrm{T}}\boldsymbol{v}^i + \sum_{j=1}^{r} \beta_j \boldsymbol{c}^{\mathrm{T}}\boldsymbol{d}^j. \tag{8.7}$$

Now, as $\boldsymbol{x}$ varies over P, the value of z clearly corresponds to variations of the weights α_i and β_j. The first term in the right-hand side of (8.7) is bounded as $\sum_{i=1}^{k} \alpha_i = 1$. The second term is lower bounded as $\boldsymbol{x}$ varies over P if and only if $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{d}^j \geq 0$ holds for all $\boldsymbol{d}^j \in D$, since otherwise we could let $\beta_j \to +\infty$ for an index j with $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{d}^j < 0$, and get that $z \to -\infty$. If $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{d}^j \geq 0$ for all $\boldsymbol{d}^j \in D$, then it is clearly optimal to choose $\beta_j = 0$ for $j = 1, \dots, r$. It remains to search for the optimal solution in the convex hull of V.

Now assume that $\boldsymbol{x} \in P$ is an optimal solution and let $\boldsymbol{x}$ be represented as in (8.6). From the above we have that we can choose $\beta_1 = \cdots = \beta_r = 0$, so we can assume that

$$\boldsymbol{x} = \sum_{i=1}^{k} \alpha_i \boldsymbol{v}^i.$$

Further, let

$$a \in \arg \underset{i \in \{1,\dots,k\}}{\text{minimum}}\, \boldsymbol{c}^{\mathrm{T}}\boldsymbol{v}^i.$$

Then,

$$\boldsymbol{c}^{\mathrm{T}}\boldsymbol{v}^a = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{v}^a \sum_{i=1}^{k} \alpha_i = \sum_{i=1}^{k} \alpha_i \boldsymbol{c}^{\mathrm{T}}\boldsymbol{v}^a \le \sum_{i=1}^{k} \alpha_i \boldsymbol{c}^{\mathrm{T}}\boldsymbol{v}^i = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x},$$

that is, the extreme point $\boldsymbol{v}^a$ is a global minimum. ■

Remark 8.11 The bounded case of Theorem 8.10 was already proved in Section 3.2. ■

8.2.3 Adjacent extreme points

Consider the polytope in Figure 8.2. Clearly, every point on the line

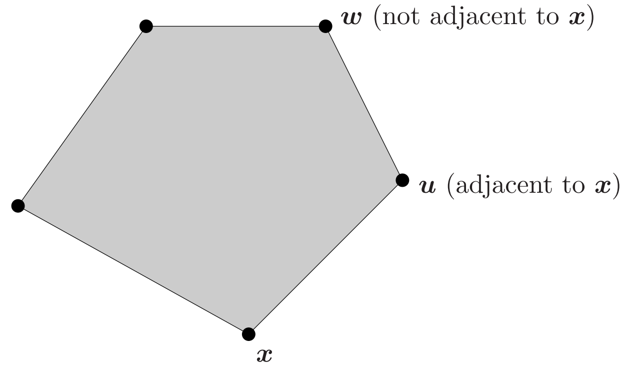


Figure 8.2: Illustration of adjacent extreme points.

segment joining the extreme points $\boldsymbol{x}$ and $\boldsymbol{u}$ cannot be written as a convex combination of any pair of points that are not on this line segment. However, this is not true for the points on the line segment between the extreme points $\boldsymbol{x}$ and $\boldsymbol{w}$. The extreme points $\boldsymbol{x}$ and $\boldsymbol{u}$ are said to be *adjacent* (while $\boldsymbol{x}$ and $\boldsymbol{w}$ are not adjacent).

Definition 8.12 (adjacent extreme points) *Two extreme points $\boldsymbol{x}$ and $\boldsymbol{u}$ of a polyhedron P are* adjacent *if each point $\boldsymbol{y}$ on the line segment between $\boldsymbol{x}$ and $\boldsymbol{u}$ has the property that if*

$$\boldsymbol{y} = \lambda\boldsymbol{v} + (1-\lambda)\boldsymbol{w},$$

where $\lambda \in (0,1)$ and $\boldsymbol{v}, \boldsymbol{w} \in P$, then both $\boldsymbol{v}$ and $\boldsymbol{w}$ must be on the line segment between $\boldsymbol{x}$ and $\boldsymbol{u}$. ■

Now, consider the polyhedron of standard form,

$$P = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}; \quad \boldsymbol{x} \geq \boldsymbol{0}^n \,\}.$$

Let $\boldsymbol{u} \in P$ be a basic feasible solution (and hence an extreme point of P) corresponding to the partition $\boldsymbol{A} = (\boldsymbol{B}^1, \boldsymbol{N}^1)$, where $\operatorname{rank} \boldsymbol{B}^1 = m$, that is,

$$\boldsymbol{u} = \begin{pmatrix} \boldsymbol{u}_{B^1} \\ \boldsymbol{u}_{N^1} \end{pmatrix} = \begin{pmatrix} (\boldsymbol{B}^1)^{-1}\boldsymbol{b} \\ \boldsymbol{0}^{n-m} \end{pmatrix}.$$

Further, let $\boldsymbol{B}^1 = (\boldsymbol{b}^1, \dots, \boldsymbol{b}^m)$ and $\boldsymbol{N}^1 = (\boldsymbol{n}^1, \dots, \boldsymbol{n}^{n-m})$ (that is, $\boldsymbol{b}^i \in \mathbb{R}^m$, $i = 1, \dots, m$, and $\boldsymbol{n}^j \in \mathbb{R}^n$, $j = 1, \dots, n-m$, are columns of $\boldsymbol{A}$). Now construct a new partition $(\boldsymbol{B}^2, \boldsymbol{N}^2)$ by replacing one column of $\boldsymbol{B}^1$, say $\boldsymbol{b}^1$, with one column of $\boldsymbol{N}^1$, say $\boldsymbol{n}^1$, that is,

$$\begin{aligned} \boldsymbol{B}^2 &= (\boldsymbol{n}^1, \boldsymbol{b}^2, \dots, \boldsymbol{b}^m), \\ \boldsymbol{N}^2 &= (\boldsymbol{b}^1, \boldsymbol{n}^2, \dots, \boldsymbol{n}^{n-m}). \end{aligned}$$

Assume that the partition $(\boldsymbol{B}^2, \boldsymbol{N}^2)$ corresponds to a basic feasible solution $\boldsymbol{v}$ (i.e., $\boldsymbol{v}$ is an extreme point), and that the two extreme points $\boldsymbol{u}$ and $\boldsymbol{v}$ corresponding to $(\boldsymbol{B}^1, \boldsymbol{N}^1)$ and $(\boldsymbol{B}^2, \boldsymbol{N}^2)$, respectively, are not equal. Then we have the following elegant result. (Since the Simplex method at each iteration performs exactly the above replacement action the theorem actually shows that the Simplex method at each non-degenerate iteration moves from one extreme point to an adjacent.)

Proposition 8.13 (algebraic characterization of adjacency) *Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be the extreme points that correspond to the partitions $(\boldsymbol{B}^1, \boldsymbol{N}^1)$ and $(\boldsymbol{B}^2, \boldsymbol{N}^2)$ described above. Then $\boldsymbol{u}$ and $\boldsymbol{v}$ are adjacent.*

Proof. If the variables of $\boldsymbol{v}$ are ordered in the same way as the variables of $\boldsymbol{u}$, then the vectors must be of the form

$$\begin{aligned} \boldsymbol{u} &= (u_1, \dots, u_m, 0, 0, \dots, 0)^{\mathrm{T}}, \\ \boldsymbol{v} &= (0, v_2, \dots, v_{m+1}, 0, \dots, 0)^{\mathrm{T}}. \end{aligned}$$

Now take a point $\boldsymbol{x}$ on the line segment between $\boldsymbol{u}$ and $\boldsymbol{v}$, that is,

$$\boldsymbol{x} = \lambda\boldsymbol{u} + (1-\lambda)\boldsymbol{v}$$

for some $\lambda \in (0, 1)$. In order to prove the theorem we must show that if $\boldsymbol{x}$ can be written as a convex combination of two feasible points, then these points must be on the line segment between $\boldsymbol{u}$ and $\boldsymbol{v}$. So assume that

$$\boldsymbol{x} = \alpha \boldsymbol{y}^1 + (1-\alpha)\boldsymbol{y}^2$$

for some feasible points $\boldsymbol{y}^1$ and $\boldsymbol{y}^2$, and $\alpha \in (0,1)$. Then it follows that $\boldsymbol{y}^1$ and $\boldsymbol{y}^2$ must be solutions to the system

$$\begin{aligned}
&y_1 \boldsymbol{b}^1 + \cdots + y_m \boldsymbol{b}^m + y_{m+1}\boldsymbol{n}^1 = \boldsymbol{b}, \\
&y_{m+2} = \ldots = y_n = 0, \\
&\boldsymbol{y} \geq \boldsymbol{0}^n,
\end{aligned}$$

or equivalently [by multiplying both sides of the first row by $(\boldsymbol{B}^1)^{-1}$],

$$\begin{aligned}
&\boldsymbol{y} = \begin{pmatrix} (\boldsymbol{B}^1)^{-1}\boldsymbol{b} \\ \boldsymbol{0}^{n-m} \end{pmatrix} + \begin{pmatrix} -y_{m+1}(\boldsymbol{B}^1)^{-1}\boldsymbol{n}^1 \\ y_{m+1} \\ \boldsymbol{0}^{n-m-1} \end{pmatrix}, \\
&\boldsymbol{y} \geq \boldsymbol{0}^n.
\end{aligned}$$

But this is in fact the line segment between $\boldsymbol{u}$ and $\boldsymbol{v}$ (if $y_{m+1} = 0$ then $\boldsymbol{y} = \boldsymbol{u}$ and if $y_{m+1} = v_{m+1}$ then $\boldsymbol{y} = \boldsymbol{v}$). In other words, $\boldsymbol{y}^1$ and $\boldsymbol{y}^2$ are on the line segment between $\boldsymbol{u}$ and $\boldsymbol{v}$, and we are done. ■

Remark 8.14 Actually the converse of Proposition 8.13 also holds. Namely, if two extreme points $\boldsymbol{u}$ and $\boldsymbol{v}$ are adjacent, then there exists a partition $(\boldsymbol{B}^1, \boldsymbol{N}^1)$ corresponding to $\boldsymbol{u}$ and a partition $(\boldsymbol{B}^2, \boldsymbol{N}^2)$ corresponding to $\boldsymbol{v}$ such that the columns of $\boldsymbol{B}^1$ and $\boldsymbol{B}^2$ are the same except for one. The proof is similar to that of Proposition 8.13. ■

8.3 Notes and further reading

The material in this chapter can be found in most books on linear programming, such as [Dan63, Chv83, Mur83, Sch86, Pad99, Van01, DaT97, DaT03].

8.4 Exercises

Exercise 8.1 (LP modelling) Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. Formulate the following problems as linear programming problems.

(a) $\text{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} \sum_{i=1}^m |(\boldsymbol{Ax} - \boldsymbol{b})_i|$ subject to $\max_{i=1,\ldots,n} |x_i| \leq 1$.

(b) $\text{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} \sum_{i=1}^m |(\boldsymbol{Ax} - \boldsymbol{b})_i| + \max_{i=1,\ldots,n} |x_i|$. ■

Exercise 8.2 (LP modelling) Consider the sets $V = \{\boldsymbol{v}^1, \dots, \boldsymbol{v}^k\} \subset \mathbb{R}^n$ and $W = \{\boldsymbol{w}^1, \dots, \boldsymbol{w}^l\} \subset \mathbb{R}^n$. Formulate the following problem as linear programming problems.

(a) Construct, if possible, a hyperplane that separates the sets V and W, that is, find $\boldsymbol{a} \in \mathbb{R}^n$, with $\boldsymbol{a} \neq \boldsymbol{0}^n$, and $b \in \mathbb{R}$ such that

$$\begin{aligned} \boldsymbol{a}^{\mathrm{T}}\boldsymbol{v} &\leq b, \quad \text{for all } \boldsymbol{v} \in V, \\ \boldsymbol{a}^{\mathrm{T}}\boldsymbol{w} &\geq b, \quad \text{for all } \boldsymbol{w} \in W. \end{aligned}$$

(b) Construct, if possible, a sphere that separates the sets V and W, that is, find a center $\boldsymbol{x}^c \in \mathbb{R}^n$ and a radius $R \geq 0$ such that

$$\begin{aligned} \|\boldsymbol{v} - \boldsymbol{x}^c\|_2 &\leq R, \quad \text{for all } \boldsymbol{v} \in V, \\ \|\boldsymbol{w} - \boldsymbol{x}^c\|_2 &\geq R, \quad \text{for all } \boldsymbol{w} \in W. \end{aligned}$$

■

Exercise 8.3 (linear-fractional programming) Consider the linear-fractional program

$$\begin{aligned} \text{minimize} \quad & f(\boldsymbol{x}) = (\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + \alpha)/(\boldsymbol{d}^{\mathrm{T}}\boldsymbol{x} + \beta) \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \end{aligned} \tag{8.8}$$

where $\boldsymbol{c}, \boldsymbol{d} \in \mathbb{R}^n$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, and $\boldsymbol{b} \in \mathbb{R}^m$. Further, assume that the polyhedron $P = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}$ is bounded and that $\boldsymbol{d}^{\mathrm{T}}\boldsymbol{x} + \beta > 0$ for all $\boldsymbol{x} \in P$. Show that (8.8) can be solved by solving the linear program

$$\begin{aligned} \text{minimize} \quad & g(\boldsymbol{y}, z) = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{y} + \alpha z \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{y} - z\boldsymbol{b} \leq \boldsymbol{0}^m, \\ & \boldsymbol{d}^{\mathrm{T}}\boldsymbol{y} + \beta z = 1, \\ & z \geq 0. \end{aligned} \tag{8.9}$$

[*Hint:* Suppose that $\boldsymbol{y}^*$ together with z^* are a solution to (8.9), and show that $z^* > 0$ and that $\boldsymbol{y}^*/z^*$ is a solution to (8.8).] ■

Exercise 8.4 (standard form) Transform the linear program

$$\begin{aligned} \text{minimize} \quad z = \quad & x_1 - 5x_2 - 7x_3 \\ \text{subject to} \quad & 5x_1 - 2x_2 + 6x_3 \geq 5, && (1) \\ & 3x_1 + 4x_2 - 9x_3 = 3, && (2) \\ & 7x_1 + 3x_2 + 5x_3 \leq 9, && (3) \\ & x_1 \geq -2, \end{aligned}$$

into standard form. ■

Exercise 8.5 (standard form) Consider the linear program

$$\begin{array}{llrrrrl} \text{minimize} & z = & 5x_1 & +3x_2 & -7x_3 & & \\ \text{subject to} & & 2x_1 & +4x_2 & +6x_3 & & = 11, \\ & & 3x_1 & -5x_2 & +3x_3 & +x_4 & = 11, \\ & & x_1, & x_2, & & x_4 & \geq 0. \end{array}$$

(a) Show how to transform this problem into standard form by eliminating the unrestricted variable x_3.

(b) Why cannot this technique be used to eliminate variables with non-negativity restrictions? ■

Exercise 8.6 (basic feasible solutions) Suppose that a linear program includes a free variable x_j. When transforming this problem into standard form, x_j is replaced by

$$\begin{aligned} x_j &= x_j^+ - x_j^-, \\ x_j^+, x_j^- &\geq 0. \end{aligned}$$

Show that no basic feasible solution can include both x_j^+ and x_j^- as non-zero basic variables. ■

Exercise 8.7 (equivalent systems) Consider the system of equations

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \qquad i = 1, \ldots, m. \tag{8.10}$$

Show that this system is equivalent to the system

$$\sum_{j=1}^{n} a_{ij} x_j \leq b_i, \quad i = 1, \ldots, m, \tag{8.11a}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} x_j \geq \sum_{i=1}^{m} b_i. \tag{8.11b}$$

■

The simplex method

IX

This chapter presents the Simplex method for solving linear programming problems. In Section 9.1 the basic algorithm is presented. First we assume that a basic feasible solution is known at the start of the algorithm, and then we describe what to do when a BFS is not known from the beginning. In Section 9.2 we discuss termination characteristics of the algorithm. It turns out that if all the BFSs of the problem are nondegenerate, then the basic algorithm terminates. However, if there exist degenerate BFSs there is a possibility that the basic algorithm cycles between degenerate BFSs and hence never terminates. Fortunately there is a simple rule, called Bland's rule, that eliminates cycling. We close the chapter by discussing the computational complexity of the Simplex algorithm. In the worst case, the algorithm visits all the extreme points of the problem, and since the number of extreme points may be exponential in the dimension of the problem, the Simplex algorithm does not belong to the desirable polynomial complexity class. The Simplex algorithm is therefore not theoretically satisfactory, but in practice it works very well and thus it frequently appears in commercial linear programming codes.

9.1 The algorithm

Assume that we have a linear program in standard form:

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n, \end{aligned}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $n > m$ and the rank of $\boldsymbol{A}$ is full, and $\boldsymbol{b} \in \mathbb{R}^m$ is such that $\boldsymbol{b} \geq \boldsymbol{0}^m$, and $\boldsymbol{c} \in \mathbb{R}^n$. (This is not a restriction, as was shown in

Section 8.2.1.) At each iteration the Simplex algorithm starts at a basic feasible solution (BFS) and moves to an adjacent BFS such that the objective function value decreases. It terminates with an optimal BFS (if there exists a finite optimal solution), or a *direction of unboundedness*, that is, a point in $C := \{\, \boldsymbol{p} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{p} = \boldsymbol{0}^m; \quad \boldsymbol{p} \geq \boldsymbol{0}^n \,\}$ along which the objective function diverges to $-\infty$. Observe that if $\boldsymbol{p} \in C$ is a direction of unboundedness and $\tilde{\boldsymbol{x}}$ is a feasible solution, then every solution $\boldsymbol{y}(\alpha)$ of the form

$$\boldsymbol{y}(\alpha) = \tilde{\boldsymbol{x}} + \alpha \boldsymbol{p}, \qquad \alpha \geq 0,$$

is feasible. Hence if $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{p} < 0$ then $z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{y}(\alpha) \to -\infty$ as $\alpha \to \infty$.

9.1.1 A BFS is known

Assume that a basic feasible solution $\boldsymbol{x}^{\mathrm{T}} = (\boldsymbol{x}_B^{\mathrm{T}}, \boldsymbol{x}_N^{\mathrm{T}})$ corresponding to the partition $\boldsymbol{A} = (\boldsymbol{B}, \boldsymbol{N})$ is known. Then we have that

$$\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{B}, \boldsymbol{N}) \begin{pmatrix} \boldsymbol{x}_B \\ \boldsymbol{x}_N \end{pmatrix} = \boldsymbol{B}\boldsymbol{x}_B + \boldsymbol{N}\boldsymbol{x}_N = \boldsymbol{b},$$

or, equivalently,

$$\boldsymbol{x}_B = \boldsymbol{B}^{-1}\boldsymbol{b} - \boldsymbol{B}^{-1}\boldsymbol{N}\boldsymbol{x}_N. \tag{9.1}$$

Further, rearrange the components of $\boldsymbol{c}$ such that $\boldsymbol{c}^{\mathrm{T}} = (\boldsymbol{c}_B^{\mathrm{T}}, \boldsymbol{c}_N^{\mathrm{T}})$ has the same ordering as $\boldsymbol{x}^{\mathrm{T}} = (\boldsymbol{x}_B^{\mathrm{T}}, \boldsymbol{x}_N^{\mathrm{T}})$. Then from (9.1) follows that

$$\begin{aligned} \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} &= \boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{x}_B + \boldsymbol{c}_N^{\mathrm{T}}\boldsymbol{x}_N \\ &= \boldsymbol{c}_B^{\mathrm{T}}(\boldsymbol{B}^{-1}\boldsymbol{b} - \boldsymbol{B}^{-1}\boldsymbol{N}\boldsymbol{x}_N) + \boldsymbol{c}_N^{\mathrm{T}}\boldsymbol{x}_N \\ &= \boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{b} + (\boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{N})\boldsymbol{x}_N. \end{aligned} \tag{9.2}$$

The principle of the Simplex algorithm is now easy to describe. Currently we are located at the BFS given by

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_B \\ \boldsymbol{x}_N \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}^{-1}\boldsymbol{b} \\ \boldsymbol{0}^{n-m} \end{pmatrix},$$

which is an extreme point according to Theorem 8.7. Proposition 8.13 gives that if we construct a new partition by replacing one column of $\boldsymbol{B}$ by one column of $\boldsymbol{N}$ such that the new partition corresponds to a basic feasible solution, $\tilde{\boldsymbol{x}}$, not equal to $\boldsymbol{x}$, then $\tilde{\boldsymbol{x}}$ is adjacent to $\boldsymbol{x}$. The principle of the Simplex algorithm is to move to an adjacent extreme point such that the objective function value decreases. From (9.2) follows that if we increase the j^{th} component of the non-basic vector $\boldsymbol{x}_N$ from 0 to 1, then the change in the objective function value becomes

$$(\tilde{\boldsymbol{c}}_N)_j := (\boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{N})_j,$$

that is, the change in the objective value resulting from a unit increase of the non-basic variable x_j from zero is given by the j^{th} component of the vector $\boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{N}$.

We call $(\tilde{\boldsymbol{c}}_N)_j$ the *reduced cost* of the non-basic variable $(\boldsymbol{x}_N)_j$ for $j = 1, \dots, n-m$. Actually, we can define the reduced cost, $\tilde{\boldsymbol{c}}^{\mathrm{T}} = (\tilde{\boldsymbol{c}}_B^{\mathrm{T}}, \tilde{\boldsymbol{c}}_N^{\mathrm{T}})$, of all the variables at the given BFS by

$$\begin{aligned}\tilde{\boldsymbol{c}}^{\mathrm{T}} &:= \boldsymbol{c}^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{A} = (\boldsymbol{c}_B^{\mathrm{T}}, \boldsymbol{c}_N^{\mathrm{T}}) - \boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{B}^{-1}(\boldsymbol{B}, \boldsymbol{N}) \\ &= ((\boldsymbol{0}^m)^{\mathrm{T}}, \boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{N});\end{aligned}$$

in particular, we see that the reduced costs of the basic variables are $\tilde{\boldsymbol{c}}_B = \boldsymbol{0}^m$.

Now, if $(\tilde{\boldsymbol{c}}_N)_j \geq 0$ for all $j = 1, \dots, n-m$, then there exists no adjacent extreme point such that the objective function value decreases and we stop; $\boldsymbol{x}$ is then an optimal solution.

Proposition 9.1 (optimality in the Simplex method) *Let $\boldsymbol{x}^*$ be the basic feasible solution that corresponds to the partition $\boldsymbol{A} = (\boldsymbol{B}, \boldsymbol{N})$. If $(\tilde{\boldsymbol{c}}_N)_j \geq 0$ for all $j = 1, \dots, n-m$, then $\boldsymbol{x}^*$ is an optimal solution.*

Proof. Since $\boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{b}$ is constant, it follows from (9.2) that the original linear program is equivalent to

$$\begin{aligned} \text{minimize} \quad z = & \qquad\qquad \tilde{\boldsymbol{c}}_N^{\mathrm{T}}\boldsymbol{x}_N \\ \text{subject to} \quad & \boldsymbol{x}_B + \boldsymbol{B}^{-1}\boldsymbol{N}\boldsymbol{x}_N = \boldsymbol{B}^{-1}\boldsymbol{b}, \\ & \boldsymbol{x}_B \qquad\qquad\quad \geq \boldsymbol{0}^m, \\ & \qquad\qquad\quad \boldsymbol{x}_N \geq \boldsymbol{0}^{n-m}, \end{aligned}$$

or equivalently [by reducing the $\boldsymbol{x}_B$ variables through (9.1)],

$$\begin{aligned} \text{minimize} \quad z = & \quad \tilde{\boldsymbol{c}}_N^{\mathrm{T}}\boldsymbol{x}_N \\ \text{subject to} \quad & \boldsymbol{B}^{-1}\boldsymbol{N}\boldsymbol{x}_N \leq \boldsymbol{B}^{-1}\boldsymbol{b}, \\ & \qquad\quad \boldsymbol{x}_N \geq \boldsymbol{0}^{n-m}. \end{aligned} \tag{9.3}$$

Since $\boldsymbol{x}^*$ is a basic feasible solution it follows that $\boldsymbol{x}_N^* = \boldsymbol{0}^{n-m}$ is a feasible solution to (9.3). But $\tilde{\boldsymbol{c}}_N \geq \boldsymbol{0}^{n-m}$ so $\boldsymbol{x}_N^* = \boldsymbol{0}^{n-m}$ is in fact an optimal solution to (9.3). (Why?) Hence

$$\boldsymbol{x}^* = \begin{pmatrix} \boldsymbol{B}^{-1}\boldsymbol{b} \\ \boldsymbol{0}^{n-m} \end{pmatrix}$$

is an optimal solution to the original problem. ■

Remark 9.2 (optimality condition) Proposition 9.1 states that if $(\tilde{\boldsymbol{c}}_N)_j \geq 0$ for all $j = 1, \dots, n-m$, then $\boldsymbol{x}^*$ is an optimal extreme point. But is it true that if $\boldsymbol{x}^*$ is an optimal extreme point, then $(\tilde{\boldsymbol{c}}_N)_j \geq 0$ for all $j = 1, \dots, n-m$? The answer to this question is *no*. Namely, if the optimal basic feasible solution $\boldsymbol{x}^*$ is degenerate, then there may exist basis representations of $\boldsymbol{x}^*$ such that $(\tilde{\boldsymbol{c}}_N)_j < 0$ for some j. However, it holds that if $\boldsymbol{x}^*$ is an optimal extreme point, then there *exists* at least one basis representation of it such that $(\tilde{\boldsymbol{c}}_N)_j \geq 0$ for all $j = 1, \dots, n-m$. That is, Proposition 9.1 can be strengthened to state that $\boldsymbol{x}^*$ *is an optimal extreme point if and only if there exists a basis representation of it such that* $\tilde{\boldsymbol{c}}_N \geq \boldsymbol{0}^{n-m}$. ■

If some of the reduced costs are strictly negative, we choose the non-basic variable with the *least* reduced cost to enter the basis. We must also choose one variable from $\boldsymbol{x}_B$ to leave the basis. Suppose that the variable $(\boldsymbol{x}_N)_j$ has been chosen to enter the basis. Then, according to (9.1), when the value of $(\boldsymbol{x}_N)_j$ is increased from zero we will move along the half-line

$$\boldsymbol{l}(\mu) := \begin{pmatrix} \boldsymbol{B}^{-1}\boldsymbol{b} \\ \boldsymbol{0}^{n-m} \end{pmatrix} + \mu \begin{pmatrix} -\boldsymbol{B}^{-1}\boldsymbol{N}_j \\ \boldsymbol{e}_j \end{pmatrix}, \qquad \mu \geq 0,$$

where $\boldsymbol{e}_j$ is the jth unit vector. In order to maintain feasibility we must have that $\boldsymbol{l}(\mu) \geq \boldsymbol{0}^n$. If $\boldsymbol{l}(\mu) \geq \boldsymbol{0}^n$ for all $\mu \geq 0$, then $z \to -\infty$ as $\mu \to \infty$, that is,

$$\boldsymbol{p} = \begin{pmatrix} -\boldsymbol{B}^{-1}\boldsymbol{N}_j \\ \boldsymbol{e}_j \end{pmatrix}$$

is a direction of unboundedness and $z \to -\infty$ along the half-line $\boldsymbol{l}(\mu)$, $\mu \geq 0$. Observe that this occurs if and only if

$$\boldsymbol{B}^{-1}\boldsymbol{N}_j \leq \boldsymbol{0}^m.$$

Otherwise, the maximal value of μ in order to maintain feasibility is given by

$$\mu^* = \underset{i \in \{\, i \,|\, (\boldsymbol{B}^{-1}\boldsymbol{N}_j)_i > 0 \,\}}{\text{minimum}} \frac{(\boldsymbol{B}^{-1}\boldsymbol{b})_i}{(\boldsymbol{B}^{-1}\boldsymbol{N}_j)_i}.$$

If $\mu^* > 0$ it follows that $\boldsymbol{l}(\mu^*)$ is an extreme point adjacent to $\boldsymbol{x}$. Actually we move to $\boldsymbol{l}(\mu^*)$ by choosing $(\boldsymbol{x}_B)_i$, where

$$i \in \arg \underset{i \in \{\, i \,|\, (\boldsymbol{B}^{-1}\boldsymbol{N}_j)_i > 0 \,\}}{\text{minimum}} \frac{(\boldsymbol{B}^{-1}\boldsymbol{b})_i}{(\boldsymbol{B}^{-1}\boldsymbol{N}_j)_i},$$

to leave the basis.

We are now ready to state the Simplex algorithm.

The Simplex Algorithm:

Step 0 (initialization: BFS). Let $\boldsymbol{x}^{\mathrm{T}} = (\boldsymbol{x}_B^{\mathrm{T}}, \boldsymbol{x}_N^{\mathrm{T}})$ be a BFS corresponding to the partition $\boldsymbol{A} = (\boldsymbol{B}, \boldsymbol{N})$.

Step 1 (descent direction generation or termination: entering variable, pricing). Calculate the reduced costs of the non-basic variables:

$$(\tilde{\boldsymbol{c}}_N)_j = (\boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{N})_j, \quad j = 1, \ldots, n - m.$$

If $(\tilde{\boldsymbol{c}}_N)_j \geq 0$ for all $j = 1, \ldots, n - m$ then stop; $\boldsymbol{x}$ is then optimal. Otherwise choose $(\boldsymbol{x}_N)_j$, where

$$j \in \arg \operatorname*{minimum}_{j \in \{1,\ldots,n-m\}} (\tilde{\boldsymbol{c}}_N)_j,$$

to enter the basis.

Step 2 (line search or termination: leaving variable). If

$$\boldsymbol{B}^{-1} \boldsymbol{N}_j \leq \boldsymbol{0}^m,$$

then the problem is unbounded, stop; $\boldsymbol{p} := ((-\boldsymbol{B}^{-1}\boldsymbol{N}_j)^{\mathrm{T}}, \boldsymbol{e}_j^{\mathrm{T}})^{\mathrm{T}}$ is then a direction of unboundedness. Otherwise choose $(\boldsymbol{x}_B)_i$, where

$$i \in \arg \operatorname*{minimum}_{i \in \{\, i \mid (\boldsymbol{B}^{-1}\boldsymbol{N}_j)_i > 0 \,\}} \frac{(\boldsymbol{B}^{-1}\boldsymbol{b})_i}{(\boldsymbol{B}^{-1}\boldsymbol{N}_j)_i},$$

to leave the basis.

Step 3 (update: change basis). Construct a new partition by swapping $(\boldsymbol{x}_B)_i$ with $(\boldsymbol{x}_N)_j$. Go to Step 1.

Remark 9.3 (the Simplex algorithm as a feasible descent method) In the above description, we have chosen to use terms similar to those that will be used for several descent methods in nonlinear optimization that are described in Parts V and VI; see, for example, the algorithm description in Section 11.1 for unconstrained nonlinear optimization problems. The Simplex method is a very special type of descent algorithm: in order to remain feasible we generate feasible descent directions $\boldsymbol{p}$ (Step 1) that follow the boundary of the polyhedron; because of the fact that the objective function is linear, a line search would yield an infinite step unless a new boundary makes such a step infeasible; this is the role of Step 2. Finally, termination at an optimal solution (Step 2) is based on a special property of linear programming which allows us to decide on global optimality based on only local information about the current BFS:s reduced costs. (Of course, the convexity of LP is a crucial property for this principle to be valid.) More on the characterization of this optimality criterion, and its relationships to the optimality principles in the Chapters 4 and 6 will be discussed in the next chapter. ∎

Remark 9.4 (calculating the reduced costs) When calculating the reduced costs of the non-basic variables at the pricing Step 1 of the Simplex algorithm, it is appropriate to first calculate

$$\boldsymbol{y}^{\mathrm{T}} := \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1},$$

through the system

$$\boldsymbol{B}^{\mathrm{T}} \boldsymbol{y} = \boldsymbol{c}_B,$$

and then calculate the reduced costs by

$$\tilde{\boldsymbol{c}}_N^{\mathrm{T}} = \boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{y}^{\mathrm{T}} \boldsymbol{N}.$$

By this procedure we avoid the matrix–matrix multiplication $\boldsymbol{B}^{-1}\boldsymbol{N}$. ∎

Remark 9.5 (alternative pricing rules) If n is very large, it can be costly to compute the reduced costs at the pricing Step 1 of the Simplex algorithm. A methodology which saves computations is *partial pricing*, in which only a subset of the elements $(\tilde{\boldsymbol{c}}_N)_j$ is calculated.

Another problem with the standard pricing rule is that the use of the criterion $\operatorname{minimize}_{j \in \{1,\dots,n-m\}} \{(\tilde{\boldsymbol{c}}_N)_j\}$ does not take into account the actual improvement that is made. In particular, a different scaling of the variables might mean that one unit change is a dramatic move in one variable, and a very small move in another. The *steepest-edge rule* eliminates this scaling problem somewhat: With $(\boldsymbol{x}_N)_j$ being the entering variable we have that

$$\begin{pmatrix} \boldsymbol{x}_B \\ \boldsymbol{x}_N \end{pmatrix}^{\text{new}} = \begin{pmatrix} \boldsymbol{x}_B \\ \boldsymbol{x}_N \end{pmatrix} + (\boldsymbol{x}_N)_j \boldsymbol{p}_j, \qquad \boldsymbol{p}_j = \begin{pmatrix} -\boldsymbol{B}^{-1}\boldsymbol{N}_j \\ \boldsymbol{e}_j \end{pmatrix}.$$

Choose j in

$$\arg \underset{j \in \{1,\dots,n-m\}}{\operatorname{minimum}} \frac{\boldsymbol{c}^{\mathrm{T}} \boldsymbol{p}_j}{\|\boldsymbol{p}_j\|},$$

that is, the usual pricing rule based on $\boldsymbol{c}^{\mathrm{T}} \boldsymbol{p}_j = \boldsymbol{c}_B^{\mathrm{T}}(-\boldsymbol{B}^{-1}\boldsymbol{N}_j) + (\boldsymbol{c}_N)_j = (\tilde{\boldsymbol{c}}_N)_j$ is replaced by a rule wherein the reduced costs are scaled by the length of the candidate search directions $\boldsymbol{p}_j$. (Other scaling factors can of course be used.) ∎

Remark 9.6 (initial basic feasible solution) Consider the linear program

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n, \end{aligned} \tag{9.4}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$ is such that $\boldsymbol{b} \geq \boldsymbol{0}^m$. By introducing slack variables $\boldsymbol{s}$ we get

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} + \boldsymbol{I}^m\boldsymbol{s} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n, \\ & \boldsymbol{s} \geq \boldsymbol{0}^m. \end{aligned} \tag{9.5}$$

Since $\boldsymbol{b} \geq \boldsymbol{0}^m$ it then follows that the partition $(\boldsymbol{I}^m, \boldsymbol{A})$ corresponds to a basic feasible solution to (9.5), that is, the slack variables $\boldsymbol{s}$ are the basic variables. (This corresponds to the origin in the problem (9.4), which is clearly feasible when $\boldsymbol{b} \geq \boldsymbol{0}^m$.)

Similarly, if we can identify an identity matrix among the columns of the constraint matrix, then (if the right-hand side is non-negative, which is the case if the problem is of standard form) we get a basic feasible solution just by taking the variables that correspond to these columns as basic variables. ∎

To illustrate the Simplex algorithm we give an example.

Example 9.7 (illustration of the Simplex method) Consider the linear program

$$\begin{array}{llrrrrl} \text{minimize} & z = & x_1 & -2x_2 & -4x_3 & +4x_4 & \\ \text{subject to} & & & -x_2 & +2x_3 & +x_4 & \leq 4, \\ & & -2x_1 & +x_2 & +x_3 & -4x_4 & \leq 5, \\ & & x_1 & -x_2 & & +2x_4 & \leq 3, \\ & & x_1, & x_2, & x_3, & x_4 & \geq 0. \end{array}$$

By introducing the slack variables x_5, x_6 and x_7 we get the problem to

$$\begin{array}{llrrrrrrrl} \text{minimize} & z = & x_1 & -2x_2 & -4x_3 & +4x_4 & & & & \\ \text{subject to} & & & -x_2 & +2x_3 & +x_4 & +x_5 & & & = 4, \\ & & -2x_1 & +x_2 & +x_3 & -4x_4 & & +x_6 & & = 5, \\ & & x_1 & -x_2 & & +2x_4 & & & +x_7 & = 3, \\ & & x_1, & x_2, & x_3, & x_4, & x_5, & x_6, & x_7 & \geq 0. \end{array}$$

According to Remark 9.6 we take $\boldsymbol{x}_B = (x_5, x_6, x_7)^{\mathrm{T}}$ and $\boldsymbol{x}_N = (x_1, x_2, x_3, x_4)^{\mathrm{T}}$ as the initial basic and non-basic vector, respectively. The reduced costs of the non-basic variables then become

$$\boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{N} = (1, -2, -4, 4),$$

and hence we choose x_3 as the entering variable. Further, we have that

$$\begin{aligned} \boldsymbol{B}^{-1}\boldsymbol{b} &= (4, 5, 3)^{\mathrm{T}}, \\ \boldsymbol{B}^{-1}\boldsymbol{N}_3 &= (2, 1, 0)^{\mathrm{T}}, \end{aligned}$$

which gives that

$$\arg \underset{i \in \{\, i \,|\, (\boldsymbol{B}^{-1}\boldsymbol{N}_3)_i > 0 \,\}}{\operatorname{minimum}} \frac{(\boldsymbol{B}^{-1}\boldsymbol{b})_i}{(\boldsymbol{B}^{-1}\boldsymbol{N}_3)_i} = \{1\},$$

so we choose x_5 to leave the basis. The new basic and non-basic vectors are $\boldsymbol{x}_B = (x_3, x_6, x_7)^{\mathrm{T}}$ and $\boldsymbol{x}_N = (x_1, x_2, x_5, x_4)^{\mathrm{T}}$, and the reduced costs of the non-basic variables become

$$\boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{N} = (1, -4, 2, 6),$$

so x_2 is the entering variable, and

$$\begin{aligned} \boldsymbol{B}^{-1}\boldsymbol{b} &= (2, 3, 3)^{\mathrm{T}}, \\ \boldsymbol{B}^{-1}\boldsymbol{N}_2 &= (-1/2, 3/2, -1)^{\mathrm{T}}, \end{aligned}$$

which gives that

$$\arg \underset{i \in \{\, i \,|\, (\boldsymbol{B}^{-1}\boldsymbol{N}_2)_i > 0 \,\}}{\operatorname{minimum}} \frac{(\boldsymbol{B}^{-1}\boldsymbol{b})_i}{(\boldsymbol{B}^{-1}\boldsymbol{N}_2)_i} = \{2\},$$

and hence x_6 is the leaving variable. The new basic and non-basic vectors become $\boldsymbol{x}_B = (x_3, x_2, x_7)^{\mathrm{T}}$ and $\boldsymbol{x}_N = (x_1, x_6, x_5, x_4)^{\mathrm{T}}$, and the reduced costs of the non-basic variables are

$$\boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{N} = (-13/3, 8/3, 2/3, -6),$$

so x_4 is the entering variable and

$$\begin{aligned} \boldsymbol{B}^{-1}\boldsymbol{b} &= (3, 2, 5)^{\mathrm{T}}, \\ \boldsymbol{B}^{-1}\boldsymbol{N}_4 &= (-1, -3, -1)^{\mathrm{T}}. \end{aligned}$$

But since $\boldsymbol{B}^{-1}\boldsymbol{N}_4 \leq \boldsymbol{0}^3$ it follows that the objective function diverges to $-\infty$ along the half-line given by

$$\boldsymbol{l}(\mu) = (x_1, x_2, x_3, x_4)^{\mathrm{T}} = (0, 2, 3, 0)^{\mathrm{T}} + \mu(0, 3, 1, 1)^{\mathrm{T}}, \qquad \mu \geq 0.$$

We conclude that the problem is unbounded. ∎

9.1.2 A BFS is not known: Phase I & II

Often a basic feasible solution is not known from the beginning. (In fact, only if the origin is feasible in (9.4) we know a BFS immediately.) However, an initial basic feasible solution can be found by solving a linear program that is a pure feasibility problem. We call this the *Phase I problem*.

Consider the following linear program in standard form:

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n. \end{aligned} \tag{9.6}$$

In order to find a basic feasible solution we introduce the *artificial variables* $\boldsymbol{a} \in \mathbb{R}^m$ and consider the *Phase I problem* to

$$\begin{aligned} \text{minimize} \quad & w = (\boldsymbol{1}^m)^{\mathrm{T}}\boldsymbol{a} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} + \boldsymbol{I}^m\boldsymbol{a} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n, \\ & \boldsymbol{a} \geq \boldsymbol{0}^m. \end{aligned} \tag{9.7}$$

In other words, we introduce an additional (artificial) variable a_i for every linear constraint $i = 1, \ldots, m$, and thus construct the unit matrix in $\mathbb{R}^{m \times m}$ sought.

We get a basic feasible solution to the Phase I problem (9.7) by taking the artificial variables $\boldsymbol{a}$ as the basic variables. (Remember that $\boldsymbol{b} \geq \boldsymbol{0}^m$; the simplicity of finding an initial BFS for the Phase I problem is in fact the reason why we require this to hold!) Then the Phase I problem (9.7) can be solved by the Simplex method stated in the previous section. Note that the Phase I problem is bounded from below (since $(\boldsymbol{1}^m)^{\mathrm{T}}\boldsymbol{a} \geq 0$) which means that an optimal solution to (9.7) always exists by Theorem 8.10.

Assume that the optimal objective function value is w^*. We observe that if and only if the part $\boldsymbol{x}^*$ of an optimal solution $((\boldsymbol{x}^*)^{\mathrm{T}}, (\boldsymbol{a}^*)^{\mathrm{T}})^{\mathrm{T}}$ to the problem (9.7) is a feasible solution to the original problem (9.6), then $((\boldsymbol{x}^*)^{\mathrm{T}}, (\boldsymbol{0}^m)^{\mathrm{T}})^{\mathrm{T}}$ is an optimal feasible solution to the Phase I problem and $w^* = 0$. Hence, if $w^* > 0$, then the original linear program is infeasible. We have the following cases:

1. If $w^* > 0$, then the original problem is infeasible.
2. If $w^* = 0$, then if the optimal basic feasible solution is $(\boldsymbol{x}^{\mathrm{T}}, \boldsymbol{a}^{\mathrm{T}})^{\mathrm{T}}$ we must have that $\boldsymbol{a} = \boldsymbol{0}^m$, and $\boldsymbol{x}$ corresponds to a basic feasible

solution to the original problem.[1]

Therefore, if there exists a feasible solution to the original problem (9.6), then a basic feasible solution is found by solving the Phase I problem (9.7). This basic feasible solution can then be used as the starting BFS in the solution of the original problem, which is called the *Phase II problem*, with the Simplex method.

Remark 9.8 (artificial variables) The purpose of introducing artificial variables is to get an identity matrix among the columns of the constraint matrix. If some of the columns of the constraint matrix of the original problem consists of only zeros except for one positive entry, then it is not necessary to introduce an artificial variable in the corresponding row. An example of a linear constraint for which an original variable naturally serves as a basic variable is a $\leq$-constraint with a positive right-hand side, in which case we can use the corresponding slack variable. ∎

Example 9.9 (Phase I & II) Consider the following linear program:

$$\begin{array}{lrrrrl}
\text{minimize} & z = 2x_1 & & & & \\
\text{subject to} & x_1 & & -x_3 & & = 3, \\
& x_1 & -x_2 & & -2x_4 & = 1, \\
& 2x_1 & & & +x_4 & \leq 7, \\
& x_1, & x_2, & x_3, & x_4 & \geq 0.
\end{array}$$

By introducing a slack variable x_5 we get the equivalent linear program in standard form:

$$\begin{array}{lrrrrrl}
\text{minimize} & z = 2x_1 & & & & & \\
\text{subject to} & x_1 & & -x_3 & & & = 3, \\
& x_1 & -x_2 & & -2x_4 & & = 1, \\
& 2x_1 & & & +x_4 & +x_5 & = 7, \\
& x_1, & x_2, & x_3, & x_4, & x_5 & \geq 0.
\end{array} \tag{9.8}$$

We cannot identify the identity matrix among the columns of the constraint matrix of the problem (9.8), but the third unit vector $\boldsymbol{e}_3$ is found in the column corresponding to the x_5-variable. Therefore, we leave the problem (9.8) for a while, and instead we introduce two artificial variables a_1 and a_2 and consider the Phase I problem to

[1] Notice that if the final BFS in the Phase I problem is degenerate then one or several artificial variables a_i may remain in the basis with value zero; in order to remove them from the basis a number of degenerate pivots may have to be performed; this is naturally always possible.

$$
\begin{array}{llrrrrrrrl}
\text{minimize} & w = & & & & & & a_1 & +a_2 & \\
\text{subject to} & & x_1 & & -x_3 & & & +a_1 & & = 3, \\
& & x_1 & -x_2 & & -2x_4 & & & +a_2 & = 1, \\
& & 2x_1 & & & +x_4 & +x_5 & & & = 7, \\
& & x_1, & x_2, & x_3, & x_4, & x_5, & a_1, & a_2 & \geq 0.
\end{array}
$$

Let $\boldsymbol{x}_B = (a_1, a_2, x_5)^{\mathrm{T}}$ and $\boldsymbol{x}_N = (x_1, x_2, x_3, x_4)^{\mathrm{T}}$ be the initial basic and non-basic vector, respectively. The reduced costs of the non-basic variables then become

$$\boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{N} = (-2, 1, 1, 2),$$

and hence we choose x_1 as the entering variable. Further, we have

$$
\begin{aligned}
\boldsymbol{B}^{-1}\boldsymbol{b} &= (3, 1, 7)^{\mathrm{T}}, \\
\boldsymbol{B}^{-1}\boldsymbol{N}_1 &= (1, 1, 2)^{\mathrm{T}},
\end{aligned}
$$

which gives that

$$\arg \underset{i \in \{\, i \,|\, (\boldsymbol{B}^{-1}\boldsymbol{N}_1)_i > 0 \,\}}{\operatorname{minimum}} \frac{(\boldsymbol{B}^{-1}\boldsymbol{b})_i}{(\boldsymbol{B}^{-1}\boldsymbol{N}_1)_i} = \{2\},$$

so we choose a_2 as the leaving variable. The new basic and non-basic vectors are $\boldsymbol{x}_B = (a_1, x_1, x_5)^{\mathrm{T}}$ and $\boldsymbol{x}_N = (a_2, x_2, x_3, x_4)^{\mathrm{T}}$, and the reduced costs of the non-basic variables become

$$\boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{N} = (2, -1, 1, -2),$$

so x_4 is the entering variable, and

$$
\begin{aligned}
\boldsymbol{B}^{-1}\boldsymbol{b} &= (2, 1, 5)^{\mathrm{T}}, \\
\boldsymbol{B}^{-1}\boldsymbol{N}_4 &= (2, -2, 5)^{\mathrm{T}},
\end{aligned}
$$

which gives that

$$\arg \underset{i \in \{\, i \,|\, (\boldsymbol{B}^{-1}\boldsymbol{N}_4)_i > 0 \,\}}{\operatorname{minimum}} \frac{(\boldsymbol{B}^{-1}\boldsymbol{b})_i}{(\boldsymbol{B}^{-1}\boldsymbol{N}_4)_i} = \{1, 3\},$$

and we choose a_1 to leave the basis. The new basic and non-basic vectors become $\boldsymbol{x}_B = (x_4, x_1, x_5)^{\mathrm{T}}$ and $\boldsymbol{x}_N = (a_2, x_2, x_3, a_1)^{\mathrm{T}}$, and the reduced costs of the non-basic variables are

$$\boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{N} = (1, 0, 0, 1),$$

so by choosing the basic variables as $\boldsymbol{x}_B = (x_4, x_1, x_5)^{\mathrm{T}}$ we get an optimal basic feasible solution of the Phase I problem, and $w^* = 0$. This means that by choosing the basic variables as $\boldsymbol{x}_B = (x_4, x_1, x_5)^{\mathrm{T}}$ we get a basic feasible solution of the Phase II problem (9.8).

We return to the problem (9.8). By letting $\boldsymbol{x}_B = (x_4, x_1, x_5)^{\mathrm{T}}$ and $\boldsymbol{x}_N = (x_2, x_3)^{\mathrm{T}}$ we get the reduced costs

$$\tilde{\boldsymbol{c}}_N^{\mathrm{T}} = \boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{N} = (0, 2),$$

which means that

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_B \\ \boldsymbol{x}_N \end{pmatrix} = \begin{pmatrix} x_4 \\ x_1 \\ x_5 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}^{-1}\boldsymbol{b} \\ \boldsymbol{0}^2 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

is an optimal basic feasible solution to the original problem. (Observe that most often the basic feasible solution found when solving the Phase I problem is not an optimal solution to the Phase II problem!) But since the reduced cost of x_2 is zero there is a possibility that there are alternative optimal solutions. Let x_2 enter the basic vector. We have that

$$\begin{aligned} \boldsymbol{B}^{-1}\boldsymbol{b} &= (1, 3, 0)^{\mathrm{T}}, \\ \boldsymbol{B}^{-1}\boldsymbol{N}_1 &= (0.5, 0, -0.5)^{\mathrm{T}}, \end{aligned}$$

which gives that

$$\arg \underset{i \in \{\, i \,|\, (\boldsymbol{B}^{-1}\boldsymbol{N}_1)_i > 0 \,\}}{\operatorname{minimum}} \frac{(\boldsymbol{B}^{-1}\boldsymbol{b})_i}{(\boldsymbol{B}^{-1}\boldsymbol{N}_1)_i} = \{1\},$$

so x_4 is the leaving variable. We get $\boldsymbol{x}_B = (x_2, x_1, x_5)^{\mathrm{T}}$ and $\boldsymbol{x}_N = (x_4, x_3)^{\mathrm{T}}$, and the reduced costs become

$$\tilde{\boldsymbol{c}}_N^{\mathrm{T}} = \boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{N} = (0, 2),$$

so

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_B \\ \boldsymbol{x}_N \end{pmatrix} = \begin{pmatrix} x_2 \\ x_1 \\ x_5 \\ x_4 \\ x_3 \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}^{-1}\boldsymbol{b} \\ \boldsymbol{0}^2 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

is an alternative optimal basic feasible solution. ∎

9.1.3 Alternative optimal solutions

As we saw in Example 9.9 there can be alternative optimal solutions to a linear program. However, this can only happen if some of the reduced costs of the non-basic variables of an optimal solution is zero.

Proposition 9.10 (unique optimal solutions in linear programming) *Consider the linear program in standard form*

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n. \end{aligned}$$

Let $\boldsymbol{x}^{\mathrm{T}} = (\boldsymbol{x}_B^{\mathrm{T}}, \boldsymbol{x}_N^{\mathrm{T}})$ *be an optimal basic feasible solution that corresponds to the partition* $\boldsymbol{A} = (\boldsymbol{B}, \boldsymbol{N})$. *If the reduced costs of the non-basic variables* $\boldsymbol{x}_N$ *are all strictly positive, then* $\boldsymbol{x}$ *is the unique optimal solution.*

Proof. As in the proof of Proposition 9.1 we have that the original linear program is equivalent to

$$\begin{aligned} \text{minimize} \quad & z = \tilde{\boldsymbol{c}}_N^{\mathrm{T}}\boldsymbol{x}_N \\ \text{subject to} \quad & \boldsymbol{x}_B + \boldsymbol{B}^{-1}\boldsymbol{N}\boldsymbol{x}_N = \boldsymbol{B}^{-1}\boldsymbol{b}, \\ & \boldsymbol{x}_B \geq \boldsymbol{0}^m, \\ & \boldsymbol{x}_N \geq \boldsymbol{0}^{n-m}. \end{aligned}$$

Now if the reduced costs of the non-basic variables are all strictly positive, that is, $\tilde{\boldsymbol{c}}_N > \boldsymbol{0}^{n-m}$, it follows that a solution for which $(\boldsymbol{x}_N)_j > 0$ for some $j = 1, \dots, n-m$ cannot be optimal. Hence

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_B \\ \boldsymbol{x}_N \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}^{-1}\boldsymbol{b} \\ \boldsymbol{0}^{n-m} \end{pmatrix}$$

is the unique optimal solution. ∎

9.2 Termination

So far we have not discussed whether the Simplex algorithm terminates in a finite number of iterations or not. Unfortunately, if there exist degenerate basic feasible solutions it can happen that the Simplex algorithm cycles between degenerate solutions and hence never terminates. However, if all of the basic feasible solutions are non-degenerate this kind of cycling never occurs.

Theorem 9.11 (finiteness of the Simplex algorithm) *If all of the basic feasible solutions are non-degenerate, then the Simplex algorithm terminates after a finite number of iterations.*

Proof. If a basic feasible solution is non-degenerate it follows that it has exactly m strictly positive components, and hence has a unique associated basis. In this case, in the minimum ratio test,

$$\mu^* = \operatorname*{minimum}_{i \in \{\, i \mid (\boldsymbol{B}^{-1}\boldsymbol{N}_j)_i > 0 \,\}} \frac{(\boldsymbol{B}^{-1}\boldsymbol{b})_i}{(\boldsymbol{B}^{-1}\boldsymbol{N}_j)_i},$$

we get that $\mu^* > 0$. Therefore, at each iteration the objective value strictly decreases, and hence a basic feasible solution that has appeared once can never reappear. Further, from Corollary 3.18 follows that the number of extreme points, hence the number of basic feasible solutions, is finite. We are done. ∎

Cycling resulting from degeneracy does not seem to occur often among the numerous degenerate linear programs encountered in practical applications. However, the fact that it *can* occur is not theoretically satisfactory. Therefore methods have been developed that avoid cycling. One of them is Bland's rule.

Theorem 9.12 (Bland's rule) *Fix an ordering of the variables. (This ordering can be arbitrary, but once it has been selected it cannot be changed.) If at each iteration step the entering and leaving variables are chosen as the first variables that are eligible*[2] *in the ordering, then the Simplex algorithm terminates after a finite number of iteration steps.* ∎

9.3 Computational complexity

The Simplex algorithm is very efficient in practice. Although the total number of basic feasible solutions can be as many as

$$\binom{n}{m} = \frac{n!}{(n-m)!\,m!}$$

[2]By eligible entering variables we mean the variables $(\boldsymbol{x}_N)_j$ for which $(\tilde{\boldsymbol{c}}_N)_j < 0$, and when we have chosen the entering variable j, the eligible leaving variables are the variables $(\boldsymbol{x}_B)_i$ such that

$$i \in \arg \operatorname*{minimum}_{i \in \{\, i \mid (\boldsymbol{B}^{-1}\boldsymbol{N}_j)_i > 0 \,\}} \frac{(\boldsymbol{B}^{-1}\boldsymbol{b})_i}{(\boldsymbol{B}^{-1}\boldsymbol{N}_j)_i}.$$

(the number of different ways m objects can be chosen from n objects), which is a number that grows exponentially, it is rare that more than $3m$ iterations are needed, and practice shows that the expected number is in the order of $3m/2$. Since each iteration costs no more than a polynomial ($O(m^3)$ for factorizations and $O(mn)$ for the pricing) the algorithm is polynomial *in practice*. Its worst-case behaviour is however very bad, in fact exponential.

The bad worst-care behaviour of the simplex method led to a huge amount of work being laid down to find polynomial algorithms for solving linear programs. Such a polynomial time competitor to the Simplex method nowadays is the class of interior point algorithms. Its main feature is that the optimal extreme points are not approached by following the edges, but by moving within the interior of the polyhedron. The famous Karmarkar algorithm is one, which however has been improved much in recent years. An analysis of interior point methods for linear programs is made in Chapter 13, as they are in fact to be seen as instances of the interior penalty algorithm in nonlinear programming.

9.4 Notes and further reading

The simplex method was developed by Danzig [Dan51]. The version of the simplex method we have presented is usually called the revised simplex method and was first described by Danzig [Dan53] and Orchard-Hays [Orc54]. The first book describing the simplex method was [Dan63].

In the (revised) simplex algorithm several computations are performed using $\boldsymbol{B}^{-1}$. The major drawback in this approach is that round-off errors accumulates as the algorithm moves from step to step. This drawback can be alleviated by using LU decomposition or Cholesky factorization. Most of the software packages for linear programming use LU decomposition. Early references on numerically stable forms of the simplex method are [BaG69, Bar71, GiM73, Sau72]. Books that discusses the subject are [Mur83, NaS96].

The first example of cycling was constructed by Hoffman [Hof53]. Different methods have been developed that avoid cycling, for example the perturbation method by Charnes [Cha52], the lexicographic method by Danzig, Orden and Wolfe [DOW55], and Bland's rule by Bland [Bla77]. In practice, however, cycling is rarely encountered. Instead, the problem is *stalling*, which means that the value of the objective function does not change (or changes very little) for a very large number of iterations[3] before it eventually starts to make substantial progress again. So in

[3] "Very large" normally refers to a number of iterations which is an exponential function of the number of variables of the LP problem.

practice, we are interested in methods that primarily prevent stalling, and only secondarily cycling (see, e.g., [GMSW89]).

In 1972, Klee and Minty [KlM72] showed that there exist problems of arbitrary size that cause the simplex method to examine every possible basis when the steepest-descent pricing rule is used, and hence showed that the simplex method is an exponential algorithm in the worst case. It is still an open question, however, whether there exists a rule for choosing entering and leaving basic variables that makes the simplex method polynomial. The first polynomial-time method for linear programming was given by Khachiyan [Kha79, Kha80], by adapting the ellipsoid method for nonlinear programming of Shor [Sho70a, Sho70b, Sho77] and Yudin and Nemirovskii [YuN77]. Karmarkar [Kar84a, Kar84b] showed that interior point methods can be used in order to solve linear programming problems in polynomial time.

General text books that discusses the simplex method are [Dan63, Chv83, Mur83, Sch86, Pad99, Van01, DaT97, DaT03].

9.5 Exercises

Exercise 9.1 (checking feasibility: phase I) Consider the system

$$\begin{aligned} 3x_1 + 2x_2 - x_3 &\leq -3, \\ -x_1 - x_2 + 2x_3 &\leq -1, \\ x_1, \quad x_2, \quad x_3 &\geq 0. \end{aligned}$$

Show that this system is infeasible. ■

Exercise 9.2 (the Simplex algorithm: phase I & II) Consider the linear program

$$\begin{aligned} \text{minimize} \quad z = \;& 3x_1 + 2x_2 + x_3 \\ \text{subject to} \quad & 2x_1 \phantom{{}+2x_2} + x_3 \geq 3, \\ & 2x_1 + 2x_2 + x_3 = 5, \\ & x_1, \quad x_2, \quad x_3 \geq 0. \end{aligned}$$

(a) Solve the linear program by using the Simplex algorithm with Phase I & II.

(b) Is the solution obtained unique? ■

Exercise 9.3 (the Simplex algorithm) Consider the linear program in standard form,

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n. \end{aligned}$$

Suppose that at a given step of the Simplex algorithm, there is only one possible entering variable, $(\boldsymbol{x}_N)_j$. Also assume that the current BFS is non-degenerate. Show that $(\boldsymbol{x}_N)_j > 0$ in any optimal solution. ■

Linear programming duality and sensitivity analysis

X

10.1 Introduction

Consider the linear program

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n, \end{aligned} \tag{10.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, and $\boldsymbol{c} \in \mathbb{R}^n$, and assume that this problem has been solved by the Simplex algorithm. Let $\boldsymbol{x}^* = (\boldsymbol{x}_B^{\mathrm{T}}, \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}}$ be an optimal basic feasible solution corresponding to the partition $\boldsymbol{A} = (\boldsymbol{B}, \boldsymbol{N})$. Introduce the vector $\boldsymbol{y}^* \in \mathbb{R}^m$ through

$$(\boldsymbol{y}^*)^{\mathrm{T}} = \boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{B}^{-1}.$$

Since $\boldsymbol{x}^*$ is an optimal solution it follows that the reduced costs of the non-basic variables are greater than or equal to zero, that is,

$$\boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{N} \geq (\boldsymbol{0}^{n-m})^{\mathrm{T}} \quad \Longleftrightarrow \quad \boldsymbol{c}_N^{\mathrm{T}} - (\boldsymbol{y}^*)^{\mathrm{T}}\boldsymbol{N} \geq (\boldsymbol{0}^{n-m})^{\mathrm{T}}.$$

Further, $\boldsymbol{c}_B^{\mathrm{T}} - (\boldsymbol{y}^*)^{\mathrm{T}}\boldsymbol{B} = \boldsymbol{c}_B^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{B} = (\boldsymbol{0}^m)^{\mathrm{T}}$, so actually we have that

$$\boldsymbol{c}^{\mathrm{T}} - (\boldsymbol{y}^*)^{\mathrm{T}}\boldsymbol{A} \geq (\boldsymbol{0}^n)^{\mathrm{T}},$$

or equivalently,

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}^* \leq \boldsymbol{c}.$$

Now, for every $\boldsymbol{y} \in \mathbb{R}^m$ such that $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} \leq \boldsymbol{c}$ and every feasible solution $\boldsymbol{x}$ to (10.1) it holds that

$$\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \geq \boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{b} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}.$$

But

$$\boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}^* = \boldsymbol{b}^{\mathrm{T}}(\boldsymbol{B}^{-1})^{\mathrm{T}}\boldsymbol{c}_B = \boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{b} = \boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{x}_B \leq \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x},$$

for every feasible solution $\boldsymbol{x}$ to (10.1) (since $\boldsymbol{x}^* = (\boldsymbol{x}_B^{\mathrm{T}}, \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}}$ is optimal), so in fact we have that $\boldsymbol{y}^*$ is an optimal solution to the linear program

$$\begin{aligned} \text{maximize} \quad & \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y} \\ \text{subject to} \quad & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} \leq \boldsymbol{c}, \\ & \boldsymbol{y} \text{ free.} \end{aligned} \tag{10.2}$$

Observe that the linear program (10.2) is exactly the Lagrangian dual problem to (10.1) (see Section 6.2.4). Also, note that the linear programs (10.1) and (10.2) have the same optimal objective value, which is in accordance with the Strong Duality Theorem 6.11 (see also Theorem 10.6 below for an independent proof).

The linear program (10.2) is called the *linear programming dual* to the linear program (10.1) (which is called the *primal linear program*). In this chapter we will study linear programming duality. In Section 10.2 we discuss how to construct the linear programming dual to a general linear program. Section 10.3 presents duality theory, such as weak and strong duality and complementary slackness. Finally, in Section 10.5 we discuss how the optimal solutions of a linear program change if the right-hand side $\boldsymbol{b}$ or the objective function coefficients $\boldsymbol{c}$ are modified.

10.2 The linear programming dual

For every linear program it is possible to construct the Lagrangian dual problem. From now on we will call this problem the *dual linear program*. It is quite tedious to construct the Lagrangian dual problem for every special case of a linear program, but fortunately the dual of a general linear program can be constructed just by following some simple rules. These rules are presented in this section. (It is, however, a good exercise to show the validity of these rules by constructing the Lagrangian dual in each case.)

10.2.1 Canonical form

When presenting the rules for constructing the linear programming dual we will use the notation of *canonical form*. The canonical form is connected with the inequalities of the problem and the objective function. If the objective is to maximize the objective function, then every inequality of type "$\leq$" is said to be of canonical form. Similarly, if the objective is to minimize the objective function, then every inequality of type "$\geq$" is said to be of canonical form. Further, we consider non-negative variables to be variables in canonical form.

Remark 10.1 (mnemonic rule for canonical form) Consider the problem to

$$\begin{array}{ll} \text{minimize} & z = x_1 \\ \text{subject to} & x_1 \leq 1. \end{array}$$

This problem is unbounded from below and hence an optimal solution does not exist. However, if the problem is to

$$\begin{array}{ll} \text{minimize} & z = x_1 \\ \text{subject to} & x_1 \geq 1, \end{array}$$

then an optimal solution exists, namely $x_1 = 1$. Hence it seems natural to consider inequalities of type "$\geq$" as canonical to minimization problems. Similarly, it is natural that inequalities of type "$\leq$" are canonical to maximization problems. ■

10.2.2 Constructing the dual

From the notation of canonical form introduced in Section 10.2.1 we can now construct the dual, (D), to a general linear program, (P), according to the following rules.

Dual variables

To each constraint of (P) a dual variable, y_i, is introduced. If the i^{th} constraint of (P) is an inequality of canonical form, then y_i is a non-negative variable, that is, $y_i \geq 0$. Similarly, if the i^{th} constraint of (P) is an inequality that is not of canonical form, then $y_i \leq 0$. Finally, if the i^{th} constraint of (P) is an equality, then the variable y_i is unrestricted.

Dual objective function

If (P) is a minimization (respectively, a maximization) problem, then (D) is a maximization (respectively, a minimization) problem. The objective function coefficient for the variable y_i in the dual problem equals the right-hand side constant of the i^{th} constraint of (P).

Constraints of the dual problem

If $\boldsymbol{A}$ is the constraint matrix of (P), then $\boldsymbol{A}^{\mathrm{T}}$ is the constraint matrix of (D). The j^{th} right-hand side constant of (D) equals the j^{th} coefficient in the objective function of (P). If the j^{th} variable of (P) has non-negativity restriction, then the j^{th} constraint of (D) is an inequality of canonical form. If the j^{th} variable of (P) has a non-positivity restriction, then the j^{th} constraint of (D) is an inequality of non-canonical form. Finally, if the j^{th} variable of (P) is unrestricted, then the j^{th} constraint of (D) is an equality.

Summary

The rules above can be summarized as follows:

primal/dual constraint		**dual/primal variable**
canonical inequality	$\iff$	≥ 0
non-canonical inequality	$\iff$	≤ 0
equality	$\iff$	unrestricted

Consider the following general linear program:

$$
\begin{aligned}
\text{minimize} \quad & z = \sum_{j=1}^{n} c_j x_j \\
\text{subject to} \quad & \sum_{j=1}^{n} a_{ij} x_j \geq b_i, \quad i \in C, \\
& \sum_{j=1}^{n} a_{ij} x_j \leq b_i, \quad i \in NC, \\
& \sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i \in E, \\
& \qquad x_j \geq 0, \quad j \in P, \\
& \qquad x_j \leq 0, \quad j \in N, \\
& \qquad x_j \text{ free}, \quad j \in F,
\end{aligned}
$$

where C stands for "canonical", NC for "non-canonical", E for "equality", P for "positive", N for "negative", and F for "free". Note that $P \cup N \cup F = \{1, \dots, n\}$ and $C \cup NC \cup E = \{1, \dots, m\}$. If we apply the rules above we get the following dual linear program:

$$
\begin{aligned}
\text{maximize} \quad & w = \sum_{i=1}^{m} b_i y_i \\
\text{subject to} \quad & \sum_{i=1}^{m} a_{ij} y_i \leq c_j, \quad j \in P, \\
& \sum_{i=1}^{m} a_{ij} y_i \geq c_j, \quad j \in N, \\
& \sum_{i=1}^{m} a_{ij} y_i = c_j, \quad j \in F, \\
& y_i \geq 0, \quad i \in C, \\
& y_i \leq 0, \quad i \in NC, \\
& y_i \text{ free}, \quad i \in E.
\end{aligned}
$$

From this it is easily established that if we construct the dual of the dual linear program, then we return to the original (primal) linear program.

Examples

In order to illustrate how to construct the dual linear program we present two examples. The first example considers a linear program with block structure. This is a usual form of linear programs and it is particularly easy to construct the dual linear program. The other example deals with the transportation problem presented in Section 8.1. The purpose of constructing the dual to this problem is to show how to handle double subscripted variables and indexed constraints.

Example 10.2 (the dual to a linear program of block form) Consider the linear program

$$
\begin{aligned}
\text{maximize} \quad & \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{d}^{\mathrm{T}}\boldsymbol{y} \\
\text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} \leq \boldsymbol{b}, \\
& \boldsymbol{D}\boldsymbol{y} = \boldsymbol{e}, \\
& \boldsymbol{x} \geq \boldsymbol{0}^{n_1}, \\
& \boldsymbol{y} \leq \boldsymbol{0}^{n_2},
\end{aligned}
$$

where $\boldsymbol{A} \in \mathbb{R}^{m_1 \times n_1}$, $\boldsymbol{D} \in \mathbb{R}^{m_2 \times n_2}$, $\boldsymbol{b} \in \mathbb{R}^{m_1}$, $\boldsymbol{e} \in \mathbb{R}^{m_2}$, $\boldsymbol{c} \in \mathbb{R}^{n_1}$, and $\boldsymbol{d} \in \mathbb{R}^{n_2}$. The dual of this linear program becomes that to

$$\begin{array}{lrl} \text{minimize} & \boldsymbol{b}^{\mathrm{T}}\boldsymbol{u} + \boldsymbol{e}^{\mathrm{T}}\boldsymbol{v} & \\ \text{subject to} & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{u} & \geq \boldsymbol{c}, \\ & \boldsymbol{B}^{\mathrm{T}}\boldsymbol{u} + \boldsymbol{D}^{\mathrm{T}}\boldsymbol{v} & \leq \boldsymbol{d}, \\ & \boldsymbol{u} & \geq \boldsymbol{0}^{m_1}, \\ & \boldsymbol{v} & \text{free.} \end{array}$$

Observe that the constraint matrix of the primal problem is

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{D} \end{pmatrix},$$

and if we transpose this matrix we get

$$\begin{pmatrix} \boldsymbol{A}^{\mathrm{T}} & \boldsymbol{0}^{\mathrm{T}} \\ \boldsymbol{B}^{\mathrm{T}} & \boldsymbol{D}^{\mathrm{T}} \end{pmatrix}.$$

Also note that the vector of objective function coefficients of the primal problem, $(\boldsymbol{c}^{\mathrm{T}}, \boldsymbol{d}^{\mathrm{T}})^{\mathrm{T}}$, is the right-hand side of the dual problem, and the right-hand side of the primal problem, $(\boldsymbol{b}^{\mathrm{T}}, \boldsymbol{e}^{\mathrm{T}})^{\mathrm{T}}$, is the vector of objective function coefficients of the dual problem. ■

Example 10.3 (the dual of the transportation problem) Consider the transportation problem (see Example 8.1) to

$$\begin{array}{ll} \text{minimize} & z = \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{M} c_{ij}x_{ij} \\ \text{subject to} & \displaystyle\sum_{j=1}^{M} x_{ij} \leq s_i, \quad i = 1, \dots, N, \\ & \displaystyle\sum_{i=1}^{N} x_{ij} \geq d_j, \quad j = 1, \dots, M, \\ & \quad x_{ij} \geq 0, \quad i = 1, \dots, N, \quad j = 1, \dots, M. \end{array}$$

The dual linear program is given by

$$\begin{array}{ll} \text{maximize} & w = \displaystyle\sum_{i=1}^{N} s_i u_i + \sum_{j=1}^{M} d_j v_j \\ \text{subject to} & u_i + v_j \leq c_{ij}, \quad i = 1, \dots, N, \quad j = 1, \dots, M, \\ & \quad u_i \leq 0, \quad i = 1, \dots, N, \\ & \quad v_j \geq 0, \quad j = 1, \dots, M. \end{array}$$

Observe that there are $N+M$ constraints in the primal problem and hence there are $N+M$ dual variables. Also, there are NM variables of the primal problem, yielding NM constraints in the dual problem. The form of the constraints in the dual problem arises from the fact that x_{ij} appears twice in the column of the constraint matrix corresponding to this variable: once in the constraints over $i = 1, \dots, N$ and once in the constraints over $j = 1, \dots, M$. Also note that all the coefficients of the constraint matrix in the primal problem equal $+1$, and since we have one dual constraint for each column, we finally get the dual constraint $u_i + v_j \leq c_{ij}$. ∎

10.3 Linear programming duality theory

In this section we present some of the most fundamental duality theorems. Throughout the section we will consider the primal linear program

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n, \end{aligned} \tag{P}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, and $\boldsymbol{c} \in \mathbb{R}^n$, and its dual linear program

$$\begin{aligned} \text{maximize} \quad & w = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y} \\ \text{subject to} \quad & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} \leq \boldsymbol{c}, \\ & \boldsymbol{y} \ \text{free}. \end{aligned} \tag{D}$$

The theorems presented actually hold for every primal–dual pair of linear programs, and the proofs are similar to those presented here. See also the more general statements on the differentiability of the perturbation function in Section 6.7.

10.3.1 Weak and strong duality

We begin by proving the Weak Duality Theorem.

Theorem 10.4 (Weak Duality Theorem) *If $\boldsymbol{x}$ is a feasible solution to* (P) and $\boldsymbol{y}$ a feasible solution to (D), then $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \geq \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}$.

Proof. We have that

$$\begin{aligned} \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} &\geq (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y})^{\mathrm{T}}\boldsymbol{x} \qquad && [\boldsymbol{c} \geq \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}, \quad \boldsymbol{x} \geq \boldsymbol{0}^n] \\ &= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{b} && [\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}] \\ &= \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}, \end{aligned}$$

and we are done. ∎

Corollary 10.5 *If $\boldsymbol{x}$ is feasible solution to* (P)*, $\boldsymbol{y}$ a feasible solution to* (D)*, and $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}$, then $\boldsymbol{x}$ is an optimal solution to* (P) *and $\boldsymbol{y}$ and optimal solution to* (D)*.* ∎

Next we show that the duality gap is zero, that is, strong duality holds. Note that this can also be established by the use of the Lagrangian duality theory in Chapter 6.

Theorem 10.6 (Strong Duality Theorem) *If one of the problems* (P) *and* (D) *has a finite optimal solution, then so does its dual, and their optimal objective values are equal.*

Proof. Suppose that $\boldsymbol{x}^* = (\boldsymbol{x}_B^{\mathrm{T}}, \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}}$ is an optimal basic feasible solution to (P) corresponding to the partition $\boldsymbol{A} = (\boldsymbol{B}, \boldsymbol{N})$. We construct an optimal solution to (D). (Actually we have already done this in detail in Section 10.1.) Let

$$(\boldsymbol{y}^*)^{\mathrm{T}} = \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1}.$$

Since $\boldsymbol{x}^*$ is an optimal basic feasible solution the reduced costs of the non-basic variables are non-negative, which gives that (for details see Section 10.1)

$$\boldsymbol{A}^{\mathrm{T}} \boldsymbol{y}^* \leq \boldsymbol{c}.$$

Hence, $\boldsymbol{y}^*$ is feasible to (D). Further, we have that

$$\boldsymbol{b}^{\mathrm{T}} \boldsymbol{y}^* = \boldsymbol{b}^{\mathrm{T}} (\boldsymbol{B}^{-1})^{\mathrm{T}} \boldsymbol{c}_B = \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{b} = \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{x}_B = \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x}^*,$$

so by Corollary 10.5 it follows that $\boldsymbol{y}^*$ is an optimal solution to (D).

Now suppose instead that the dual problem has a finite optimal solution. Convert (D) into standard form:

$$\begin{aligned}
\text{minimize} \quad & \tilde{w} = -\boldsymbol{b}^{\mathrm{T}} \boldsymbol{y}^+ + \boldsymbol{b}^{\mathrm{T}} \boldsymbol{y}^- \\
\text{subject to} \quad & \begin{pmatrix} \boldsymbol{A}^{\mathrm{T}} & -\boldsymbol{A}^{\mathrm{T}} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{y}^+ \\ \boldsymbol{y}^- \\ \boldsymbol{s} \end{pmatrix} = \boldsymbol{c}, \\
& \boldsymbol{y}^+, \boldsymbol{y}^- \geq \boldsymbol{0}^m, \\
& \boldsymbol{s} \geq \boldsymbol{0}^n.
\end{aligned} \tag{D'}$$

If there exists an optimal solution $\boldsymbol{y}^*$ to (D), then there exists an optimal basic feasible solution to (D') (with $\tilde{w}^* = -w^*$). As in the first part of the proof we then get that there exists an $\boldsymbol{x}^* \in \mathbb{R}^n$ such that

$$\begin{pmatrix} \boldsymbol{A} \\ -\boldsymbol{A} \\ \boldsymbol{I} \end{pmatrix} \boldsymbol{x}^* \leq \begin{pmatrix} -\boldsymbol{b} \\ \boldsymbol{b} \\ \boldsymbol{0}^n \end{pmatrix},$$

and

$$\tilde{w}^* = \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x}^*.$$

Hence, by Corollary 10.5, we get that $-\boldsymbol{x}^*$ is an optimal solution to (P). We are done. ∎

Remark 10.7 (dual solution from the primal solution) Note that the proof of Theorem 10.6 is constructive. We actually construct an optimal dual solution from an optimal basic feasible solution by

$$(\boldsymbol{y}^*)^{\mathrm{T}} = \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1}. \tag{10.3}$$

When a linear program is solved by the Simplex method we obtain an optimal basic feasible solution (if the LP is not unbounded or infeasible). Hence from (10.3) we then also—without any additional effort—obtain an optimal dual solution. In fact the dual solution is calculated in the pricing step of the Simplex algorithm. ∎

Interpretation of the optimal dual solution

We have from (10.3) that

$$\boldsymbol{b}^{\mathrm{T}} \boldsymbol{y}^* = \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{b},$$

for any optimal basic feasible solution to (P). If $\boldsymbol{x}_B > \boldsymbol{0}^m$, then a small change in $\boldsymbol{b}$ does not change the basis, and so the optimal value of (D) (and (P)), namely

$$v(\boldsymbol{b}) := \boldsymbol{b}^{\mathrm{T}} \boldsymbol{y}^*$$

is linear at, and locally around, the value $\boldsymbol{b}$. If, however, some $(\boldsymbol{x}_B)_i = 0$, then in this degenerate case it could be that the basis changes in a non-differentiable manner with $\boldsymbol{b}$. We summarize:

Theorem 10.8 (shadow price) *If, for a given $\boldsymbol{b}$, the optimal solution to* (P) *corresponds to a non-degenerate basic feasible solution, then its optimal value is differentiable at $\boldsymbol{b}$, with*

$$\frac{\partial v(\boldsymbol{b})}{\partial b_i} = y_i^*, \qquad i = 1, \dots, m,$$

that is, $\nabla v(\boldsymbol{b}) = \boldsymbol{y}^*$. ■

Remark 10.9 (shadow price) The optimal dual solution is indeed the shadow price for the constraints. If a unit change in one right-hand side b_i does not change the optimal basis, then the above states that the optimal value will change exactly with the amount y_i^*.

It is also clear that non-degeneracy at $\boldsymbol{x}^*$ in (P) implies that the optimal solution in (D) must be unique. Namely, we can show that the function v is convex on its effective domain (why?) and the non-degeneracy property clearly implies that v is also finite in a neighbourhood of $\boldsymbol{b}$. Then, its differentiability at $\boldsymbol{b}$ is equivalent to the uniqueness of its subgradients at $\boldsymbol{b}$; cf. Proposition 6.16(c). ■

Farkas' Lemma

In Section 3.2 we proved Farkas' Lemma 3.30 by using the Separation Theorem 3.24. However, Farkas' Lemma 3.30 can easily be proved by using the Strong Duality Theorem 10.6.

Theorem 10.10 (Farkas' Lemma) *Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. Then, exactly one of the systems*

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} &= \boldsymbol{b}, \\ \boldsymbol{x} &\geq \boldsymbol{0}^n, \end{aligned} \tag{I}$$

and

$$\begin{aligned} \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} &\leq \boldsymbol{0}^n, \\ \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y} &> 0, \end{aligned} \tag{II}$$

has a feasible solution, and the other system is inconsistent.

Proof. If (I) has a solution $\boldsymbol{x}$, then

$$\boldsymbol{b}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} > 0.$$

But $\boldsymbol{x} \geq \boldsymbol{0}^n$, so $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} \leq \boldsymbol{0}^n$ cannot hold, which means that (II) is infeasible.

Assume that (II) is infeasible. Consider the linear program

$$\begin{aligned} \text{maximize} \quad & \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y} \\ \text{subject to} \quad & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} \leq \boldsymbol{0}^n, \\ & \boldsymbol{y} \quad \text{free}, \end{aligned} \tag{10.4}$$

and its dual program

$$\begin{aligned} \text{minimize} \quad & (\boldsymbol{0}^n)^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n. \end{aligned} \tag{10.5}$$

Since (II) is infeasible, $\boldsymbol{y} = \boldsymbol{0}^m$ is an optimal solution to (10.4). Hence the Strong Duality Theorem 10.6 implies that there exists an optimal solution to (10.5). This solution is feasible to (I).

What we have proved above is the equivalence

$$\text{(I)} \qquad \Longleftrightarrow \qquad \neg\text{(II)}.$$

Logically, this is equivalent to the statement that

$$\neg\text{(I)} \qquad \Longleftrightarrow \qquad \text{(II)}.$$

We have hence established that precisely one of the two systems (I) and (II) has a solution. We are done. ∎

10.3.2 Complementary slackness

A further relationship between (P) and (D) at an optimal solution is given by the Complementary Slackness Theorem.

Theorem 10.11 (Complementary Slackness Theorem) *Let $\boldsymbol{x}$ be a feasible solution to* (P) *and $\boldsymbol{y}$ a feasible solution to* (D). *Then $\boldsymbol{x}$ is optimal to* (P) *and $\boldsymbol{y}$ optimal to* (D) *if and only if*

$$x_j(c_j - \boldsymbol{A}_{\cdot j}^{\mathrm{T}}\boldsymbol{y}) = 0, \qquad j = 1, \dots, n, \tag{10.6}$$

where $\boldsymbol{A}_{\cdot j}$ is the j^{th} column of $\boldsymbol{A}$.

Proof. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are feasible we get

$$\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \geq (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y})^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}. \tag{10.7}$$

Further, by the Strong Duality Theorem 10.6 and the Weak Duality Theorem 10.4, $\boldsymbol{x}$ and $\boldsymbol{y}$ are optimal if and only if $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}$, so in fact (10.7) holds with equality, that is,

$$\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y})^{\mathrm{T}}\boldsymbol{x} \quad \Longleftrightarrow \quad \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{c} - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}) = 0.$$

Since $\boldsymbol{x} \geq \boldsymbol{0}^n$ and $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} \leq \boldsymbol{c}$, $\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{c} - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}) = 0$ is equivalent to that each term in the sum is zero, that is, (10.6) holds. ∎

Often the Complementary Slackness Theorem is stated for the primal–dual pair given by

$$\begin{aligned} \text{maximize} \quad & \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n, \end{aligned} \tag{10.8}$$

and

$$\begin{aligned} \text{minimize} \quad & \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y} \\ \text{subject to} \quad & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} \geq \boldsymbol{c}, \\ & \boldsymbol{y} \geq \boldsymbol{0}^m. \end{aligned} \tag{10.9}$$

The Complementary Slackness Theorem then becomes as follows. (Its proof is similar to that of Theorem 10.11.)

Theorem 10.12 (Complementary Slackness Theorem) *Let $\boldsymbol{x}$ be a feasible solution to* (10.8) *and $\boldsymbol{y}$ a feasible solution to* (10.9). *Then $\boldsymbol{x}$ is optimal to* (10.8) *and $\boldsymbol{y}$ optimal to* (10.9) *if and only if*

$$x_j(c_j - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}_{\cdot j}) = 0, \qquad j = 1, \dots, n, \tag{10.10a}$$
$$y_i(\boldsymbol{A}_{i\cdot}\boldsymbol{x} - b_i) = 0, \qquad i = 1 \dots, m, \tag{10.10b}$$

where $\boldsymbol{A}_{\cdot j}$ is the j^{th} column of $\boldsymbol{A}$ and $\boldsymbol{A}_{i\cdot}$ the i^{th} row of $\boldsymbol{A}$. ∎

Remark 10.13 (interpretation of the Complementary Slackness Theorem) From the Complementary Slackness Theorem follows that, for an optimal primal dual pair of solutions, if there is slack in one constraint, then the respective variable in the other problem is zero. Further, if a variable is positive, then there is no slack in the respective constraint in the other problem. ∎

The consequence of the Complementary Slackness Theorem is the following characterization of an optimal solution to a linear program. We state it for the primal–dual pair given by (10.8) and (10.9), but it holds as well for each primal dual pair of linear programs.

Theorem 10.14 (necessary and sufficient conditions for global optimality) *Take a vector $\boldsymbol{x} \in \mathbb{R}^n$. For $\boldsymbol{x}$ to be an optimal solution to the linear program* (10.8), *it is both necessary and sufficient that*

(a) *$\boldsymbol{x}$ is a feasible solution to* (10.8)*;*

(b) *corresponding to $\boldsymbol{x}$ there is a dual feasible solution $\boldsymbol{y} \in \mathbb{R}^m$ to* (10.9)*; and*

(c) *the pair $(\boldsymbol{x}, \boldsymbol{y})$ satisfies the complementarity conditions* (10.10).

■

The simplex method is very well adapted to these conditions. After Phase I, (a) holds. Every basic solution (feasible or not) satisfies (c), since if x_j is in the basis, then $\tilde{c}_j = c_j - \boldsymbol{y}^{\mathrm{T}} \boldsymbol{A}_{\cdot j} = 0$, and if $\tilde{c}_j \neq 0$, then $x_j = 0$. So, the only condition that the Simplex method does not satisfy for every basic feasible solution is (b). The proof of the Strong Duality Theorem 10.6 shows that it is satisfied exactly at an optimal basic feasible solution. The entering criterion is based on trying to better satisfy it. Indeed, by choosing as an entering variable x_j such that

$$j \in \arg \underset{j \in \{1,\dots,n\}}{\operatorname{minimum}} \tilde{c}_j,$$

we actually identify a dual constraint

$$\sum_{i=1}^{m} a_{ij} y_i \leq c_j,$$

which is among *the most* violated at the complementary solution $\boldsymbol{y}^{\mathrm{T}} = \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1}$ given by the current BFS. After the basis change we will have equality in this dual constraint, and hence the basis change corresponds to making a currently most violated dual constraint feasible!

Example 10.15 (illustration of complementary slackness) Consider the primal–dual pair given by

$$\begin{array}{llrrl} \text{maximize} & z = & 3x_1 & +2x_2 & \\ \text{subject to} & & x_1 & +x_2 & \leq 80, \\ & & 2x_1 & +x_2 & \leq 100, \\ & & x_1 & & \leq 40, \\ & & x_1, & x_2 & \geq 0, \end{array} \tag{10.11}$$

and

$$\begin{array}{llrrrl} \text{minimize} & w = & 80y_1 & +100y_2 & +40y_3 & \\ \text{subject to} & & y_1 & +2y_2 & +y_3 & \geq 3, \\ & & y_1 & +y_2 & & \geq 2, \\ & & y_1, & y_2, & y_3 & \geq 0. \end{array} \tag{10.12}$$

We use Theorem 10.14 to show that $\boldsymbol{x}^* = (20, 60)^{\mathrm{T}}$ is an optimal solution to (10.11).

(a) (primal feasibility) Obviously $\boldsymbol{x}^*$ is a feasible solution to (10.11).

(c) (complementary) The complementary conditions must hold, that is,

$$\begin{aligned}
y_1^*(x_1^* + x_2^* - 80) &= 0 \\
y_2^*(2x_1^* + x_2^* - 100) &= 0 \\
y_3^*(x_1^* - 40) &= 0 \quad \Longrightarrow \quad y_3^* = 0 \quad [x_1^* = 20 \neq 40] \\
x_1^*(y_1^* + 2y_2^* + y_3^* - 3) &= 0 \quad \Longrightarrow \quad y_1^* + 2y_2^* = 3 \quad [x_1^* > 0] \\
x_2^*(y_1^* + y_2^* - 2) &= 0 \quad \Longrightarrow \quad y_1^* + y_2^* = 2 \quad [x_2^* > 0]
\end{aligned}$$

which gives that $y_1^* = 1$, $y_2^* = 1$ and $y_3^* = 0$.

(b) (dual feasibility) Obviously $\boldsymbol{y}^* = (1, 1, 0)^{\mathrm{T}}$ is a feasible solution to (10.12).

From Theorem 10.14 it then follows that $\boldsymbol{x}^* = (20, 60)^{\mathrm{T}}$ is an optimal solution to (10.11) and $\boldsymbol{y}^* = (1, 1, 0)^{\mathrm{T}}$ an optimal solution to (10.12). ■

10.4 The Dual Simplex method

The Simplex method presented in Chapter 9, which we here refer to as the *primal Simplex method*, starts with a basic feasible solution to the primal linear program and then iterates until the primal optimality conditions are fulfilled, that is, until a basic feasible solution is found such that the reduced costs

$$\bar{\boldsymbol{c}}_N^{\mathrm{T}} := \boldsymbol{c}_N^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{N} \geq (\boldsymbol{0}^{n-m})^{\mathrm{T}}.$$

This is equivalent to the dual feasibility condition

$$\boldsymbol{A}^{\mathrm{T}} \boldsymbol{y} \leq \boldsymbol{c},$$

where $\boldsymbol{y} = (\boldsymbol{B}^{-1})^{\mathrm{T}} \boldsymbol{c}_B$. We call a basis such that all of the reduced costs are greater than or equal to zero a *dual feasible basis*; otherwise we call it a *dual infeasible basis*. Hence, the primal Simplex method starts with a primal feasible basis and then moves through a sequence of dual infeasible (but primal feasible) bases until a dual (and primal) feasible basis is found.

The *Dual Simplex method* is a variant of the primal Simplex method that works in a dual manner in the sense that it starts with a dual feasible basis and then moves through a sequence of primal infeasible (but dual feasible) bases until a primal (and dual) feasible basis is found.

In order to derive the Dual Simplex algorithm, let $\boldsymbol{x}_B$ be a dual feasible basis with the corresponding partition $(\boldsymbol{B}, \boldsymbol{N})$. If

$$\bar{\boldsymbol{b}} = \boldsymbol{B}^{-1}\boldsymbol{b} \geq \boldsymbol{0}^m,$$

then $\boldsymbol{x}_B$ is primal feasible and since it is also dual feasible all of the reduced costs are greater than or equal to zero; hence, $\boldsymbol{x}_B$ is an optimal BFS. Otherwise some of the components of $\tilde{\boldsymbol{b}}$ is strictly negative, say $\bar{b}_1$, that is,

$$(\boldsymbol{x}_B)_1 + \sum_{j=1}^{n-m} (\boldsymbol{B}^{-1}\boldsymbol{N})_{1j}(\boldsymbol{x}_N)_j = \bar{b}_1 < 0,$$

so $(\boldsymbol{x}_B)_1 < 0$ in the current basis and will be the leaving variable. Now, if

$$(\boldsymbol{B}^{-1}\boldsymbol{N})_{1j} \geq 0, \quad j = 1, \dots, n-m, \tag{10.13}$$

then there exists no primal feasible solution to the problem. (Why?) Hence, if (10.13) is fulfilled, then we say that the *primal infeasibility criterion* is satisfied. Otherwise $(\boldsymbol{B}^{-1}\boldsymbol{N})_{1j} < 0$ for some $j = 1, \dots, n-m$. Assume that $(\boldsymbol{B}^{-1}\boldsymbol{N})_{1k} < 0$ and choose $(\boldsymbol{x}_N)_k$ to replace $(\boldsymbol{x}_B)_1$ in the basis. (Note that this yields that $(\boldsymbol{x}_N)_k = \bar{b}_1/(\boldsymbol{B}^{-1}\boldsymbol{N})_{1k} > 0$ in the new basis.) The new reduced costs then become

$$\begin{aligned}
&(\tilde{\boldsymbol{c}}_B)_1 = -\frac{1}{(\boldsymbol{B}^{-1}\boldsymbol{N})_{1k}},\\
&(\tilde{\boldsymbol{c}}_B)_j = 0, \quad j = 2, \dots, m,\\
&(\tilde{\boldsymbol{c}}_N)_j = (\bar{\boldsymbol{c}}_N)_j - (\bar{\boldsymbol{c}}_N)_k \frac{(\boldsymbol{B}^{-1}\boldsymbol{N})_{1j}}{(\boldsymbol{B}^{-1}\boldsymbol{N})_{1k}}, \quad j = 1, \dots, n-m.
\end{aligned}$$

Since we want the new basis to be dual feasible it must hold that all of the new reduced costs are non-negative, that is,

$$(\bar{\boldsymbol{c}}_N)_j \geq (\bar{\boldsymbol{c}}_N)_k \frac{(\boldsymbol{B}^{-1}\boldsymbol{N})_{1j}}{(\boldsymbol{B}^{-1}\boldsymbol{N})_{1k}}, \quad j = 1, \dots, n-m,$$

or, equivalently,

$$\frac{(\bar{\boldsymbol{c}}_N)_k}{(\boldsymbol{B}^{-1}\boldsymbol{N})_{1k}} \geq \frac{(\bar{\boldsymbol{c}}_N)_j}{(\boldsymbol{B}^{-1}\boldsymbol{N})_{1j}}, \quad \text{for all } j \text{ such that } (\boldsymbol{B}^{-1}\boldsymbol{N})_{1j} < 0.$$

Therefore, in order to preserve dual feasibility, as entering variable we must choose $(\boldsymbol{x}_N)_k$ such that

$$k \in \arg \operatorname*{maximum}_{i \in \{\, j \,|\, (\boldsymbol{B}^{-1}\boldsymbol{N})_{1j} < 0 \,\}} \frac{(\bar{\boldsymbol{c}}_N)_j}{(\boldsymbol{B}^{-1}\boldsymbol{N})_{1j}}.$$

We have now derived an infeasibility criterion and criteria for how to choose the leaving and the entering variables, and are ready to state the Dual Simplex algorithm:

The Dual Simplex Algorithm:

Step 0 (initialization: DFS) Assume that $\boldsymbol{x}^{\mathrm{T}} = (\boldsymbol{x}_B^{\mathrm{T}}, \boldsymbol{x}_N^{\mathrm{T}})$ is a dual feasible basis corresponding to the partition $\boldsymbol{A} = (\boldsymbol{B}, \boldsymbol{N})$.

Step 1 (leaving variable or termination) Calculate

$$\bar{\boldsymbol{b}} = \boldsymbol{B}^{-1}\boldsymbol{b}.$$

If $\bar{\boldsymbol{b}} \geq \boldsymbol{0}^m$, then stop; the current basis is optimal. Otherwise, choose an s such that $\bar{b}_s < 0$, and let $(\boldsymbol{x}_B)_s$ be the leaving variable.

Step 2 (entering variable or termination) If

$$(\boldsymbol{B}^{-1}\boldsymbol{N})_{sj} \geq 0, \quad j = 1, \ldots, n - m,$$

then stop; the (primal) problem is infeasible. Otherwise, choose a k such that

$$k \in \arg \underset{i \in \{\, j \mid (\boldsymbol{B}^{-1}\boldsymbol{N})_{sj} < 0 \,\}}{\operatorname{maximum}} \frac{(\bar{\boldsymbol{c}}_N)_j}{(\boldsymbol{B}^{-1}\boldsymbol{N})_{sj}},$$

and let $(\boldsymbol{x}_N)_k$ be the entering variable.

Step 3 (update: change basis) Construct a new partition by swapping $(\boldsymbol{x}_B)_s$ with $(\boldsymbol{x}_N)_k$. Go to Step 1.

Similarly to the primal Simplex Algorithm it can be shown that the Dual Simplex Algorithm terminates in a finite number of steps if cycling is avoided. Also, there exist rules for choosing the leaving and entering variables (among the eligible ones) such that cycling is avoided.

If a dual feasible solution is not available from the start, it is possible to add a constraint to the original problem, that makes it possible to construct a dual feasible basis, and then run the Dual Simplex Algorithm on this modified problem (see Exercise 10.13).

Remark 10.16 (unboundedness of the primal problem) Since the dual problem is known to be feasible, the primal problem cannot be unbounded by the Weak Duality Theorem 10.4. Hence the Dual Simplex Algorithm terminates with a basis that satisfies either the optimality criterion or the primal infeasibility criterion. ∎

Example 10.17 (illustration of the Dual Simplex Algorithm) Consider the linear program

$$\begin{array}{llllllll}
\text{minimize} & 3x_1 & +4x_2 & +2x_3 & +x_4 & +5x_5 & & \\
\text{subject to} & x_1 & -2x_2 & -x_3 & +x_4 & +x_5 & \leq & -3, \\
 & -x_1 & -x_2 & -x_3 & +x_4 & +x_5 & \leq & -2, \\
 & x_1 & +x_2 & -2x_3 & +2x_4 & -3x_5 & \leq & 4, \\
 & x_1, & x_2, & x_3, & x_4, & x_5 & \geq & 0.
\end{array}$$

By introducing the slack variables x_6, x_7, x_8, we get the following linear program:

$$\begin{array}{lllllllllll}
\text{minimize} & 3x_1 & +4x_2 & +2x_3 & +x_4 & +5x_5 & & & & & \\
\text{subject to} & x_1 & -2x_2 & -x_3 & +x_4 & +x_5 & +x_6 & & & = & -3, \\
 & -x_1 & -x_2 & -x_3 & +x_4 & +x_5 & & +x_7 & & = & -2, \\
 & x_1 & +x_2 & -2x_3 & +2x_4 & -3x_5 & & & +x_8 & = & 4, \\
 & x_1, & x_2, & x_3, & x_4, & x_5, & x_6, & x_7, & x_8 & \geq & 0.
\end{array}$$

We see that the basis $\boldsymbol{x}_B = (x_6, x_7, x_8)^{\mathrm{T}}$ is dual feasible, but primal infeasible. Hence we use the Dual Simplex Algorithm to solve the problem. We have that

$$\bar{\boldsymbol{b}} = \boldsymbol{B}^{-1}\boldsymbol{b} = (-3, -2, 4)^{\mathrm{T}},$$

so we choose $(\boldsymbol{x}_B)_1 = x_6$ to leave the basis. Further we have that

$$\begin{aligned}
\bar{\boldsymbol{c}}^{\mathrm{T}} &= (3, 4, 2, 1, 5, 0, 0, 0), \\
(\boldsymbol{B}^{-1}\boldsymbol{A})_{1,\cdot} &= (1, -2, -1, 1, 1, 1, 0, 0), \qquad \text{[the 1st row of } \boldsymbol{B}^{-1}\boldsymbol{A}\text{]}
\end{aligned}$$

so we choose x_2 as the entering variable. The new basis becomes $\boldsymbol{x}_B = (x_2, x_7, x_8)^{\mathrm{T}}$. We get that

$$\bar{\boldsymbol{b}} = \boldsymbol{B}^{-1}\boldsymbol{b} = (1.5, -0.5, 2.5)^{\mathrm{T}}.$$

Hence, we choose $(\boldsymbol{x}_B)_2 = x_7$ as the leaving variable. Further

$$\begin{aligned}
\bar{\boldsymbol{c}}^{\mathrm{T}} &= (5, 0, 0, 3, 7, 2, 0, 0), \\
(\boldsymbol{B}^{-1}\boldsymbol{A})_{2,\cdot} &= (-1.5, 0, -0.5, 0.5, 0.5, -0.5, 1, 0),
\end{aligned}$$

which gives that x_3 is the entering variable. The new basis becomes $\boldsymbol{x}_B = (x_2, x_3, x_8)^{\mathrm{T}}$. We get that

$$\bar{\boldsymbol{b}} = \boldsymbol{B}^{-1}\boldsymbol{b} = (1, 1, 5)^{\mathrm{T}},$$

which means that the optimality criterion (primal feasibility) is satisfied, and an optimal solution to the original problem is given by

$$\boldsymbol{x}^* = (x_1, x_2, x_3, x_4, x_5)^{\mathrm{T}} = (0, 1, 1, 0, 0)^{\mathrm{T}}.$$

Check that this is indeed true, for example by using Theorem 10.12. ∎

10.5 Sensitivity analysis

In this section we study two kinds of perturbations of a linear program in standard form,

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n, \end{aligned} \tag{10.14}$$

namely

1. perturbations in the objective function coefficients c_j; and
2. perturbations in the right-hand side coefficients b_i.

We assume that $\boldsymbol{x}^* = (\boldsymbol{x}_B^{\mathrm{T}}, \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}} = ((\boldsymbol{B}^{-1}\boldsymbol{b})^{\mathrm{T}}, (\boldsymbol{0}^{n-m})^{\mathrm{T}})^{\mathrm{T}}$ is an optimal basic feasible solution to (10.14) with the corresponding partition $\boldsymbol{A} = (\boldsymbol{B}, \boldsymbol{N})$.

10.5.1 Perturbations in the objective function

Assume that the objective function coefficients of the linear program (10.14) are perturbed by the vector $\boldsymbol{p} \in \mathbb{R}^n$, that is, we consider the perturbed problem to

$$\begin{aligned} \text{minimize} \quad & \tilde{z} = (\boldsymbol{c} + \boldsymbol{p})^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n. \end{aligned} \tag{10.15}$$

The optimal solution $\boldsymbol{x}^*$ to the unperturbed problem (10.14) is obviously a feasible solution to (10.15), but is it still optimal? To answer this question, we note that a basic feasible solution is optimal if the reduced costs of the non-basic variables are greater than or equal to zero. The reduced costs for the non-basic variables of the perturbed problem (10.15) are given by [let $\boldsymbol{p}^{\mathrm{T}} = (\boldsymbol{p}_B^{\mathrm{T}}, \boldsymbol{p}_N^{\mathrm{T}})$]

$$\bar{\boldsymbol{c}}_N^{\mathrm{T}} = (\boldsymbol{c}_N + \boldsymbol{p}_N)^{\mathrm{T}} - (\boldsymbol{c}_B + \boldsymbol{p}_B)^{\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{N}.$$

Hence, $\bar{\boldsymbol{c}}_N \geq \boldsymbol{0}^{n-m}$ is sufficient for $\boldsymbol{x}^*$ to be an optimal solution to the perturbed problem (10.15). (Observe, however, that this is not a necessary condition unless $\boldsymbol{x}^*$ is non-degenerate.)

Perturbations of a non-basic cost coefficient

If only one component of $\boldsymbol{c}_N$ is perturbed, that is,

$$\boldsymbol{p} = \begin{pmatrix} \boldsymbol{p}_B \\ \boldsymbol{p}_N \end{pmatrix} = \begin{pmatrix} \boldsymbol{0}^m \\ \varepsilon \boldsymbol{e}_j \end{pmatrix},$$

for some $\varepsilon \in \mathbb{R}$ and $j \in \{1, \dots, n-m\}$, then we have that $\boldsymbol{x}^*$ is an optimal solution to the perturbed problem if

$$(\boldsymbol{c}_N)_j + \varepsilon - \boldsymbol{c}_B^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{N}_j \geq 0 \quad \iff \quad \varepsilon + (\tilde{\boldsymbol{c}}_N)_j \geq 0,$$

so in this case we only have to check that the perturbation ε is not less than $-(\tilde{\boldsymbol{c}}_N)_j$ in order to guarantee that $\boldsymbol{x}^*$ is an optimal solution to the perturbed problem.

Perturbations of a basic cost coefficient

If only one component of $\boldsymbol{c}_B$ is perturbed, that is,

$$\boldsymbol{p} = \begin{pmatrix} \boldsymbol{p}_B \\ \boldsymbol{p}_N \end{pmatrix} = \begin{pmatrix} \varepsilon \boldsymbol{e}_j \\ \boldsymbol{0}^{n-m} \end{pmatrix},$$

for some $\varepsilon \in \mathbb{R}$ and $j \in \{1, \dots, m\}$, then we have that $\boldsymbol{x}^*$ is an optimal solution to the perturbed problem if

$$(\boldsymbol{c}_N)^{\mathrm{T}} - (\boldsymbol{c}_B^{\mathrm{T}} + \varepsilon \boldsymbol{e}_j^{\mathrm{T}}) \boldsymbol{B}^{-1} \boldsymbol{N} \geq (\boldsymbol{0}^{n-m})^{\mathrm{T}} \quad \iff \quad \varepsilon \boldsymbol{e}_j^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{N} + \tilde{\boldsymbol{c}}_N \geq \boldsymbol{0}^{n-m}.$$

In this case all of the reduced costs of the non-basic variables may change, and we must check that the perturbation ε multiplied by the j^{th} row of $\boldsymbol{B}^{-1}\boldsymbol{N}$ plus the original reduced costs $\tilde{\boldsymbol{c}}_N$ is a vector whose components all are greater than or equal to zero.

Perturbations that makes $\boldsymbol{x}^*$ non-optimal

If the perturbation $\boldsymbol{p}$ is such that some of the reduced costs of the perturbed problem becomes strictly negative for the basis $\boldsymbol{x}_B$, then $\boldsymbol{x}^*$ is perhaps not an optimal solution anymore. If this happens, let some of the variables with strictly negative reduced cost enter the basis and continue the Simplex algorithm until an optimal solution is found.

10.5.2 Perturbations in the right-hand side coefficients

Now, assume that the right-hand side $\boldsymbol{b}$ of the linear program (10.14) is perturbed by the vector $\boldsymbol{p} \in \mathbb{R}^m$, that is, we consider the perturbed problem to

$$\begin{aligned} \text{minimize} \quad & \tilde{z} = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} + \boldsymbol{p}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n. \end{aligned} \tag{10.16}$$

The original reduced costs do not change as the right-hand side is perturbed, so the basic feasible solution given by the partition $\boldsymbol{A} = (\boldsymbol{B}, \boldsymbol{N})$ is optimal to the perturbed problem (10.16) if and only if it is feasible, that is,

$$\begin{pmatrix} \boldsymbol{x}_B \\ \boldsymbol{x}_N \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}^{-1}(\boldsymbol{b} + \boldsymbol{p}) \\ \boldsymbol{0}^{n-m} \end{pmatrix} \geq \boldsymbol{0}^n,$$

which means that we have to check that $\boldsymbol{B}^{-1}(\boldsymbol{b} + \boldsymbol{p}) \geq \boldsymbol{0}^m$.

Perturbations of one component of the right-hand side

Suppose that only one of the components of the right-hand side is perturbed, that is,

$$\boldsymbol{p} = \varepsilon \boldsymbol{e}_j,$$

for some $\varepsilon \in \mathbb{R}$ and $j \in \{1, \dots, m\}$. The basic feasible solution corresponding to the partition $\boldsymbol{A} = (\boldsymbol{B}, \boldsymbol{N})$ is then feasible if and only if

$$\boldsymbol{B}^{-1}(\boldsymbol{b} + \varepsilon \boldsymbol{e}_j) \geq \boldsymbol{0}^m \quad \Longleftrightarrow \quad \varepsilon \boldsymbol{B}^{-1}\boldsymbol{e}_j + \boldsymbol{B}^{-1}\boldsymbol{b} \geq \boldsymbol{0}^m,$$

so it must hold that ε multiplied by the j^{th} column of $\boldsymbol{B}^{-1}$ plus the vector $\boldsymbol{B}^{-1}\boldsymbol{b}$ equals a vector whose components all are greater than or equal to zero.

Perturbations that makes $\boldsymbol{x}^*$ infeasible

If the perturbation $\boldsymbol{p}$ is such that the basis $\boldsymbol{x}_B$ becomes infeasible, then some of the components in the updated right-hand side, $\boldsymbol{B}^{-1}(\boldsymbol{b} + \boldsymbol{p})$, is strictly negative. However, the reduced costs are independent of $\boldsymbol{p}$, so the basis $\boldsymbol{x}_B$ is still a *dual* feasible basis. Hence, we can continue with the Dual Simplex algorithm until an optimal solution is found (or until the primal infeasibility criterion is satisfied).

10.6 Notes and further reading

10.7 Exercises

Exercise 10.1 (constructing the LP dual) Consider the linear program

$$\begin{array}{llrrrrl}
\text{maximize} & z = & 6x_1 & -3x_2 & -2x_3 & +5x_4 & \\
\text{subject to} & & 4x_1 & +3x_2 & -8x_3 & +7x_4 & = 11, \\
& & 3x_1 & +2x_2 & +7x_3 & +6x_4 & \geq 23, \\
& & 7x_1 & +4x_2 & +3x_3 & +2x_4 & \leq 12, \\
& & x_1, & x_2 & & & \geq 0, \\
& & & & x_3 & & \leq 0, \\
& & & & & x_4 & \text{free.}
\end{array}$$

Construct the linear programming dual. ∎

Exercise 10.2 (constructing the LP dual) Consider the linear program

$$\begin{array}{ll}
\text{minimize} & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\
\text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\
& \boldsymbol{l} \leq \boldsymbol{x} \leq \boldsymbol{u}.
\end{array}$$

(a) Construct the linear programming dual.

(b) Show that the dual problem is always feasible (independent of $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{l}$, and $\boldsymbol{u}$). ∎

Exercise 10.3 (constructing an optimal dual solution from an optimal BFS) Consider the linear program in standard form

$$\begin{array}{ll}
\text{minimize} & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\
\text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\
& \boldsymbol{x} \geq \boldsymbol{0}^n.
\end{array} \tag{P}$$

Assume that an optimal BFS, $\boldsymbol{x}^* = (\boldsymbol{x}_B^{\mathrm{T}}, \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}}$, is given by the partition $\boldsymbol{A} = (\boldsymbol{B}, \boldsymbol{N})$. Show that

$$\boldsymbol{y} = (\boldsymbol{B}^{-1})^{\mathrm{T}}\boldsymbol{c}_B$$

is an optimal solution to the LP dual problem. ∎

Exercise 10.4 (application of the Weak and Strong Duality Theorems) Consider the linear program

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \qquad\qquad (\mathrm{P}) \\ \text{subject to} \quad & \boldsymbol{Ax} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n, \end{aligned}$$

and the perturbed problem to

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \qquad\qquad (\mathrm{P'}) \\ \text{subject to} \quad & \boldsymbol{Ax} = \tilde{\boldsymbol{b}}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n. \end{aligned}$$

Show that if (P) has an optimal solution, then the perturbed problem (P') cannot be unbounded (independent of $\tilde{\boldsymbol{b}}$). ■

Exercise 10.5 (application of the Weak and Strong Duality Theorems) Consider the linear program

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \qquad\qquad (\mathrm{P}) \\ \text{subject to} \quad & \boldsymbol{Ax} \leq \boldsymbol{b}. \end{aligned}$$

Assume that the objective function vector $\boldsymbol{c}$ cannot be written as a linear combination of the rows of $\boldsymbol{A}$. Show that (P) cannot have an optimal solution. ■

Exercise 10.6 (application of the Weak and Strong Duality Theorems) Consider the linear program

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \qquad\qquad (\mathrm{P}) \\ \text{subject to} \quad & \boldsymbol{Ax} \geq \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n. \end{aligned}$$

Construct a polyhedron that equals the set of optimal solutions to (P). ■

Exercise 10.7 (application of the Weak and Strong Duality Theorems) Consider the linear program

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \qquad\qquad (\mathrm{P}) \\ \text{subject to} \quad & \boldsymbol{Ax} \leq \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n. \end{aligned}$$

Let $\boldsymbol{x}^*$ be an optimal solution to (P) with the optimal objective function value z^*, and let $\boldsymbol{y}^*$ be an optimal solution to the LP dual of (P). Show that

$$z^* = (\boldsymbol{y}^*)^{\mathrm{T}} \boldsymbol{A}\boldsymbol{x}^*.$$

■

Exercise 10.8 (linear programming primal-dual optimality conditions) Consider the linear program

$$\begin{array}{llrrrrrl} \text{maximize} & z = & & -4x_2 & +3x_3 & +2x_4 & -8x_5 & \\ \text{subject to} & & 3x_1 & +x_2 & +2x_3 & +x_4 & & = 3, \\ & & x_1 & -x_2 & & +x_4 & -x_5 & \geq 2, \\ & & x_1, & x_2, & x_3, & x_4, & x_5 & \geq 0. \end{array}$$

Find an optimal solution by using the LP primal-dual optimality conditions. ■

Exercise 10.9 (linear programming primal-dual optimality conditions) Consider the linear program (the continuous knapsack problem)

$$\begin{aligned} \text{maximize} \quad & z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} && \text{(P)} \\ \text{subject to} \quad & \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} \leq b, \\ & \boldsymbol{x} \leq \mathbf{1}^n, \\ & \boldsymbol{x} \geq \mathbf{0}^n, \end{aligned}$$

where $\boldsymbol{c} > \mathbf{0}^n$, $\boldsymbol{a} > \mathbf{0}^n$, $b > 0$, and

$$\frac{c_1}{a_1} \geq \frac{c_2}{a_2} \geq \cdots \geq \frac{c_n}{a_n}.$$

Show that the feasible solution $\boldsymbol{x}$ given by

$$x_j = 1, \; j = 1, \dots, r-1, \quad x_r = \frac{b - \sum_{j=1}^{r-1} a_j}{a_r}, \quad x_j = 0, \; j = r+1, \dots, n,$$

where r is such that $\sum_{j=1}^{r-1} a_j < b$ and $\sum_{j=1}^{r} a_j > b$, is an optimal solution. ■

Exercise 10.10 (characterizations of optimal solutions in linear programming) Assume that $\boldsymbol{x}$ is feasible to (10.8) and $\boldsymbol{y}$ is feasible to (10.9). Show that the following are equivalent:

(1) $\boldsymbol{x}$ is an optimal solution to (10.8) and $\boldsymbol{y}$ is an optimal solution to (10.9);

(2) $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}$; and

(3) the complementary slackness conditions (10.10) hold. ■

Exercise 10.11 (KKT versus LP primal-dual optimality conditions) Consider the linear program

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \end{aligned} \tag{P}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{c} \in \mathbb{R}^n$, and $\boldsymbol{b} \in \mathbb{R}^m$. Show that the KKT conditions are equivalent to the LP primal-dual optimality conditions. ■

Exercise 10.12 (Lagrangian primal-dual versus LP primal-dual) Consider the linear program

$$\begin{aligned} \text{minimize} \quad & z = \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}. \end{aligned}$$

Show that the Lagrangian primal-dual optimality conditions are equivalent to the LP primal-dual optimality conditions. ■

Exercise 10.13 (the Dual Simplex Method) Show that by adding the constraint

$$x_1 + \cdots + x_n \leq M,$$

where M is very large positive number, to a linear program in standard form, it is always possible to construct a dual feasible basis. ■

Exercise 10.14 (sensitivity analysis: perturbations in the objective function) Consider the linear program

$$\begin{array}{lrrrrl} \text{maximize} \quad z = & -x_1 & +18x_2 & +c_3x_3 & +c_4x_4 & \\ \text{subject to} & x_1 & +2x_2 & +3x_3 & +4x_4 & \leq 3, \\ & -3x_1 & +4x_2 & -5x_3 & -6x_4 & \leq 1, \\ & x_1, & x_2, & x_3, & x_4 & \geq 0. \end{array}$$

Find the values of c_3 and c_4 such that the basic solution that corresponds to the partition $\boldsymbol{x}_B = (x_1, x_2)^{\mathrm{T}}$ is an optimal basic feasible solution to the problem. ■

Exercise 10.15 (sensitivity analysis: perturbations in the right-hand side) Consider the linear program

$$\begin{array}{lrrrl} \text{minimize} \quad z = & -x_1 & +2x_2 & +x_3 & \\ \text{subject to} & 2x_1 & +x_2 & -x_3 & \leq 7, \\ & -x_1 & +2x_2 & +3x_3 & \geq 3 + \delta, \\ & x_1, & x_2, & x_3 & \geq 0. \end{array}$$

(a) Show that the basic solution that corresponds to the partition $\boldsymbol{x}_B = (x_1, x_3)^{\mathrm{T}}$ is an optimal solution to the problem when $\delta = 0$.

(b) Find the values of the perturbation $\delta \in \mathbb{R}$ such that the above BFS is optimal.

(c) Find an optimal solution when $\delta = -7$. ■

Part V

Optimization over Convex Sets

Unconstrained optimization

XI

11.1 Introduction

We consider throughout this chapter the unconstrained optimization problem to

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\operatorname{minimize}} \; f(\boldsymbol{x}), \tag{11.1}$$

where $f \in C^0$ on $\mathbb{R}^n$ (f is continuous). Mostly, we will assume that $f \in C^1$ holds (f is continuously differentiable), in some cases even $f \in C^2$.

What are the methods of choice for this problem? It depends on many factors:

- what is the size of the problem (n)?
- are $\nabla f(\boldsymbol{x})$ and/or $\nabla^2 f(\boldsymbol{x})$ available, and if so to what cost?
- what it is the solution requirement? (Do we need a global minimum or a local minimum or simply a stationary point?)
- What are the convexity properties of f?
- Do we have a good estimate of the location of a stationary point $\boldsymbol{x}^*$? (Can we use locally-only convergent methods?)

We will discuss some basic approaches to the problem (11.1) and refer to questions such as the ones just mentioned during the development.

Example 11.1 (non-linear least squares data fitting) Suppose that we have m data points (t_i, b_i) which we believe are related through an algebraic expression of the form

$$x_1 + x_2 \exp(x_3 t_i) + x_4 \exp(x_5 t_i) = b_i, \qquad i = 1, \dots, m,$$

where however the parameters $x_1, \ldots, x_5$ are unknown. (Here, $\exp(x) = e^x$.) In order to best describe the above model, we minimize the total "residual error" given by the norm of the residual

$$f_i(\boldsymbol{x}) := b_i - [x_1 + x_2 \exp(x_3 t_i) + x_4 \exp(x_5 t_i)], \qquad i = 1, \ldots, m.$$

A minimization will then yield the best fit with respect to the data points available. The following then is the resulting optimization problem to be solved:

$$\underset{\boldsymbol{x} \in \mathbb{R}^5}{\text{minimize}}\ f(\boldsymbol{x}) := \sum_{i=1}^{m} |f_i(\boldsymbol{x})|^2 = \sum_{i=1}^{m} [f_i(\boldsymbol{x})]^2.$$

This type of problem is very often solved within numerical analysis and mathematical statistics. Note that the 2-norm is not the only measure of the residual used; some times the maximum norm is used. ■

What is the typical form of an algorithm in unconstrained optimization (in fact, for almost every problem class)? Take a look at Figure 11.1 which depicts the level curves of a convex, quadratic function, the below description, and the flow chart in Figure 11.2 of a complete iteration.

Descent algorithm:

Step 0 (initialization). Determine a *starting point* $\boldsymbol{x}_0 \in \mathbb{R}^n$. Set $k := 0$.

Step 1 (descent direction generation). Determine a *search direction* $\boldsymbol{p}_k \in \mathbb{R}^n$.

Step 2 (line search). Determine a *step length* $\alpha_k > 0$ such that $f(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k) < f(\boldsymbol{x}_k)$ holds.

Step 3 (update). Update: let $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$.

Step 4 (termination check). If a *termination criterion* is fulfilled, then stop! Otherwise, let $k := k + 1$ and go to step 1.

This type of algorithm is inherently local, since we cannot in general use more than the information that can be calculated at the current point $\boldsymbol{x}_k$, that is, $f(\boldsymbol{x}_k)$, $\nabla f(\boldsymbol{x}_k)$, and $\nabla^2 f(\boldsymbol{x}_k)$. As far as our local "sight" is concerned, we sometimes call this type of method (for *maximization* problems) the "near-sighted mountain climber," reflecting the distinct possibility that the mountain climber is in a deep fog and can only check her barometer for the height and feel the steepness of the slope under her feet. Notice then that Figure 11.1 was plotted using several thousands of function evaluations; in reality—and definitely in higher dimension than two—this type of orienteering map *never* exists when we want to solve a problem.

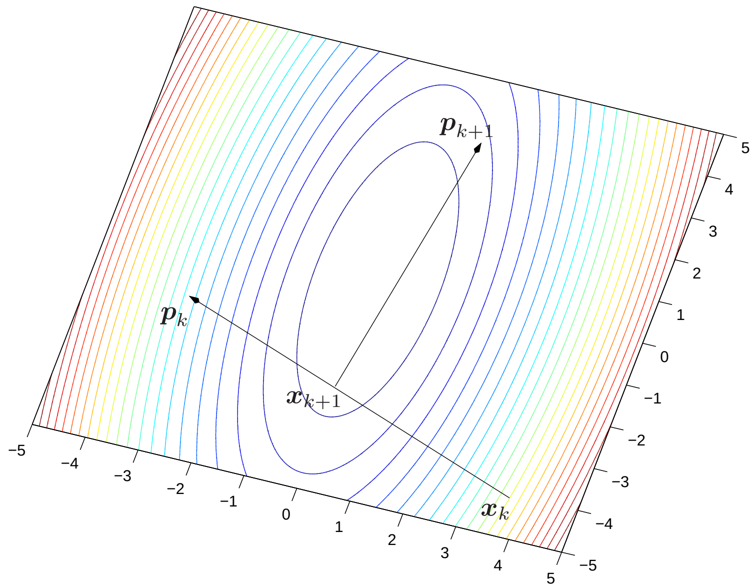


Figure 11.1: At $\boldsymbol{x}_k$, the search direction $\boldsymbol{p}_k$ is generated. A step α_k is taken in this direction, producing $\boldsymbol{x}_{k+1}$. At this point, a new search direction $\boldsymbol{p}_{k+1}$ is generated, and so on.

We begin by analyzing Step 1, the most important step of the above-described algorithm. Based on the result in Proposition 4.15 it makes good sense to generate $\boldsymbol{p}_k$ such that it is a direction of descent.

11.2 Descent directions

11.2.1 Basic ideas

The definition of a direction of descent at a given point is given by Definition 4.14. Usually, we have many choices for directions of descent, see for example Proposition 4.15 for a sufficient criterion in case we deal with a continuously differentiable function. In this section we discuss some details on how descent directions should be generated, depending on a particular situation.

Example 11.2 (example descent directions) (a) Let $f \in C^1(N)$ in some neighborhood N of $\boldsymbol{x}_k \in \mathbb{R}^n$. If $\nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0}^n$, then $\boldsymbol{p} = -\nabla f(\boldsymbol{x}_k)$ is a

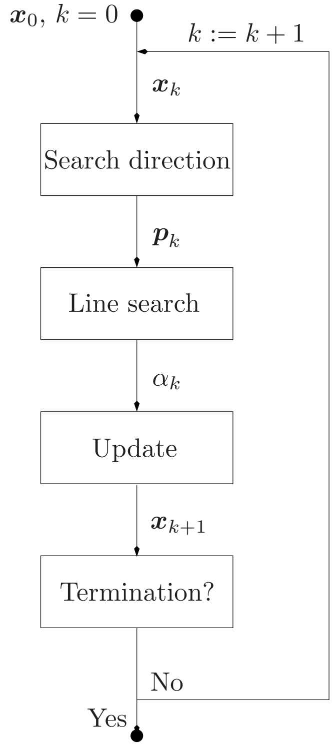


Figure 11.2: Flow chart of an iteration of the general algorithm.

descent direction for f at $\boldsymbol{x}_k$ (this follows directly from Proposition 4.15). This is exactly the search direction used in the steepest descent method, and it naturally bears the name of *steepest descent direction* because it solves the minimization problem to[1]

$$\underset{\boldsymbol{p} \in \mathbb{R}^n : \|\boldsymbol{p}\|=1}{\text{minimize}} \; \nabla f(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{p}. \tag{11.2}$$

(b) Let $f \in C^2(N)$ in some neighborhood N of $\boldsymbol{x}_k$. If $\nabla f(\boldsymbol{x}_k) = \mathbf{0}^n$ we cannot use the steepest descent direction anymore. However, we can work with second order information provided by the Hessian to find a descent direction in this case also, provided that f is non-convex at $\mathbf{0}^n$.

[1] We have that $\nabla f(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{p} = \|\nabla f(\boldsymbol{x})\| \cdot \|\boldsymbol{p}\| \cos\theta$, where θ is the angle between the vectors $\nabla f(\boldsymbol{x})$ and $\boldsymbol{p}$; this expression is clearly minimized by making $\cos\theta = -1$, that is, by letting $\boldsymbol{p}$ have the angle $180°$ with $-\nabla f(\boldsymbol{x})$; in other words, $\boldsymbol{p} = -\nabla f(\boldsymbol{x})/\|\nabla f(\boldsymbol{x})\|$.

Assume that $\nabla^2 f(\boldsymbol{x}_k)$ is not positive semidefinite (otherwise, $\boldsymbol{x}_k$ is likely to be at the locally optimal solution, see Theorem 4.16). If $\nabla^2 f(\boldsymbol{x}_k)$ is indefinite we call the stationary point $\boldsymbol{x}_k$ a *saddle point* of f. Let $\boldsymbol{p}$ be an eigenvector corresponding to a negative eigenvalue of $\nabla^2 f(\boldsymbol{x}_k)$. Then, we call $\boldsymbol{p}$ a *direction of negative curvature* for f at $\boldsymbol{x}_k$, and it can be demonstrated that it is a descent direction for f at this point [the proof is similar to the one of Proposition 4.15, but uses Taylor expansion (2.5) instead of (2.4)].

(c) Assume the conditions of (a), and let $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ be an arbitrary symmetric, positive definite matrix. Then $\boldsymbol{p} = -\boldsymbol{Q}\nabla f(\boldsymbol{x}_k)$ is a descent direction for f at $\boldsymbol{x}_k$: $\nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{p} = -\nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{Q}\nabla f(\boldsymbol{x}_k) < 0$, due to the positive definiteness of $\boldsymbol{Q}$. (This is of course true only if $\boldsymbol{x}_k$ is nonstationary, as assumed.)

Pre-multiplying by $\boldsymbol{Q}$ may be interpreted as a scaling of ∇f if we choose a diagonal matrix $\boldsymbol{Q}$; the use of more general matrices is of course possible and leads to exceptionally good computational results for clever choices of $\boldsymbol{Q}$. Newton and quasi-Newton methods are based on constructing directions in this way. Note that setting $\boldsymbol{Q} = \boldsymbol{I}^n$ (the identity matrix in $\mathbb{R}^{n\times n}$), we obtain the steepest descent direction. ■

To find some arbitrary direction of descent is not a very difficult task as demonstrated by Example 11.2 [in fact, the situation when $\nabla f(\boldsymbol{x}_k) = \boldsymbol{0}^n$ appearing in (b) is quite an exotic one already, so typically one can always use directions constructed in (a), or, more generally (c), as descent directions]. However, in order to secure the convergence of numerical algorithms we must provide descent directions that "behave well" numerically. Typical requirements, additional to the basic requirement of being a direction of descent, are:

$$|\nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{p}_k| \geq s_1\|\nabla f(\boldsymbol{x}_k)\|^2, \quad \text{and} \quad \|\boldsymbol{p}_k\| \leq s_2\|\nabla f(\boldsymbol{x}_k)\|, \tag{11.3}$$

or

$$-\frac{\nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{p}_k}{\|\nabla f(\boldsymbol{x}_k)\| \cdot \|\boldsymbol{p}_k\|} \geq s_1, \quad \text{and} \quad \|\boldsymbol{p}_k\| \geq s_2\|\nabla f(\boldsymbol{x}_k)\|, \tag{11.4}$$

where $s_1, s_2 > 0$, and $\boldsymbol{x}_k$ and $\boldsymbol{p}_k$ are, respectively, iterates and search directions of some iterative algorithm. (In the next section, we shall provide the basic form of an iterative algorithm.)

The purpose of these condition is to prevent the descent directions to deteriorate in quality in terms of always providing good enough descent. For example, the first condition in (11.3) states that if the directional derivative of f tends to zero then it must be that the gradient of f also tends to zero, while the second condition makes sure that a bad direction in terms of the directional derivative is not compensated by the search direction becoming extremely long in norm. The first condition

$$-\frac{\nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{p}_k}{\|\nabla f(\boldsymbol{x}_k)\| \cdot \|\boldsymbol{p}_k\|} \geq s_1 \tag{11.5}$$

in (11.4) is equivalent to the requirement that the cosine of the angle between $-\nabla f(\boldsymbol{x}_k)$ and $\boldsymbol{p}_k$ is positive and bounded away from zero by the value of s_1, that is, the angle must be acute and not too close to $\pi/2$; this is another way of saying that the direction $\boldsymbol{p}_k$ must be steep enough. The purpose of the second condition in (11.4) then is to ensure that if the search direction vanishes then so does the gradient. Methods satisfying (11.3), (11.4) are some times referred to as *gradient related*, since they cannot be based on search directions that are very far from those of the steepest descent method.

The choice

$$\boldsymbol{p}_k = -\nabla f(\boldsymbol{x}_k)$$

fulfills the above conditions, with $s_1 = s_2 = 1$.

Another example is as follows: set $\boldsymbol{p}_k = -\boldsymbol{Q}_k \nabla f(\boldsymbol{x}_k)$, where $\boldsymbol{Q}_k \in \mathbb{R}^{n \times n}$ is a symmetric and *positive definite* matrix such that $m\|\boldsymbol{s}\|^2 \leq \boldsymbol{s}^{\mathrm{T}}\boldsymbol{Q}_k\boldsymbol{s} \leq M\|\boldsymbol{s}\|^2$, for all $\boldsymbol{s} \in \mathbb{R}^n$, holds. [All eigenvalues of $\boldsymbol{Q}_k$ lie in the interval $[m, M] \subset (0, \infty)$.] Then, the requirement (11.3) is verified with $s_1 = m$, $s_2 = M$, and (11.4) holds with $s_1 = m/M$, $s_2 = m$.

11.2.2 Less basic ideas

What should a good descent direction do? Roughly speaking, it should provide as large descent as possible, that is, minimize $f(\boldsymbol{x} + \boldsymbol{p}) - f(\boldsymbol{x})$ over some large enough region of $\boldsymbol{p}$ around the origin. In principle, this is the idea behind the optimization problem (11.2), because, according to (2.1), $f(\boldsymbol{x} + \boldsymbol{p}) - f(\boldsymbol{x}) \approx \nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{p}$.

Therefore, more insights into how the scaling matrices $\boldsymbol{Q}$ appearing in Example 11.2(c) should be constructed and, in particular, reasons why the steepest descent direction is not a very wise choice, can be gained if we consider more general approximations than the ones given by (2.1). Namely, assume that f is C^1 near $\boldsymbol{x}$, and that for some positive definite matrix $\boldsymbol{Q}$ it holds that

$$f(\boldsymbol{x} + \boldsymbol{p}) - f(\boldsymbol{x}) \approx \varphi_{\boldsymbol{x}}(\boldsymbol{p}) = \nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^{\mathrm{T}}\boldsymbol{Q}^{-1}\boldsymbol{p}. \tag{11.6}$$

For example, if $f \in C^2$, $\nabla^2 f(\boldsymbol{x}) \succ \boldsymbol{0}^{n \times n}$, and assuming $o(\|\boldsymbol{p}\|^2) \approx 0$ [cf. (2.3)] we may use $\boldsymbol{Q}^{-1} = \nabla^2 f(\boldsymbol{x})$.

Using the optimality conditions, we can easily check that the search direction defined in Example 11.2(c) is a solution to the following optimization problem:

$$\underset{\boldsymbol{p}\in\mathbb{R}^n}{\text{minimize}}\ \varphi_{\boldsymbol{x}}(\boldsymbol{p}), \tag{11.7}$$

where $\varphi_{\boldsymbol{x}}(\boldsymbol{p})$ is defined by (11.6). The closer $\varphi_{\boldsymbol{x}}(\boldsymbol{p})$ approximates $f(\boldsymbol{x}+\boldsymbol{p}) - f(\boldsymbol{x})$, the better the quality of search directions generated by Example 11.2(c) we can expect.

As was already mentioned, setting $\boldsymbol{Q} = \boldsymbol{I}^n$, which absolutely fails to take into account any information about f (that is, it is a "one-size-fits-all" approximation), gives us the steepest descent direction. (Cases can easily be constructed such that the algorithm converges extremely slowly; convergence can actually be so bad that the authors of the book [BGLS03] decree that the steepest descent method should be *forbidden*!) On the other hand, the "best" second-order approximation is given by the Taylor expansion (2.3), and therefore we would like to set $\boldsymbol{Q} = [\nabla^2 f(\boldsymbol{x})]^{-1}$; this is exactly the choice made in the Newton method.

Remark 11.3 (a motivation for the descent property in Newton's method) Recall that the search direction in Newton's method is based on the solution of the following linear system of equations: find $\boldsymbol{p} \in \mathbb{R}^n$ such that

$$\nabla_{\boldsymbol{p}}\varphi_{\boldsymbol{x}}(\boldsymbol{p}) = \nabla f(\boldsymbol{x}) + \nabla^2 f(\boldsymbol{x})\boldsymbol{p} = \boldsymbol{0}^n.$$

Consider the case of $n = 1$. We should then solve

$$f'(x) + f''(x)p = 0. \tag{11.8}$$

It is obvious that unless $f'(x) = 0$ (whence we are at a stationary point and $p = 0$ solves the equation) we cannot solve (11.8) unless $f''(x) \neq 0$. Then, the solution $\bar{p} := -f'(x)/f''(x)$ to (11.8) is well-defined. We distinguish between two cases:

(a) $f''(x) > 0$. In this case, the derivative of the second-order approximation $p \mapsto f'(x)p + \frac{1}{2}f''(x)p^2$ has a positive slope. Hence, if $f'(x) > 0$ then $\bar{p} < 0$, and if $f'(x) < 0$ then instead $\bar{p} > 0$ holds. In both cases, therefore, the directional derivative $f'(x)\bar{p} < 0$, that is, $\bar{p}$ is a descent direction.

(b) $f''(x) < 0$. In this case, the derivative of the second-order approximation $p \mapsto f'(x)p + \frac{1}{2}f''(x)p^2$ has a negative slope. Hence, if $f'(x) > 0$ then $\bar{p} > 0$, and if $f'(x) < 0$ then instead $\bar{p} < 0$ holds. In both cases, therefore, the directional derivative $f'(x)\bar{p} > 0$, that is, $\bar{p}$ is an ascent direction.

From this derivation it becomes clear that Newton's method (for $n = 1$ it is often referred to as the *Newton–Raphson method*) (cf. Section 4.6.4.2) provides the same search direction regardless of whether the optimization problem is a minimization or a maximization problem; the reason is that the search direction is based on the stationarity of the second-order approximation and not its minimization/maximization. We also see that the Newton direction $\bar{p}$ is a descent direction if the function f is of the strictly convex type around x [that is, if $f''(x) > 0$], and an ascent direction if it is of the strictly concave type around x [that is, if $f''(x) < 0$]. In other words, if the objective function is (strictly) convex or concave, the Newton equation will give us the right direction, if it gives us a direction at all. Translated to the case of $n > 1$, Newton's method acts as a descent method if the Hessian matrix $\nabla^2 f(\boldsymbol{x})$ is positive definite, and as an ascent method if it is negative definite, which is appropriate. ∎

An essential problem arises of course if the above-described is *not* what we want; for example, it may be that we are interested in maximizing a function which is neither convex or concave, and around a current point the function is of strictly convex type (that is, the Hessian is positive definite). In this case the Newton direction will not point in an ascent direction, but instead the opposite. How to solve a problem with a Newton-type method in a non-convex world is the main topic of what follows. As always, we consider minimization to be the direction of interest for f.

So, why might one want to choose a matrix $\boldsymbol{Q}$ differing from the "best" choice $[\nabla^2 f(\boldsymbol{x})]^{-1}$? There are several reasons:

Lack of positive definiteness The matrix $\nabla^2 f(\boldsymbol{x})$ might not be positive definite. As a result, the problem (11.7) may even lack optimal solutions and $-[\nabla^2 f(\boldsymbol{x})]^{-1}\nabla f(\boldsymbol{x})$ might in any case not be a direction of descent.

This problem can be cured by adding to $\nabla^2 f(\boldsymbol{x})$ a diagonal matrix $\boldsymbol{E}$, so that $\nabla^2 f(\boldsymbol{x}) + \boldsymbol{E}$ is positive definite. For example, $\boldsymbol{E} = \gamma \boldsymbol{I}^n$, for $-\gamma$ smaller than all the non-positive eigenvalues of $\nabla^2 f(\boldsymbol{x})$, may be used because such a modification "shifts" the original eigenvalues of $\nabla^2 f(\boldsymbol{x})$ by $\gamma > 0$. The value of γ needed will automatically be found when solving the "Newton equation" $\nabla^2 f(\boldsymbol{x})\boldsymbol{p} = -\nabla f(\boldsymbol{x})$, since eigenvalues of $\nabla^2 f(\boldsymbol{x})$ are pivot elements in Gaussian-elimination procedures. This modification bears the name *Levenberg–Marquardt*.

[*Note:* as γ becomes large, $\boldsymbol{p}$ resembles more and more the steepest descent direction.]

Lack of enough differentiability The function f might not be twice differentiable, or the matrix of second derivatives might be too costly to compute/evaluate.

Either being the case, quasi-Newton methods approximate the Newton equation by replacing $\nabla^2 f(\boldsymbol{x}_k)$ with a matrix $\boldsymbol{B}_k$ that is cheaper to compute, typically by only using values of ∇f at the current and some previous points.

Using a first-order Taylor expansion (2.1) for $\nabla f(\boldsymbol{x}_k)$ we know that

$$\nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) \approx \nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_{k-1}),$$

so the matrix $\boldsymbol{B}_k$ is taken to satisfy the similar system

$$\boldsymbol{B}_k(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) = \nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_{k-1}).$$

[*Note:* For $n = 1$, this corresponds to the *secant method*, in which at iteration k we approximate the second derivative as

$$f''(x_k) \approx \frac{f'(x_k) - f'(x_{k-1})}{x_k - x_{k-1}}$$

in Newton's method.]

However, the matrix $\boldsymbol{B}_k$ (that is, n^2 unknowns) is under-determined in these n equations, so additional requirements, such as ones that make sure that $\boldsymbol{B}_k$ is symmetric and positive definite, result in particular quasi-Newton methods. Typically, starting from $\boldsymbol{B}_0 = \boldsymbol{I}^n$, $\boldsymbol{B}_{k+1}$ is calculated from $\boldsymbol{B}_k$ using a rank-one or rank-two update; in particular, this allows us to update the factorization of $\boldsymbol{B}_k$ to efficiently obtain the factorization of $\boldsymbol{B}_{k+1}$ using standard algorithms in linear algebra.

There are infinitely many choices that may be used, and the following (called the Broyden–Fletcher–Goldfarb–Shanno, or BFGS, method after the original publications [Bro70, Fle70, Gol70, Sha70]) is considered to be the most effective:

$$\boldsymbol{B}_{k+1} = \boldsymbol{B}_k - \frac{(\boldsymbol{B}_k\boldsymbol{s}_k)(\boldsymbol{B}_k\boldsymbol{s}_k)^{\mathrm{T}}}{\boldsymbol{s}_k^{\mathrm{T}}\boldsymbol{B}_k\boldsymbol{s}_k} + \frac{\boldsymbol{y}_k\boldsymbol{y}_k^{\mathrm{T}}}{\boldsymbol{y}_k^{\mathrm{T}}\boldsymbol{s}_k},$$

where $\boldsymbol{s}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$, and $\boldsymbol{y}_k = \nabla f(\boldsymbol{x}_{k+1}) - \nabla f(\boldsymbol{x}_k)$. Interestingly enough, should f be quadratic, $\boldsymbol{B}_k$ will be identical to the Hessian of f after a finite number of steps (namely, n).

Quasi-Newton methods with various update rules for $\boldsymbol{B}_k$ are very popular for unconstrained optimization.

See Section 11.9 for more details on quasi-Newton methods.

Computational burden The solution of a linear system $\boldsymbol{B}_k\boldsymbol{p}_k = -\nabla f(\boldsymbol{x}_k)$, or, which is the same if we identify $\boldsymbol{Q}^{-1} = \boldsymbol{B}_k$, finding the optimum of (11.7), may be too costly. This is exactly the situation when one would like to use the steepest descent method, which avoids any such calculations.

Other possibilities are: (a) In a quasi-Newton method, keep the matrix $\boldsymbol{B}_k$ (and, obviously, its factorization) fixed for $k_0 > 1$ subsequent steps; in this way, we need to perform matrix factorization (the most computationally consuming part) only every k_0 steps, k_0 being a small integer.

(b) Solve the optimization problem (11.7) only approximately; based on the following arguments. Assume that $\boldsymbol{x}_k$ violates the second order necessary optimality conditions for f, and consider the problem (11.7) where we replace the matrix $\boldsymbol{Q}^{-1}$ with an iteration-dependent, perhaps only positive semi-definite matrix $\boldsymbol{B}_k$. As a first example, suppose we consider the Newton method, whence we choose $\boldsymbol{B}_k = \nabla^2 f(\boldsymbol{x}_k)$. Then, by the assumption that the second order necessary optimality conditions are violated, $\boldsymbol{p} = \boldsymbol{0}^n$ is not a minimum of $\varphi_{\boldsymbol{x}_k}(\boldsymbol{p})$ in the problem (11.7). Let $\widetilde{\boldsymbol{p}} \neq \boldsymbol{0}^n$ be any vector with $\varphi_{\boldsymbol{x}_k}(\widetilde{\boldsymbol{p}}) < \varphi_{\boldsymbol{x}_k}(\boldsymbol{0}^n) = 0$. Then,

$$\varphi_{\boldsymbol{x}_k}(\widetilde{\boldsymbol{p}}) = \nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\widetilde{\boldsymbol{p}} + \underbrace{\frac{1}{2}\widetilde{\boldsymbol{p}}^{\mathrm{T}}\boldsymbol{B}_k\widetilde{\boldsymbol{p}}}_{\geq 0} < 0 = \varphi_{\boldsymbol{x}_k}(\boldsymbol{0}^n),$$

which implies that $\nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\widetilde{\boldsymbol{p}} < 0$. This means that if the Newton equations are solved inexactly, a descent direction is still obtained. This can of course be generalized for quasi-Newton methods as well, since we only assumed that the matrix $\boldsymbol{B}_k$ is positive semi-definite.

We summarize the above development of search directions in Table 11.1. At some iteration k the iterate is $\boldsymbol{x}_k$; for each algorithm, we describe the linear system solved in order to generate the search direction $\boldsymbol{p}_k$. In the table $\gamma_k > 0$ and $\boldsymbol{B}_k \in \mathbb{R}^{n\times n}$ is a symmetric and positive definite matrix.

Algorithm	*Linear system*
Steepest descent	$\boldsymbol{p}_k = -\nabla f(\boldsymbol{x}_k)$
Newton's method	$\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{p}_k = -\nabla f(\boldsymbol{x}_k)$
Levenberg–Marquardt	$[\nabla^2 f(\boldsymbol{x}_k) + \gamma_k \boldsymbol{I}^n]\boldsymbol{p}_k = -\nabla f(\boldsymbol{x}_k)$
Quasi-Newton	$\boldsymbol{B}_k\boldsymbol{p}_k = -\nabla f(\boldsymbol{x}_k)$

Table 11.1: Search directions.

11.3 Line searches

11.3.1 Introduction

Executing Step 2 in the iterative algorithm is naturally done by finding an approximate solution to the one-dimensional problem to

$$\underset{\alpha \geq 0}{\text{minimize}}\ \varphi(\alpha) := f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k). \tag{11.9}$$

Its optimality conditions are that[2]

$$\varphi'(\alpha^*) \geq 0, \qquad \alpha^* \cdot \varphi'(\alpha^*) = 0, \qquad \alpha^* \geq 0, \tag{11.11}$$

that is,

$$\nabla f(\boldsymbol{x}_k + \alpha^* \boldsymbol{p}_k)^{\mathrm{T}} \boldsymbol{p}_k \geq 0, \quad \alpha^* \cdot \nabla f(\boldsymbol{x}_k + \alpha^* \boldsymbol{p}_k)^{\mathrm{T}} \boldsymbol{p}_k = 0, \quad \alpha^* \geq 0,$$

holds. So, if $\alpha^* > 0$, then $\varphi'(\alpha^*) = 0$ must hold, which therefore means that $\nabla f(\boldsymbol{x}_k + \alpha^* \boldsymbol{p}_k)^{\mathrm{T}} \boldsymbol{p}_k = 0$, that is, that the search direction $\boldsymbol{p}_k$ is orthogonal to the gradient of f at the point $\boldsymbol{x}_k + \alpha^* \boldsymbol{p}_k$.

Figure 11.3 shows an example of the one-dimensional function φ along a descent direction with a well-defined minimum.

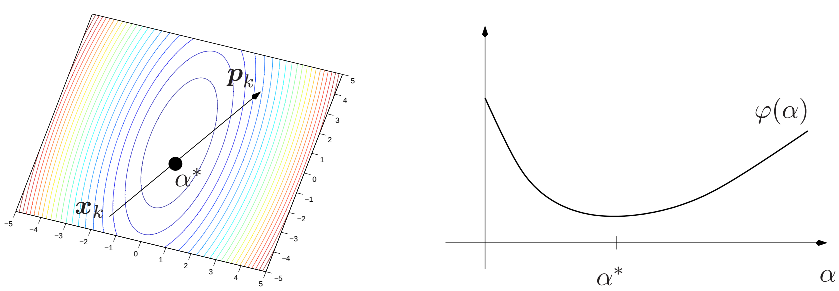


Figure 11.3: A line search in a descent direction.

---

[2]These conditions are the same as those in Proposition 4.22(b). To establish this fact, let's suppose first that we satisfy (4.10) which here becomes the statement that

$$\varphi'(\alpha^*)(\alpha - \alpha^*) \geq 0, \qquad \alpha \geq 0. \tag{11.10}$$

Setting first $\alpha = 0$ in (11.10), then $\alpha^* \cdot \varphi'(\alpha^*) \leq 0$ follows. On the other hand, setting $\alpha = 2\alpha^*$ in (11.10), then $\alpha^* \cdot \varphi'(\alpha^*) \geq 0$ follows. So, $\alpha^* \cdot \varphi'(\alpha^*) = 0$ must hold. Also, setting $\alpha = \alpha^* + 1$ in (11.10), we obtain that $\varphi'(\alpha^*) \geq 0$. This establishes that (11.11) follows from (4.10). To establish the reverse conclusion and therefore prove that the two conditions are the same, we note that if we satisfy (11.11), then it follows that for every $\alpha \geq 0$, $\varphi'(\alpha^*)(\alpha - \alpha^*) = \alpha\varphi'(\alpha^*) \geq 0$, and we are done.

In the quest for a stationary point it is of relatively minor importance to do a line search accurately—the stationary point is most probably not situated somewhere along that half-line anyway. Therefore, most line search strategies used in practice are approximate. It should also be noted that if the function f is non-convex then so is probably the case with φ as well, and globally minimizing a non-convex function is difficult even in one variable.

11.3.2 Approximate line search strategies

First, we consider the case where f is quadratic; this is the only general case where an accurate line search is practical.

Let $f(\boldsymbol{x}) = (1/2)\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{q}^{\mathrm{T}}\boldsymbol{x} + a$, where the dimensions of $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$, $\boldsymbol{q} \in \mathbb{R}^n$ and $a \in \mathbb{R}$ are given data. Suppose we wish to minimize the function φ for this special case. Then, we can solve for $\varphi'(\alpha^*) = 0$ analytically:

$$\begin{aligned}
\varphi'(\alpha) &= \nabla f(\boldsymbol{x} + \alpha\boldsymbol{p})^{\mathrm{T}}\boldsymbol{p} \\
&= [\boldsymbol{Q}(\boldsymbol{x} + \alpha\boldsymbol{p}) - \boldsymbol{q}]^{\mathrm{T}}\boldsymbol{p} \\
&= \alpha\boldsymbol{p}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{p} - (\boldsymbol{q} - \boldsymbol{Q}\boldsymbol{x})^{\mathrm{T}}\boldsymbol{p} \\
&= 0 \\
&\Leftrightarrow \\
\alpha &= (\boldsymbol{q} - \boldsymbol{Q}\boldsymbol{x})^{\mathrm{T}}\boldsymbol{p}/\boldsymbol{p}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{p}.
\end{aligned}$$

Let's check the validity and meaning of this solution. We suppose naturally that $\boldsymbol{p}$ is a descent direction, whence $\varphi'(0) = (\boldsymbol{q} - \boldsymbol{Q}\boldsymbol{x})^{\mathrm{T}}\boldsymbol{p} < 0$ holds. Therefore, if $\boldsymbol{Q}$ is positive definite, we are guaranteed that the value of α will be positive.

Among the classic approximate line searches we mention very briefly the following:

Interpolation Take $f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{p}_k$ to model a quadratic function approximating f along $\boldsymbol{p}_k$. Minimize it by using the analytic formula above.

Newton's method Repeat the improvements gained from a quadratic approximation: $\alpha := \alpha - \varphi'(\alpha)/\varphi''(\alpha)$.

Golden Section The golden section method is a derivative-free method for minimizing *unimodal* functions.[3] The method reduces an in-

[3] φ is *unimodal* in an interval $[a, b]$ of $\mathbb{R}$ if it has a unique global minimum in $[a, b]$, and is strictly increasing to the left as well as to the right of the minimum. This notion is equivalent to that of φ having a minimum over $[a, b]$ and being strictly quasi-convex there.

terval wherein the reduction is based only on evaluating φ. The portion left of the length of the previous interval after reduction is exactly the golden section, $\frac{\sqrt{5}-1}{2} \approx 0.618$.

An approximate line search methodology often used is known as the *Armijo* step length rule. The idea is to quickly generate a step length α which provides a "sufficient" decrease in the value of f. Note that $f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k) \approx f(\boldsymbol{x}_k) + \alpha \cdot \nabla f(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{p}_k$ for very small values of $\alpha > 0$. The requirement of the step length rule is that we get a decrease in the left-hand side of the above approximate relation which is at least a fraction of that predicted in the right-hand side.

Let $\mu \in (0, 1)$ be the fraction of decrease required. Then, the step lengths accepted by the Armijo step length rule are the positive values α which satisfy the inequality

$$\varphi(\alpha) - \varphi(0) \leq \mu\alpha\varphi'(0), \tag{11.12a}$$

that is,

$$f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k) - f(\boldsymbol{x}_k) \leq \mu\alpha\nabla f(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{p}_k. \tag{11.12b}$$

Figure 11.4 illustrates the Armijo step length rule.

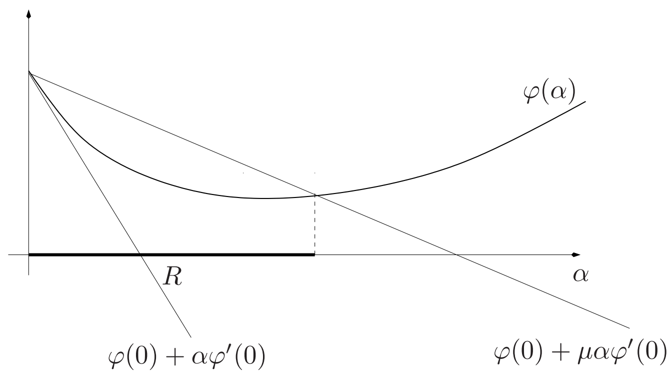


Figure 11.4: The interval, denoted R, accepted by the Armijo step length rule.

The typical choices are the following: choose μ small [$\mu \in (0.001, 0.01)$], and take $\alpha = 1$. If $\alpha = 1$ does not satisfy the inequality (11.12), then take $\alpha := \alpha/2$, and check the inequality (11.12) again. The choice of initial trial step $\alpha = 1$ is especially of interest in Newton-type methods, where, locally around a stationary point $\boldsymbol{x}^*$ where $\nabla^2 f(\boldsymbol{x}^*)$ is positive definite, local convergence with step length one is guaranteed. (See also Section 4.6.4.2.)

In theory, however, we can select any starting guess $\bar{\alpha} > 0$ and any fraction $\beta \in (0, 1)$ in place of the choice $\beta = \frac{1}{2}$ made above.

The Armijo condition is satisfied for any sufficiently small step length, provided that the direction $\boldsymbol{p}_k$ is a direction of descent. (See Exercise 11.1.) In itself it therefore does not guarantee that the next iterate is much better in terms of the objective value than the current one. Often, therefore, it is combined with a condition such that

$$|\varphi'(\alpha_k)| \leq \eta |\varphi'(0)|,$$

that is,

$$|\nabla f(x_k + \alpha p_k)^{\mathrm{T}} p_k| \leq \eta |\nabla f(x_k)^{\mathrm{T}} p_k|,$$

holds for some $\eta \in [0, 1)$. This is called the *Wolfe condition*. A relaxed condition, the *weak Wolfe condition*, of the form

$$\varphi'(\alpha_k) \geq \eta \varphi'(0)$$

is often preferred, since the former takes more computations to fulfill. The choice $0 < \mu < \eta < 1$ leads to interesting descent algorithms when the Armijo and weak Wolfe conditions are combined, and it is possible (Why?) to find positive step lengths that satisfy these two conditions provided only that f is bounded from below and $\boldsymbol{p}_k$ is a direction of descent.

11.4 Convergent algorithms

11.4.1 Basic convergence results

This section presents two basic convergence results for descent methods under different step length rules.

Theorem 11.4 (convergence of a gradient related algorithm) *Suppose that $f \in C^1$, and that for the starting point $\boldsymbol{x}_0$ it holds that the level set $\operatorname{lev}_f(f(\boldsymbol{x}_0)) = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0) \,\}$ is bounded. Consider the iterative algorithm defined by the description in Section 11.1. In this algorithm, suppose we make the following choices that hold for each iteration k:*

- *the search direction $\boldsymbol{p}_k$ satisfies the sufficient descent condition (11.4);*
- *$\|\boldsymbol{p}_k\| \leq M$, where M is some positive constant; and*

- *the Armijo step length rule* (11.12) *is used.*

Then, the sequence $\{\boldsymbol{x}_k\}$ *is bounded, the sequence* $\{f(\boldsymbol{x}_k)\}$ *is descending, lower bounded and therefore has a limit, and every limit point of* $\{\boldsymbol{x}_k\}$ *is stationary.*

Proof. That $\{\boldsymbol{x}_k\}$ is bounded follows since the algorithm, as stated, is a descent method, and we assumed that the level set of f at the starting point is bounded; therefore, the sequence of iterates must remain in that set and is therefore bounded.

The rest of the proof is by contradiction. Suppose that $\bar{\boldsymbol{x}}$ is a limit point of $\{\boldsymbol{x}_k\}$ but that $\nabla f(\bar{\boldsymbol{x}}) \neq \boldsymbol{0}^n$. It is clear that by the continuity of f, the whole sequence $\{f(\boldsymbol{x}_k)\}$ converges to the value $f(\bar{\boldsymbol{x}})$. Hence, $\{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1})\} \to 0$ must hold. According to the Armijo rule, then, $\{\alpha_k \nabla f(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{p}_k\} \to 0$. Here, there are two possibilities. Suppose that $\{\alpha_k\} \to 0$. Then, there must be some iteration $\bar{k}$ after which the initial step length is not accepted by the inequality (11.12), and therefore,

$$f(\boldsymbol{x}_k + (\alpha_k/\beta)\boldsymbol{p}_k) - f(\boldsymbol{x}_k) > \mu(\alpha_k/\beta)\nabla f(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{p}_k, \qquad k \geq \bar{k}.$$

Dividing both sides by $2\alpha_k$ we obtain in the limit that

$$\nabla f(\bar{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{p}^{\infty} \geq 0,$$

for any limit point $\boldsymbol{p}^{\infty}$ of the bounded sequence $\{\boldsymbol{p}_k\}$. But in the limit of the inequality (11.4) we then clearly reach a contradiction. So, in fact, we must have that $\{\alpha_k\} \not\to 0$. In this case, then, by the above we must have that $\{\nabla f(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{p}_k\} \to 0$ holds, so by letting k tend to infinity we obtain that

$$\nabla f(\bar{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{p}^{\infty} = 0,$$

which again produces a contradiction to the initial claim because of (11.4). We conclude that $\nabla f(\bar{\boldsymbol{x}}) = \boldsymbol{0}^n$ must therefore hold, and we are done. ∎

We note that since the resulting step length from an exact line search in particular must satisfy the Armijo rule (11.12), the above proof can be used to also establish the result of such a modification of the algorithm given in the theorem. We further note that there is no guarantee that the limit points $\bar{\boldsymbol{x}}$ is a local minimum; it may also be a *saddle point*, that is, a stationary point where $\nabla^2 f(\bar{\boldsymbol{x}})$ is indefinite, if it exists.

Another result is cited below from [BeT00]. It allows the Armijo step length rule to be replaced by a much simpler type of step length rule which is also used to minimize a class of non-differentiable functions (cf. Section 6.5). The proof requires the addition of a technical assumption:

Definition 11.5 (Lipschitz continuity) *A C^1 function $f : \mathbb{R}^n \to \mathbb{R}$ is said to have a* Lipschitz continuous *gradient mapping on $\mathbb{R}^n$ if there exists a scalar $L \geq 0$ such that*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\| \tag{11.13}$$

holds for every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. ∎

Check that the gradient of a C^2 function f is Lipschitz continuous whenever its Hessian matrix is bounded over $\mathbb{R}^n$.

Theorem 11.6 (on the convergence of gradient related methods) *Let $f \in C^1$. Consider the sequence $\{\boldsymbol{x}_k\}$ generated by the formula $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$. Suppose that:*

- *∇f is Lipschitz continuous on $\mathbb{R}^n$;*
- *$c_1\|\nabla f(\boldsymbol{x}_k)\|^2 \leq -\nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{p}_k, \quad c_1 > 0$;*
- *$\|\boldsymbol{p}_k\| \leq c_2\|\nabla f(\boldsymbol{x}_k)\|, \quad c_2 > 0$;*
- *$\alpha_k > 0$ satisfies that $\{\alpha_k\} \to 0$ and $\lim_{k\to\infty} \sum_{s=1}^{k} \alpha_s = \infty$.*

Then, either $\lim_{k\to\infty} f(\boldsymbol{x}_k) = -\infty$ holds, or $\lim_{k\to\infty} f(\boldsymbol{x}_k) = \bar{f}$ and $\lim_{k\to\infty} \nabla f(\boldsymbol{x}_k) = \mathbf{0}^n$ holds. ∎

In Theorem 11.4 convergence is only established in terms of that of *subsequences*, and the requirements include a level set boundedness condition that can be difficult to check. A strong convergence result is available for the case of *convex* functions f whenever we know that there exists at least one optimal solution. It follows readily from Theorem 12.4 on the gradient projection method for differentiable optimization over convex sets, whence we will not establish it here. In fact, for the special case of the steepest descent algorithm, we have already seen such a result in Theorem 6.23 for possibly even non-differentiable convex functions.

Theorem 11.7 (convergence of gradient related algorithms under convexity) *Suppose the function $f \in C^1$ on $\mathbb{R}^n$. Suppose further that f is convex and that the problem* (11.1) *has at least one optimal solution. Consider the iterative algorithm defined by the description in Section 11.1, under the three additional conditions stated in Theorem* 12.4*, and where the step length α_k is determined by the Armijo step length rule. Then, the sequence $\{\boldsymbol{x}_k\}$ converges to some optimal solution to* (11.1). ∎

We have so far neglected Step 4 in the algorithm description in Section 11.1 in that we assume in the above results that the sequence $\{\boldsymbol{x}_k\}$ is infinite. A termination criterion must obviously be applied if we are to obtain a result in a finite amount of time. This is the subject of the next section.

11.5 Finite termination criteria

As noted above, convergence to a stationary point is only asymptotic. How does one know when it is wise to terminate? A criterion based only on a small size of $\|\nabla f(\boldsymbol{x}_k)\|$ is no good—why? Because we compare with 0!

The recommendation is the combination of the following:

1. $\|\nabla f(\boldsymbol{x}_k)\| \leq \varepsilon_1(1 + |f(\boldsymbol{x}_k)|)$, $\varepsilon_1 > 0$ small;
2. $f(\boldsymbol{x}_{k-1}) - f(\boldsymbol{x}_k) \leq \varepsilon_2(1 + |f(\boldsymbol{x}_k)|)$, $\varepsilon_2 > 0$ small; and
3. $\|\boldsymbol{x}_{k-1} - \boldsymbol{x}_k\| \leq \varepsilon_3(1 + \|\boldsymbol{x}_k\|)$, $\varepsilon_3 > 0$ small.

The right-hand sides are constructed in order to eliminate some of the possible influence of bad scaling of the variable values, of the objective function, and of the gradient, and also of the possibility that some values are zero at the optimum solution.

Notice that using the criterion 2. only might mean that we terminate too soon if f is very flat; similarly with 3., we terminate prematurely if f is extremely steep around the stationary point we are approaching. The presence of the constant 1 is to remove the dependency of the criterion on the absolute values of f and $\boldsymbol{x}_k$, particularly if they are near zero.

We also note that using the $\|\cdot\|_2$ norm may not be good when n is very large: suppose that $\nabla f(\bar{\boldsymbol{x}}) = (\gamma, \gamma, \dots, \gamma)^{\mathrm{T}} = \gamma(1, 1, \dots, 1)^{\mathrm{T}}$. Then, $\|\nabla f(\bar{\boldsymbol{x}})\|_2 = \sqrt{n} \cdot \gamma$, which illustrates that the dimension of the problem may enter the norm. Better then is to use the ∞-norm: $\|\nabla f(\bar{\boldsymbol{x}})\|_\infty := \max_{1 \leq j \leq n} |\frac{\partial f(\bar{\boldsymbol{x}})}{\partial x_j}| = |\gamma|$, which does *not* depend on n.

Norms have other bad effects. Suppose that

$$\begin{aligned}
\boldsymbol{x}_{k-1} &= (1.44453, 0.00093, 0.0000079)^{\mathrm{T}}, \\
\boldsymbol{x}_k &= (1.44441, 0.00012, 0.0000011)^{\mathrm{T}};
\end{aligned}$$

then,

$$\begin{aligned}
\|\boldsymbol{x}_{k-1} - \boldsymbol{x}_k\|_\infty &= \|(0.00012, 0.00081, 0.0000068)^{\mathrm{T}}\|_\infty \\
&= 0.00081.
\end{aligned}$$

Here, the termination test would possibly pass, although the number of significant digits is very small (the first significant digit is still changing!) Norms emphasize larger elements, so small ones may have bad relative accuracy. This is a case where *scaling* is needed.

Suppose we know that $\boldsymbol{x}^* = (1, 10^{-4}, 10^{-6})^{\mathrm{T}}$. If, by transforming the space, we obtain that the optimal solution is $\hat{\boldsymbol{x}}^* = (1, 1, 1)^{\mathrm{T}}$, then the same relative accuracy would be possible to achieve for all variables. Let then

$$\hat{\boldsymbol{x}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 10^4 & 0 \\ 0 & 0 & 10^6 \end{pmatrix}}_{\boldsymbol{D}} \boldsymbol{x}.$$

Let

$$f(\boldsymbol{x}) := \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x},$$

$$\boldsymbol{Q} := \begin{pmatrix} 8 & 3\cdot 10^4 & 0 \\ 3\cdot 10^4 & 4\cdot 10^8 & 10^{10} \\ 0 & 10^{10} & 6\cdot 10^{12} \end{pmatrix},$$

$$\boldsymbol{c} := \begin{pmatrix} 11 \\ 8\cdot 10^4 \\ 7\cdot 10^6 \end{pmatrix}.$$

Hence, $\boldsymbol{x}^* = \boldsymbol{Q}^{-1}\boldsymbol{c} = (1, 10^{-4}, 10^{-6})^{\mathrm{T}}$.

With $\hat{\boldsymbol{x}} = \boldsymbol{D}\boldsymbol{x}$, we get the transformed problem to minimize $\hat{f}(\hat{\boldsymbol{x}}) := \frac{1}{2}\hat{\boldsymbol{x}}^{\mathrm{T}}(\boldsymbol{D}^{-1}\boldsymbol{Q}\boldsymbol{D}^{-1})\hat{\boldsymbol{x}} - (\boldsymbol{D}^{-1}\boldsymbol{c})^{\mathrm{T}}\hat{\boldsymbol{x}}$, with

$$\boldsymbol{D}^{-1}\boldsymbol{Q}\boldsymbol{D}^{-1} = \begin{pmatrix} 8 & 3 & 0 \\ 3 & 4 & 1 \\ 0 & 1 & 6 \end{pmatrix}; \qquad \boldsymbol{D}^{-1}\boldsymbol{c} = \begin{pmatrix} 11 \\ 8 \\ 7 \end{pmatrix},$$

and $\hat{\boldsymbol{x}}^* = (1,1,1)^{\mathrm{T}}$. Notice the change in condition number in the matrix!

The steepest descent algorithm takes only $\nabla f(\boldsymbol{x})$ into account, not $\nabla^2 f(\boldsymbol{x})$. Therefore, if the problem is badly scaled, it will suffer from a poor convergence behaviour. Introducing elements of $\nabla^2 f(\boldsymbol{x})$ into the search direction helps in this respect. This is the precisely the effect of using second-order (Newton-type) algorithms.

11.6 A comment on non-differentiability

The subject of non-differentiable optimization will not be taken up in generality here; it has been analyzed more fully for Lagrangian dual problems in Chapter 6. The purpose of this discussion is to explain, by means of an example, that things can go terribly wrong if we apply methods for the minimization of differentiable function when the function is non-differentiable.

The famous theorem of Rademacher states that a function that is Lipschitz continuous [cf. (11.13) for a statement of the Lipschitz condition for gradients] automatically is differentiable almost everywhere. That seems to imply that we should not worry about differentiability, because it is very unlikely that a non-differentiable point will be "hit" by mistake. This is certainly true if the subject is simply to pick points at random, but the subject of optimization deals with searching for a particular, *extremal* point in the sense of the objective function, and such points tend to be non-differentiable with a higher probability than zero! Suppose for example that we consider the convex (Why?) function

$$f(\boldsymbol{x}) := \underset{i \in \{1,\ldots,m\}}{\operatorname{maximum}} \{\boldsymbol{c}_i^{\mathrm{T}} \boldsymbol{x} + b_i\}, \qquad \boldsymbol{x} \in \mathbb{R}^n,$$

that is, a max-function defined by affine functions. It has the appearance shown in Figure 11.5.

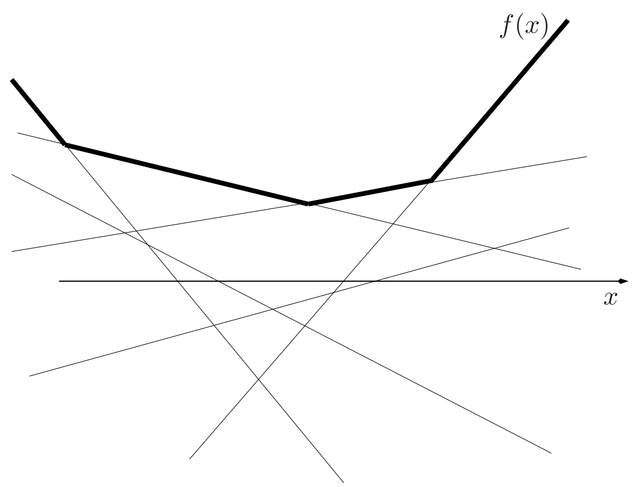


Figure 11.5: A piece-wise affine convex function.

Clearly, the minimum of this function is located at a point where it is non-differentiable.

We next look at a specific problem to which we will apply the method of steepest descent. Suppose that we are given the following objective function:[4]

$$f(x_1, x_2) := \begin{cases} 5(9x_1^2 + 16x_2^2)^{1/2}, & \text{if } x_1 > |x_2|, \\ 9x_1 + 16|x_2|, & \text{if } x_1 \le |x_2|. \end{cases}$$

[4]This example is due to Wolfe [Wol75].

For $x_1 > 0$, f is actually continuously differentiable! It is also convex, by the way. (Checking these facts is a nice exercise.)

If we start at a point $\boldsymbol{x}_0$ anywhere in the region $x_1 > |x_2| > (9/16)^2|x_1|$ then we obtain a sequence generated by steepest descent with exact line searches that defines a polygonal path with successive orthogonal segments, converging to $\bar{\boldsymbol{x}} = (0,0)^{\mathrm{T}}$.

But $\bar{\boldsymbol{x}}$ is not a stationary point! What is wrong here is that the gradients calculated say very little about the behaviour of f at the limit point $(0,0)^{\mathrm{T}}$. In fact, f is non-differentiable there. In this example, it in fact holds that $\lim_{x_1 \to -\infty} f(x_1, 0) = -\infty$, so steepest descent has failed miserably.

In order to resolve this problem, we need to take some necessary measures:

a) At a non-differentiable point, $\nabla f(x)$ must be replaced by a well-defined extension. Usually, we would replace it with a *subgradient*, that is, one of the vectors that define a supporting hyperplane to the graph of f. At $\bar{\boldsymbol{x}}$ it is the set defined by the convex hull of the two vectors $(9,16)^{\mathrm{T}}$ and $(9,-16)^{\mathrm{T}}$.

b) The step lengths must be chosen differently; exact line searches are clearly forbidden, as we have just seen.

From such considerations, we may develop algorithms that find optima to non-differentiable problems. They are referred to a *subgradient algorithms*, and are analyzed in Section 6.5.

11.7 Trust region methods

Trust region methods use quadratic models like Newton-type methods do, but avoid a line search by instead bounding the length of the search direction, thereby also influencing its direction.

Let $\psi_k(\boldsymbol{p}) := f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^{\mathrm{T}}\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{p}$. We say that the model ψ_k is *trusted* in a neighbourhood of $\boldsymbol{x}_k : \|\boldsymbol{p}\| \leq \Delta_k$. The use of this bound is apparent when $\nabla^2 f(\boldsymbol{x}_k)$ is not positive semi-definite. The problem to minimize $\psi_k(\boldsymbol{p})$ subject to $\|\boldsymbol{p}\| \leq \Delta_k$ can be solved (approximately) quite efficiently. The idea is that when $\nabla^2 f(\boldsymbol{x}_k)$ is badly conditioned, the value of Δ_k should be kept low—thus turning the algorithm more into a steepest descent-like method [recall (11.2)]—while if $\nabla^2 f(\boldsymbol{x}_k)$ is well conditioned, Δ_k should become large and allow unit steps to be taken. (Prove that the direction of $\boldsymbol{p}_k$ tends to that of the steepest descent method when $\Delta_k \to 0$!)

The vector $\boldsymbol{p}_k$ which solves the trust region problem satisfies $[\nabla^2 f(\boldsymbol{x}_k)+\gamma_k \boldsymbol{I}^n]\boldsymbol{p}_k = -\nabla f(\boldsymbol{x}_k)$ for some $\gamma_k \geq 0$ such that $\nabla^2 f(\boldsymbol{x}_k) + \gamma \boldsymbol{I}^n$ is positive semidefinite. The bounding enforced hence has a similar effect to that of the Levenberg–Marquardt strategy discussed in Section 11.2.2. Provided that the value of Δ_k is low enough, $f(\boldsymbol{x}_k + \boldsymbol{p}_k) < f(\boldsymbol{x}_k)$ holds. Even if $\nabla f(\boldsymbol{x}_k) = \boldsymbol{0}^n$ holds, $f(\boldsymbol{x}_k + \boldsymbol{p}_k) < f(\boldsymbol{x}_k)$ still holds, if $\nabla^2 f(\boldsymbol{x}_k)$ is not positive definite. So, progress is made also from stationary points if they are saddle points or local maxima. The robustness and strong convergence characteristics have made trust region methods quite popular.

The update of the trust region size is based on the following measure of similarity between the model ψ_k and f: Let

$$\rho_k = \frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{p}_k)}{f(\boldsymbol{x}_k) - \psi_k(\boldsymbol{p}_k)} = \frac{\text{actual reduction}}{\text{predicted reduction}}.$$

If $\rho_k \leq \mu$ let $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$ (unsuccessful step), else
$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{p}_k$ (successful step).

The value of Δ_k is updated in the following manner, depending on the value of ρ_k:

$$\begin{aligned} \rho_k \leq \mu &\Longrightarrow \Delta_{k+1} = \tfrac{1}{2}\Delta_k, \\ \mu < \rho_k < \eta &\Longrightarrow \Delta_{k+1} = \Delta_k, \\ \rho_k \geq \eta &\Longrightarrow \Delta_{k+1} = 2\Delta_k. \end{aligned}$$

Here, $0 < \mu < \eta < 1$, with typical choices being $\mu = \frac{1}{4}$ and $\eta = \frac{3}{4}$; μ is a bound used for deciding when the model can or cannot be trusted even within the region given, while η is used for deciding when the model is good enough to be used in a larger neighbourhood.

Figure 11.6 illustrates the trust region subproblem.

11.8 Conjugate gradient methods

When applied to nonlinear unconstrained optimization problems conjugate direction methods are methods intermediate between the steepest descent and Newton methods. The motivation behind them is similar to that for quasi-Newton methods: accelerating the steepest descent method but avoid the evaluation, storage and inversion of the Hessian matrix. They are analyzed for quadratic problems only; extensions to non-quadratic problems utilize that close to an optimal solution every problem is nearly quadratic. Even for non-quadratic problems, the last few decades of developments have resulted in conjugate direction methods being one of the most efficient general methodologies available.

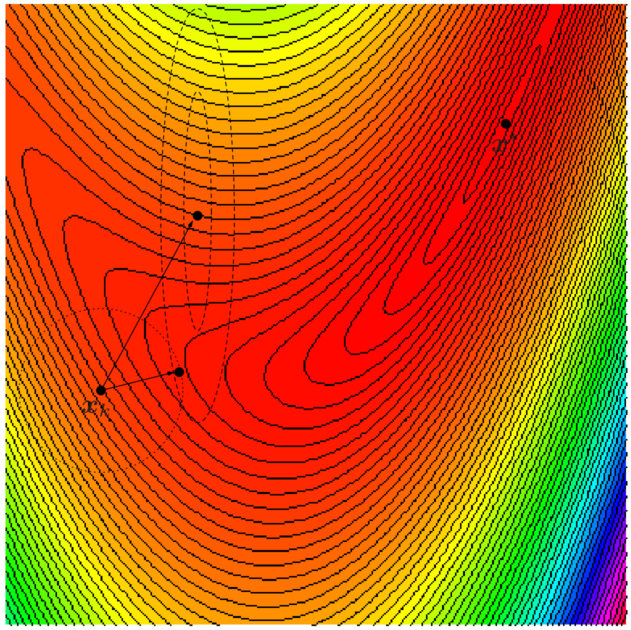


Figure 11.6: Trust region and line search step. The dashed ellipses are two level curves of the quadratic model constructed at $\boldsymbol{x}_k$, while the dotted circle is the boundary of the trust region. A step to the minimum of the quadratic model is here clearly inferior to the step taken within the trust region.

11.8.1 Conjugate directions

Definition 11.8 (conjugate direction) *Let $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ be symmetric. Two vectors $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ in $\mathbb{R}^n$ are $\boldsymbol{Q}$-*orthogonal*, or,* conjugate with respect to $\boldsymbol{Q}$, *if $\boldsymbol{p}_1^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_2 = 0$.* ■

Note that if $\boldsymbol{Q}$ is the zero matrix then every pair of vectors in $\mathbb{R}^n$ are conjugate; when $\boldsymbol{Q}$ is the unit matrix, conjugacy reduces to orthogonality. The following result is easy to prove (see Exercise 11.14).

Proposition 11.9 (conjugate vectors are linearly independent) *If $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ is positive definite and the collection $\boldsymbol{p}_1, \boldsymbol{p}_2, \dots, \boldsymbol{p}_k$ are mutually conjugate with respect to $\boldsymbol{Q}$, then they are also linearly independent.* ■

The usefulness of conjugate directions for the quadratic problem to

$$\operatorname*{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}) := \boldsymbol{x}^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{x} - \boldsymbol{q}^{\mathrm{T}} \boldsymbol{x}, \tag{11.14}$$

where from now on $\boldsymbol{Q}$ is symmetric and positive definite, is clear from the following identification: if the vectors $\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_{n-1}$ are $\boldsymbol{Q}$-orthogonal, then Proposition 11.9 implies that there exists a vector $\boldsymbol{w} \in \mathbb{R}^n$ with

$$\boldsymbol{x}^* = \sum_{i=0}^{n-1} w_i \boldsymbol{p}_i; \tag{11.15}$$

multiplying the equation by $\boldsymbol{Q}$ and scalar multiplying the result by $\boldsymbol{p}_i$ yields

$$w_i = \frac{\boldsymbol{p}_i^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{x}^*}{\boldsymbol{p}_i^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_i} = \frac{\boldsymbol{p}_i^{\mathrm{T}} \boldsymbol{q}}{\boldsymbol{p}_i^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_i}, \tag{11.16}$$

so that

$$\boldsymbol{x}^* = \sum_{i=0}^{n-1} \frac{\boldsymbol{p}_i^{\mathrm{T}} \boldsymbol{q}}{\boldsymbol{p}_i^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_i} \boldsymbol{p}_i. \tag{11.17}$$

Two ideas are embedded in (11.17): by selecting a proper set of orthogonal vectors $\boldsymbol{p}_i$, and by taking the appropriate scalar product all terms but i in (11.15) disappear. This could be accomplished by using any n orthogonal vectors, but (11.16) shows that by making them $\boldsymbol{Q}$-orthogonal we can express w_i without knowing $\boldsymbol{x}^*$.

11.8.2 Conjugate direction methods

The corresponding conjugate direction method for (11.14) is given by

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k, \qquad k = 0, \ldots, n-1,$$

where $\boldsymbol{x}_0 \in \mathbb{R}^n$ is arbitrary and α_k is obtained from an exact line search with respect to f in the direction of $\boldsymbol{p}_k$; cf. (11.9). The principal result about conjugate direction methods is that successive iterates minimize f over a progressively expanding linear manifold, or subspace, that after at most n iterations includes the minimizer of f over $\mathbb{R}^n$. In other words, defining

$$M_k := \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x} = \boldsymbol{x}_0 + \text{ subspace spanned by } \{\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_{n-1}\}\,\},$$

$$\{\boldsymbol{x}_{k+1}\} = \arg \underset{\boldsymbol{x} \in M_k}{\operatorname{minimum}}\ f(\boldsymbol{x}) \tag{11.18}$$

holds.

To show this, note that by the exact line search rule, for all i,

$$\left. \frac{\partial f(\boldsymbol{x}_i + \alpha \boldsymbol{p}_i)}{\partial \alpha} \right|_{\alpha = \alpha_i} = \nabla f(\boldsymbol{x}_{i+1})^{\mathrm{T}} \boldsymbol{p}_i = 0.$$

and for $i = 0, 1, \dots, k-1$,

$$\begin{aligned}\nabla f(\boldsymbol{x}_{k+1})^{\mathrm{T}} \boldsymbol{p}_i &= (\boldsymbol{Q}\boldsymbol{x}_{k+1} - \boldsymbol{q})^{\mathrm{T}} \boldsymbol{p}_i \\ &= \left(\boldsymbol{x}_{i+1} + \sum_{j=i+1}^{k} \alpha_j \boldsymbol{p}_j \right)^{\mathrm{T}} \boldsymbol{Q}\boldsymbol{p}_i - \boldsymbol{q}^{\mathrm{T}} \boldsymbol{p}_i \\ &= \boldsymbol{x}_{i+1}^{\mathrm{T}} \boldsymbol{Q}\boldsymbol{p}_i - \boldsymbol{q}^{\mathrm{T}} \boldsymbol{p}_i \\ &= \nabla f(\boldsymbol{x}_{i+1})^{\mathrm{T}} \boldsymbol{p}_i,\end{aligned}$$

where we used the conjugacy of $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$, $j \neq i$. Hence, $\nabla f(\boldsymbol{x}_{k+1})^{\mathrm{T}} \boldsymbol{p}_i = 0$ for every $i = 0, 1, \dots, k$, which verifies (11.18).

It is easy to get a picture of what is going on if we look at the case where $\boldsymbol{Q} = \boldsymbol{I}^n$ and $\boldsymbol{q} = \boldsymbol{0}^n$; since the level curves are circles, minimizing over the n coordinates one by one gives us $\boldsymbol{x}^*$ in n steps; in each iteration we also identify the optimal value of one of the variables. Conjugate directions in effect does this, although in a transformed space.[5]

The discussion so far has been based on an arbitrary selection of conjugate directions. There are many ways in which conjugate directions could be generated. For example, we could let the vectors $\boldsymbol{p}_i$, $i = 0, \dots, n-1$ be defined by the eigenvectors of $\boldsymbol{Q}$, as they are mutually orthogonal as well as conjugate with respect to $\boldsymbol{Q}$. (Why?) Such a procedure would however be too costly in large-scale applications. The remarkable feature of the conjugate gradient method to be presented below is that the new vector $\boldsymbol{p}_k$ can be generated directly from the vector $\boldsymbol{p}_{k-1}$—there is no need to remember any of the vectors $\boldsymbol{p}_0, \dots, \boldsymbol{p}_{k-2}$, and yet $\boldsymbol{p}_k$ will be conjugate to them all.

11.8.3 Generating conjugate directions

Given a set of linearly independent vectors $\boldsymbol{d}_0, \boldsymbol{d}_1, \dots, \boldsymbol{d}_k$ we can generate a set of mutually $\boldsymbol{Q}$-orthogonal vectors $\boldsymbol{p}_0, \boldsymbol{p}_1, \dots, \boldsymbol{p}_k$ such that they span the same subspace, by using the Gram–Schmidt procedure. We start the recursion with $\boldsymbol{p}_0 = \boldsymbol{d}_0$. Suppose that for $i < k$ we have $\boldsymbol{d}_0, \boldsymbol{d}_1, \dots, \boldsymbol{d}_i$ such that they span the same subspace as $\boldsymbol{p}_0, \boldsymbol{p}_1, \dots, \boldsymbol{p}_i$. Then, let $\boldsymbol{p}_{i+1}$ take the following form:

$$\boldsymbol{p}_{i+1} = \boldsymbol{d}_{i+1} + \sum_{m=0}^{i} c_m^{i+1} \boldsymbol{d}_m,$$

[5] Compare this to Newton's method as applied to the problem (11.14); its convergence in one step corresponds to the convergence in one step of the steepest descent method when we first have performed a coordinate transformation such that the level curves become circular.

choosing c_m^{i+1} so that $\boldsymbol{p}_{i+1}$ is $\boldsymbol{Q}$-orthogonal to $\boldsymbol{p}_0, \boldsymbol{p}_1, \dots, \boldsymbol{p}_i$. This will be true if, for $j = 0, 1, \dots, i$,

$$\boldsymbol{p}_{i+1}^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_j = \boldsymbol{d}_{i+1}^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_j + \left(\sum_{m=0}^{i} c_m^{i+1} \boldsymbol{p}_m \right)^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_j = 0.$$

Since $\boldsymbol{p}_0, \boldsymbol{p}_1, \dots, \boldsymbol{p}_i$ are $\boldsymbol{Q}$-orthogonal we have that $\boldsymbol{p}_m^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_j = 0$ if $m \neq j$, so

$$c_j^{i+1} = -\frac{\boldsymbol{d}_{i+1}^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_j}{\boldsymbol{p}_j^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_j}, \qquad j = 0, 1, \dots, i.$$

Some notes are in order regarding the above development: (a) it holds that $\boldsymbol{p}_j^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_j \neq 0$. (b) $\boldsymbol{p}_{i+1} \neq \boldsymbol{0}^n$; otherwise it would contradict the linear independence of $\boldsymbol{d}_0, \boldsymbol{d}_1, \dots, \boldsymbol{d}_k$. (c) Finally, $\boldsymbol{d}_{i+1}$ lies in the subspace spanned by $\boldsymbol{p}_0, \boldsymbol{p}_1, \dots, \boldsymbol{p}_{i+1}$, while $\boldsymbol{p}_{i+1}$ lies in the subspace spanned by $\boldsymbol{d}_0, \boldsymbol{d}_1, \dots, \boldsymbol{d}_{i+1}$, since these vectors span the same space. Therefore, the subspace identification above is true for $i + 1$, and we have shown that the Gram–Schmidt procedure has the property asked for.

11.8.4 Conjugate gradient methods

The conjugate gradient method applies the above Gram–Schmidt procedure to the vectors

$$\boldsymbol{d}_0 = -\nabla f(\boldsymbol{x}_0), \quad \boldsymbol{d}_1 = -\nabla f(\boldsymbol{x}_1), \quad \dots, \quad \boldsymbol{d}_{n-1} = -\nabla f(\boldsymbol{x}_{n-1}).$$

Thus, the conjugate gradient method is to take $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$, where α_k is determined through an exact line search and $\boldsymbol{p}_k$ is obtained through step k of the Gram–Schmidt procedure to the vector $\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$ and the previous vectors $\boldsymbol{p}_0, \boldsymbol{p}_1, \dots, \boldsymbol{p}_{k-1}$. In particular,

$$\boldsymbol{p}_k = -\nabla f(\boldsymbol{x}_k) + \sum_{j=0}^{k-1} \frac{\nabla f(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_j}{\boldsymbol{p}_j^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_j} \boldsymbol{p}_j. \tag{11.19}$$

It holds that $\boldsymbol{p}_0 = -\nabla f(\boldsymbol{x}_0)$, and termination occurs at step k if $\nabla f(\boldsymbol{x}_k) = \boldsymbol{0}^n$; the latter happens exactly when $\boldsymbol{p}_k = \boldsymbol{0}^n$. (Why?)

[*Note:* the search directions are based on negative gradients of f,

$$-\nabla f(\boldsymbol{x}_k) = \boldsymbol{q} - \boldsymbol{Q} \boldsymbol{x}_k,$$

which are identical to the residual in the linear system

$$\boldsymbol{Q} \boldsymbol{x} = \boldsymbol{q}$$

that identifies the optimal solution to (11.14).]

The formula (11.19) can in fact be simplified. The reason is that, because of the successive optimization over subspaces, $\nabla f(\boldsymbol{x}_k)$ is orthogonal to the subspace spanned by $\boldsymbol{p}_0, \boldsymbol{p}_1, \dots, \boldsymbol{p}_{k-1}$.

Proposition 11.10 (the conjugate gradient method) *The directions of the conjugate gradient method are generated by*

$$\boldsymbol{p}_0 = -\nabla f(\boldsymbol{x}_0); \tag{11.20a}$$
$$\boldsymbol{p}_k = -\nabla f(\boldsymbol{x}_k) + \beta_k \boldsymbol{p}_{k-1}, \qquad k = 1, 2, \dots, n-1, \tag{11.20b}$$

where

$$\beta_k = \frac{\nabla f(\boldsymbol{x}_k)^{\mathrm{T}} \nabla f(\boldsymbol{x}_k)}{\nabla f(\boldsymbol{x}_{k-1})^{\mathrm{T}} \nabla f(\boldsymbol{x}_{k-1})}. \tag{11.20c}$$

Moreover, the method terminates after at most n steps.

Proof. We first use induction to show that the gradients $\nabla f(\boldsymbol{x}_k)$ are linearly independent. It is clearly true for $k = 0$. Suppose that the method has not terminated after k steps, and that $\nabla f(\boldsymbol{x}_0), \nabla f(\boldsymbol{x}_1), \dots, \nabla f(\boldsymbol{x}_{k-1})$ are linearly independent. Being a conjugate gradient method we know that the subspace spanned by these vectors is the same as that spanned by the vectors $\boldsymbol{p}_0, \boldsymbol{p}_1, \dots, \boldsymbol{p}_{k-1}$:

$$\operatorname{span}(\boldsymbol{p}_0, \boldsymbol{p}_1, \dots, \boldsymbol{p}_{k-1}) = \operatorname{span}(\nabla f(\boldsymbol{x}_0), \nabla f(\boldsymbol{x}_1), \dots, \nabla f(\boldsymbol{x}_{k-1})). \tag{11.21}$$

Two cases are possible: either $\nabla f(\boldsymbol{x}_k) = \mathbf{0}^n$, whence the algorithm terminates at the optimal solution, or $\nabla f(\boldsymbol{x}_k) \neq \mathbf{0}^n$, in which case (by the expanding manifold property) it is orthogonal to $\boldsymbol{p}_0, \boldsymbol{p}_1, \dots, \boldsymbol{p}_{k-1}$. By (11.21) $\nabla f(\boldsymbol{x}_k)$ is linearly independent of $\nabla f(\boldsymbol{x}_0), \nabla f(\boldsymbol{x}_1), \dots, \nabla f(\boldsymbol{x}_{k-1})$, completing the induction. Since we have at most n linearly independent vectors in $\mathbb{R}^n$ the algorithm must stop after at most n steps.

The proof is completed by showing that the simplification in (11.20c) is possible. For all j with $\nabla f(\boldsymbol{x}_j) \neq \mathbf{0}^n$ we have that

$$\nabla f(\boldsymbol{x}_{j+1}) - \nabla f(\boldsymbol{x}_j) = \boldsymbol{Q}(\boldsymbol{x}_{j+1} - \boldsymbol{x}_j) = \alpha_j \boldsymbol{Q} \boldsymbol{p}_j,$$

and, since $\alpha_j \neq 0$,

$$\begin{aligned}
\nabla f(\boldsymbol{x}_i)^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_j &= \frac{1}{\alpha_j} \nabla f(\boldsymbol{x}_i)^{\mathrm{T}} [\nabla f(\boldsymbol{x}_{j+1}) - \nabla f(\boldsymbol{x}_j)] \\
&= \begin{cases} 0, & \text{if } j = 0, 1, \dots, i-2, \\ \frac{1}{\alpha_j} \nabla f(\boldsymbol{x}_i)^{\mathrm{T}} \nabla f(\boldsymbol{x}_i), & \text{if } j = i-1, \end{cases}
\end{aligned}$$

and also that

$$\boldsymbol{p}_j^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{p}_j = \frac{1}{\alpha_j}\boldsymbol{p}_j^{\mathrm{T}}[\nabla f(\boldsymbol{x}_{j+1}) - \nabla f(\boldsymbol{x}_j)].$$

Substituting these two relations into the Gram–Schmidt formula, we obtain that (11.20b) holds, with

$$\beta_k = \frac{\nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\nabla f(\boldsymbol{x}_k)}{\boldsymbol{p}_{k-1}^{\mathrm{T}}(\boldsymbol{p}_k - \boldsymbol{p}_{k-1})}.$$

From (11.20b) follows that $\boldsymbol{p}_{k-1} = -\nabla f(\boldsymbol{x}_{k-1}) + \beta_{k-1}\boldsymbol{p}_{k-2}$. Using this equation and the orthogonality of $\nabla f(\boldsymbol{x}_k)$ and $\nabla f(\boldsymbol{x}_{k-1})$ we can write the denominator in the expression for β_k as desired. We are done. ∎

We can deduce also further interesting properties of the algorithm. If the matrix $\boldsymbol{Q}$ has the eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, we have the following estimate of the distance to the optimal solution after iteration $k+1$:

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_{\boldsymbol{Q}}^2 \leq \left(\frac{\lambda_{n-k} - \lambda_1}{\lambda_{n-k} + \lambda_1}\right)^2 \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_{\boldsymbol{Q}}^2,$$

where $\|\boldsymbol{z}\|_{\boldsymbol{Q}}^2 = \boldsymbol{z}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{z}$, $\boldsymbol{z} \in \mathbb{R}^n$. What does this estimate tell us about the behaviour of the conjugate gradient algorithm? Suppose that we have a situation where the matrix $\boldsymbol{Q}$ has m large eigenvalues, and the remaining $n-m$ eigenvalues all are approximately equal to 1. Then the above tells us that after $m+1$ steps of the conjugate gradient algorithm,

$$\|\boldsymbol{x}_{m+1} - \boldsymbol{x}^*\|_{\boldsymbol{Q}} \approx (\lambda_{n-m} - \lambda_1)\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_{\boldsymbol{Q}}.$$

For a small value of $\lambda_{n-m} - \lambda_1$ this implies that the algorithm gives a good estimate of $\boldsymbol{x}^*$ already after $m+1$ steps. The conjugate gradient algorithm hence eliminates the effect of the largest eigenvalues first, as the convergence rate after the first $m+1$ steps does not depend on the $m+1$ largest eigenvalues.

The exercises offer additional insight into this convergence theory.

This is in sharp contrast with the convergence rate of the steepest descent algorithm, which is known to be

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_{\boldsymbol{Q}}^2 \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_{\boldsymbol{Q}}^2;$$

in other words, the rate of convergence worsens as the condition number of the matrix $\boldsymbol{Q}$, $\kappa(\boldsymbol{Q}) := \lambda_n/\lambda_1$, increases.[6]

[6]This type of bound on the convergence rate of the steepest descent method can

Nevertheless, the conjugate gradient method often comes with a *preconditioning*, which means that the system system solved is not $\boldsymbol{Qx} = \boldsymbol{q}$ but $\boldsymbol{MQx} = \boldsymbol{Mq}$ for some invertible square matrix $\boldsymbol{M}$, constructed such that the eigenvalues of $\boldsymbol{MQ}$ are better clustered than $\boldsymbol{Q}$ itself. (In other words, the condition number is reduced.)

11.8.5 Extension to non-quadratic problems

Due to the orthogonality of $\nabla f(\boldsymbol{x}_k)$ and $\nabla f(\boldsymbol{x}_{k-1})$, we could rewrite (11.20c) as

$$\beta_k = \frac{\nabla f(\boldsymbol{x}_k)^{\mathrm{T}}[\nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_{k-1})]}{\nabla f(\boldsymbol{x}_{k-1})^{\mathrm{T}}\nabla f(\boldsymbol{x}_{k-1})}. \tag{11.22}$$

The formula (11.20c) is often referred to as the *Fletcher–Reeves formula* (after the paper [FlR64]), while the formula (11.22) is referred to as the *Polak–Ribière formula* (after the paper [PoR69]).

For the quadratic programming problem, the two formulas are identical. However, they would not produce the same sequence of iterates if f were non-quadratic, and the conjugate gradient method has been extended also to such cases. The normal procedure is then to utilize the above algorithm for $k < n$ steps, after which a restart is made at the current iterate using the steepest descent direction; that is, we use the conjugate gradient algorithm several times in succession, in order to not lose conjugacy. The algorithm is not any more guaranteed to terminate after n steps, of course, but the algorithm has been observed to be quite efficient when the objective function and gradient values are cheap to evaluate; especially, this is true when comparing the algorithm class to that of quasi-Newton. (See [Lue84, Ber99] for further discussions on such computational issues.) It is also remarked in several sources that the Polak–Ribière formula (11.22) is preferable in the non-quadratic case.

---

also be extended to non-quadratic problems: suppose $\boldsymbol{x}^*$ is the unique optimal solution to the problem of minimizing the C^2 function f and that $\nabla^2 f(\boldsymbol{x}^*)$ is positive definite. Then, with $0 < \lambda_1 \leq \cdots \leq \lambda_n$ being the eigenvalues of $\nabla^2 f(\boldsymbol{x}^*)$ we have that for all k,

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*) \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 [f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)].$$

11.9 A quasi-Newton method

11.9.1 Introduction

As we have already touched upon in Section 11.2.2, most quasi-Newton methods are based on the idea to try to construct the (inverse) Hessian, or an approximation of it, through the use of information gathered in the process of solving the problem; the algorithm then works as a deflected gradient method where the matrix scaling of the negative of the gradient vector is the current approximation of the inverse Hessian matrix.

The BFGS updating formula that was given in Section 11.2.2 is a rank-two update of the Hessian matrix. There are several other versions of the quasi-Newton method, the most popular being based on rank-two updates but of the inverse of the Hessian rather than the Hessian matrix itself. We present one such method below.

11.9.2 The Davidon–Fletcher–Powell method

This algorithm is given in the two papers [Dav59, FlP63]. The algorithm is of interest to us here especially because we can show that through a special choice of matrix update, the quasi-Newton method implemented with an exact line search works exactly like a conjugate gradient method! Moreover, since quasi-Newton methods do not rely on exact line searches for convergence, we learn that quasi-Newton methods are, in this sense, more general than conjugate gradient methods.

The algorithm can be explained like this: start with a positive definite matrix $\boldsymbol{H}_0 \in \mathbb{R}^{n \times n}$, a point $\boldsymbol{x}_0 \in \mathbb{R}^n$, and with $k = 0$; then set

$$\boldsymbol{p}_k = -\boldsymbol{H}_k \nabla f(\boldsymbol{x}_k); \tag{11.23a}$$

$$\{\alpha_k\} = \arg\operatorname*{minimum}_{\alpha \geq 0} \; f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k); \tag{11.23b}$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k; \tag{11.23c}$$

$$\boldsymbol{q}_k = \nabla f(\boldsymbol{x}_{k+1}) - \nabla f(\boldsymbol{x}_k); \tag{11.23d}$$

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{\boldsymbol{p}_k \boldsymbol{p}_k^{\mathrm{T}}}{\boldsymbol{p}_k^{\mathrm{T}} \boldsymbol{q}_k} - \frac{(\boldsymbol{H}_k \boldsymbol{q}_k)(\boldsymbol{q}_k^{\mathrm{T}} \boldsymbol{H}_k)}{\boldsymbol{q}_k^{\mathrm{T}} \boldsymbol{H}_k \boldsymbol{q}_k}, \tag{11.23e}$$

and with $k := k + 1$ repeat.

We note that the matrix update in (11.23e) is a rank two update, since the two matrices added to $\boldsymbol{H}_k$ both are defined by the outer product of a given vector with itself.

We first demonstrate that the matrices $\boldsymbol{H}_k$ are positive definite. For any $\boldsymbol{x} \in \mathbb{R}^n$ we have

$$\boldsymbol{x}^{\mathrm{T}} \boldsymbol{H}_{k+1} \boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{H}_k \boldsymbol{x} + \frac{(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{p}_k)^2}{\boldsymbol{p}_k \boldsymbol{q}_k} - \frac{(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{H}_k \boldsymbol{q}_k)^2}{\boldsymbol{q}_k^{\mathrm{T}} \boldsymbol{H}_k \boldsymbol{q}_k}.$$

Defining $\boldsymbol{a} = \boldsymbol{H}_k^{1/2} \boldsymbol{x}$ and $\boldsymbol{b} = \boldsymbol{H}_k^{1/2} \boldsymbol{q}_k$ we can write this as

$$\boldsymbol{x}^{\mathrm{T}} \boldsymbol{H}_{k+1} \boldsymbol{x} = \frac{(\boldsymbol{a}^{\mathrm{T}} \boldsymbol{a})(\boldsymbol{b}^{\mathrm{T}} \boldsymbol{b}) - (\boldsymbol{a}^{\mathrm{T}} \boldsymbol{b})^2}{\boldsymbol{b}^{\mathrm{T}} \boldsymbol{b}} + \frac{(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{p}_k)^2}{\boldsymbol{p}_k^{\mathrm{T}} \boldsymbol{q}_k}.$$

We also have that

$$\boldsymbol{p}_k^{\mathrm{T}} \boldsymbol{q}_k = \boldsymbol{p}_k^{\mathrm{T}} \nabla f(\boldsymbol{x}_{k+1}) - \boldsymbol{p}_k^{\mathrm{T}} \nabla f(\boldsymbol{x}_{k+1}) = -\boldsymbol{p}_k^{\mathrm{T}} \nabla f(\boldsymbol{x}_k),$$

since

$$\boldsymbol{p}_k^{\mathrm{T}} \nabla f(\boldsymbol{x}_{k+1}) = 0 \tag{11.24}$$

due to the line search being exact. Therefore, by the definition of $\boldsymbol{p}_k$,

$$\boldsymbol{p}_k^{\mathrm{T}} \nabla f(\boldsymbol{x}_k) = \alpha_k \nabla f(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{H}_k \nabla f(\boldsymbol{x}_k),$$

and hence

$$\boldsymbol{x}^{\mathrm{T}} \boldsymbol{H}_{k+1} \boldsymbol{x} = \frac{(\boldsymbol{a}^{\mathrm{T}} \boldsymbol{a})(\boldsymbol{b}^{\mathrm{T}} \boldsymbol{b}) - (\boldsymbol{a}^{\mathrm{T}} \boldsymbol{b})^2}{\boldsymbol{b}^{\mathrm{T}} \boldsymbol{b}} + \frac{(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{p}_k)^2}{\alpha_k \nabla f(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{H}_k \nabla f(\boldsymbol{x}_k)}.$$

Both terms in the right-hand side are non-negative (Why?). We must finally show that not both can be zero at the same time. The first term disappears precisely when $\boldsymbol{a}$ and $\boldsymbol{b}$ are proportional. This in turn implies that $\boldsymbol{x}$ and $\boldsymbol{q}_k$ are proportional, say, $\boldsymbol{x} = \beta \boldsymbol{q}_k$ for some $\beta \in \mathbb{R}$. But this would mean that

$$\boldsymbol{p}_k^{\mathrm{T}} \boldsymbol{x} = \beta \boldsymbol{p}_k^{\mathrm{T}} \boldsymbol{q}_k = \beta \alpha_k \nabla f(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{H}_k \nabla f(\boldsymbol{x}_k) \neq 0,$$

whence $\boldsymbol{x}^{\mathrm{T}} \boldsymbol{H}_{k+1} \boldsymbol{x} > 0$ holds.

Notice that the fact that the line search is exact is not actually used; it is enough that the α_k chosen yields that $\boldsymbol{p}_k^{\mathrm{T}} \boldsymbol{q}_k > 0$.

The following proposition shows that the Davidon–Fletcher–Powell (DFP) algorithm (11.23) is a conjugate gradient algorithm which provides an optimal solution to (11.14) in at most n steps.

Theorem 11.11 (finite convergence of the DFP algorithm) *Consider the algorithm* (11.23) *for the problem* (11.14)*. Then,*

$$\boldsymbol{p}_i^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_j = 0, \qquad 0 \leq i < j \leq k, \tag{11.25a}$$

$$\boldsymbol{H}_{k+1}^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{p}_i = \boldsymbol{p}_i, \qquad 0 \leq i \leq k \tag{11.25b}$$

holds.

Proof. We have that

$$\boldsymbol{q}_k = \nabla f(\boldsymbol{x}_{k+1}) - \nabla f(\boldsymbol{x}_k) = \boldsymbol{Q}\boldsymbol{x}_{k+1} - \boldsymbol{Q}\boldsymbol{x}_k = \boldsymbol{Q}\boldsymbol{p}_k, \qquad (11.26)$$

and

$$\boldsymbol{H}_{k+1}\boldsymbol{Q}\boldsymbol{p}_k = \boldsymbol{H}_{k+1}\boldsymbol{q}_k = \boldsymbol{p}_k, \qquad (11.27)$$

the latter from (11.23e).

Proving (11.25) by induction, we see from the above equation that it is true for $k = 0$. Assume (11.25) true for $k - 1$. We have that

$$\nabla f(\boldsymbol{x}_k) = \nabla f(\boldsymbol{x}_{i+1}) + \boldsymbol{Q}(\boldsymbol{p}_{i+1} + \cdots + \boldsymbol{p}_{k-1}).$$

Therefore, from (11.25a) and (11.24),

$$\boldsymbol{p}_i^{\mathrm{T}} \nabla f(\boldsymbol{x}_k) = \boldsymbol{p}_i^{\mathrm{T}} \nabla f(\boldsymbol{x}_{i+1}) = 0, \qquad 0 \le i < k.$$

Hence from (11.25b)

$$\boldsymbol{p}_i^{\mathrm{T}} \boldsymbol{Q}\boldsymbol{H}_k\boldsymbol{p}_k = 0, \qquad i < k, \qquad (11.28)$$

which proves (11.25a) for k.

Now, since from (11.25b) for $k - 1$, (11.26), and (11.28)

$$\boldsymbol{q}_k^{\mathrm{T}} \boldsymbol{H}_k\boldsymbol{Q}\boldsymbol{p}_i = \boldsymbol{q}_k^{\mathrm{T}} \boldsymbol{p}_i = \boldsymbol{p}_k^{\mathrm{T}} \boldsymbol{Q}\boldsymbol{p}_i = 0, \qquad 0 \le i < k,$$

we have that

$$\boldsymbol{H}_{k+1}\boldsymbol{Q}\boldsymbol{p}_i = \boldsymbol{H}_k\boldsymbol{Q}\boldsymbol{p}_i = \boldsymbol{p}_i, \qquad 0 \le i < k.$$

This together with (11.27) proves (11.25b) for k. ■

Since the $\boldsymbol{p}_k$-vectors are $\boldsymbol{Q}$-orthogonal and since we minimize f successively over these directions, the DFP algorithm is a conjugate direction method. Especially, if the initial matrix $\boldsymbol{H}_0$ is taken to be the unit matrix, it becomes the conjugate gradient method. In any case, however, convergence is obtained after at most n steps.

Finally, we note that (11.25b) shows that the vectors $\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_k$ are eigenvectors corresponding to unity eigenvalues of the matrix $\boldsymbol{H}_{k+1}\boldsymbol{Q}$. These eigenvectors are linearly independent, since they are $\boldsymbol{Q}$-orthogonal, and therefore we have that

$$\boldsymbol{H}_n = \boldsymbol{Q}^{-1}.$$

In other words, with any choice of initial matrix $\boldsymbol{H}_0$ (as long as it is positive definite) n steps of the 2-rank updates in (11.23e) result in the final matrix being identical to the inverse of the Hessian.

11.10 Convergence rates

The *local convergence rate* is a statement about the speed in which one iteration takes the guess closer to the solution.

Definition 11.12 (local convergence rate) *Suppose that* $\mathbb{R}^n \supset \{\boldsymbol{x}_k\} \to \boldsymbol{x}^*$. *Consider for large* k *the quotients*

$$q_k := \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|}.$$

(a) [linear convergence rate] *We say that the speed of convergence is* linear *if*

$$\limsup_{k\to\infty} q_k < 1.$$

A linear convergence rate is roughly equivalent to the statement that we get one new correct digit per iteration.

(b) [superlinear convergence rate] *We say that the speed of convergence is* superlinear *if*

$$\lim_{k\to\infty} q_k = 0.$$

(c) [quadratic convergence rate] *We say that the speed of convergence is* quadratic *if*

$$\limsup_{k\to\infty} \frac{q_k}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|} \leq c, \qquad c \geq 0.$$

A linear convergence rate is roughly equivalent to the statement that the number of correct digits is doubled in every iteration. ∎

The steepest descent method has, at most, a linear rate of convergence, moreover often with a constant q_k near unity. Newton-like algorithms have, however, superlinear convergence if $\nabla^2 f(\boldsymbol{x}^*)$ is positive definite, and even quadratic local convergence can be achieved for Newton's method if $\nabla^2 f$ is Lipschitz continuous in a neighbourhood of $\boldsymbol{x}^*$.

11.11 Implicit functions

Suppose that the value of $f(\boldsymbol{x})$ is given through a simulation procedure:

$$\xrightarrow{\quad \boldsymbol{x} \in \mathbb{R}^n \quad} \boxed{\text{Simulation}} \xrightarrow{\qquad\qquad} \boldsymbol{y} \in \mathbb{R}^m, f(\boldsymbol{x}, \boldsymbol{y}(\boldsymbol{x}))$$

If the response $\boldsymbol{y}(\boldsymbol{x})$ from the input $\boldsymbol{x}$ is unknown explicitly, then we cannot differentiate $\boldsymbol{x} \mapsto f(\boldsymbol{x}, \boldsymbol{y}(\boldsymbol{x}))$ with respect to $\boldsymbol{x}$. If, however, we believe that $\boldsymbol{y}(\cdot)$ *is differentiable*, which means that $\boldsymbol{y}$ is very stable with respect to changes in $\boldsymbol{x}$, then $\nabla_{\boldsymbol{x}} \boldsymbol{y}(\boldsymbol{x})$, and hence $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}(\boldsymbol{x}))$ can be calculated *numerically*. The use of the Taylor expansion technique that follows is only practical if $\boldsymbol{y}(\boldsymbol{x})$ is "cheap"; if it takes an hour or more to run the simulation, then it is probably too costly.

Let $\boldsymbol{e}_i = (0, 0, \dots, 0, 1, 0, \dots, 0)^{\mathrm{T}}$ be the unit vector in $\mathbb{R}^n$ where the only non-zero entry is in position i. Then,

$$\begin{aligned} f(\boldsymbol{x} + \alpha \boldsymbol{e}_i) &= f(\boldsymbol{x}) + \alpha \boldsymbol{e}_i^{\mathrm{T}} \nabla f(\boldsymbol{x}) + (\alpha^2/2) \boldsymbol{e}_i^{\mathrm{T}} \nabla^2 f(\boldsymbol{x}) \boldsymbol{e}_i + \dots \\ &= f(\boldsymbol{x}) + \alpha \partial f(\boldsymbol{x}) / \partial x_i + (\alpha^2/2) \partial^2 f(\boldsymbol{x}) / \partial x_i^2 + \dots \end{aligned}$$

So, for small $\alpha > 0$,

$$\begin{aligned} \frac{\partial f(\boldsymbol{x})}{\partial x_i} &\approx \frac{f(\boldsymbol{x} + \alpha \boldsymbol{e}_i) - f(\boldsymbol{x})}{\alpha} \quad \text{(forward difference)} \\ \frac{\partial f(\boldsymbol{x})}{\partial x_i} &\approx \frac{f(\boldsymbol{x} + \alpha \boldsymbol{e}_i) - f(\boldsymbol{x} - \alpha \boldsymbol{e}_i)}{2\alpha} \quad \text{(central difference)} \end{aligned}$$

The value of α is typically set to a function of the machine precision; if chosen too large, we get a bad approximation of the partial derivative, while a too small value might result in numerical cancellation.

11.12 Notes and further reading

The material of this chapter is mostly classic; text books covering similar material in more depth include [DeS83, Lue84, Fle87, BSS93, BGLS03]. Line search methods were first developed by Newton [New1687], and the steepest descent method is due to Cauchy [Cau1847]. The Armijo rule is due to Armijo [Arm66], and the Wolfe condition is due to Wolfe [Wol69]. The classic book by Brent [Bre73] analyzes algorithms that do not use derivatives, especially line search methods.

Rademacher's Theorem, which states that a Lipschitz continuous function is differentiable everywhere except on sets of Lebesgue measure zero, is due to Rademacher [Rad19]. The Lipschitz condition is due to Lipschitz [Lip1877]. Algorithms for the minimization of non-differentiable convex functions are given in [Sho85, HiL93, Ber99, BGLS03].

Trust region methods are given a thorough treatment in the book [CGT00]. The material on the conjugate gradient and BFGS methods was collected from [Lue84, Ber99]; another good source is [NoW99].

An increasingly popular class of algorithms for problems with an implicit objective function is the class of *pattern search methods*. With such algorithms the search for a good gradient-like direction is replaced by calculations of the objective function along directions specified by a pattern of possible points. For a good introduction to this field, see [KLT03].

11.13 Exercises

Exercise 11.1 (well-posedness of the Armijo rule) Establish the following, through an argument by contradiction: If $f \in C^1$, $\boldsymbol{x}_k \in \mathbb{R}^n$ and $\boldsymbol{p}_k \in \mathbb{R}^n$ satisfies $\nabla f(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{p}_k < 0$, then for every choice of $\mu \in (0,1)$ there exists $\bar{\alpha} > 0$ such that every $\alpha \in (0, \bar{\alpha}]$ satisfies (11.12). In other words, which ever positive first trial step length α we choose, we will find a step length that satisfies (11.12) in a finite number of trials. ■

Exercise 11.2 (descent direction) Investigate whether the direction of $\boldsymbol{p} = (2, -1)^{\mathrm{T}}$ is a direction of descent with respect to the function

$$f(\boldsymbol{x}) = x_1^2 + x_1 x_2 - 4x_2^2 + 10$$

at $\boldsymbol{x} = (1,1)^{\mathrm{T}}$. ■

Exercise 11.3 (Newton's method) Suppose that you wish to solve the unconstrained problem to

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\operatorname{minimize}} \; f(\boldsymbol{x}),$$

where f is twice continuously differentiable. You are naturally interested in using Newton's method (with line searches).

(a) At some iteration you get the error message, "Step length is zero." Which reason(s) can there be for such a message?

(b) At some iteration you get the error message, "Search direction does not exist." Which reason(s) can there be for such a message?

(c) Describe at least one means to modify Newton's method such that neither of the above two error message will ever appear. ■

Exercise 11.4 (Steepest descent) Is it possible to reach the (unique) optimal solution to the problem of minimizing the function $f(\boldsymbol{x}) := (x_1-2)^2+5(x_2+6)^2$ over $\mathbb{R}^2$ by the use of the steepest descent algorithm, if we first perform a variable substitution? If so, perform it and thus find the optimal solution. ■

Exercise 11.5 (Steepest descent with exact line search) Consider the problem to

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\operatorname{minimize}} \; f(\boldsymbol{x}) := (2x_1^2 - x_2)^2 + 3x_1^2 - x_2.$$

(a) Perform one iteration of the steepest descent method, starting at $\boldsymbol{x}_0 := (1/2, 5/4)^{\mathrm{T}}$.

(b) Is the function convex around $\boldsymbol{x}_1$?

(c) Will the method converge to a global optimum? Why/why not? ■

Exercise 11.6 (Newton's method with exact line search) Consider the problem to

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\operatorname{minimize}}\ f(\boldsymbol{x}) := (x_1 + 2x_2 - 3)^2 + (x_1 - 2)^2.$$

(a) Start from $\boldsymbol{x}_0 := (0, 0)^{\mathrm{T}}$, and perform one iteration of Newton's method with an exact line search.

(b) Are there any descent directions from $\boldsymbol{x}_1$?

(c) Is $\boldsymbol{x}_1$ optimal? Why/why not? ■

Exercise 11.7 (Newton's method with Armijo line search) Consider the problem to

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\operatorname{minimize}}\ f(\boldsymbol{x}) := \frac{1}{2}(x_1 - 2x_2)^2 + x_1^4.$$

(a) Start from $\boldsymbol{x}_0 := (2, 1)^{\mathrm{T}}$, and perform one iteration of Newton's method with the Armijo rule, using the fraction requirement $\mu = 0.1$.

(b) Determine the values of $\mu \in (0, 1)$ such that the step length $\alpha = 1$ will be accepted. ■

Exercise 11.8 (Newton's method for nonlinear equations) Suppose the function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable and consider the following system of nonlinear equations:

$$\boldsymbol{f}(\boldsymbol{x}) = \mathbf{0}^n.$$

Newton's method for the solution of unconstrained optimization problems has its correspondence for the above problem.

Given an iterate $\boldsymbol{x}_k$ we construct a linear approximation of the nonlinear function; this approximation results in an approximate *linear* system of equations of the form

$$\boldsymbol{f}(\boldsymbol{x}_k) + \nabla \boldsymbol{f}(\boldsymbol{x}_k)(\boldsymbol{x} - \boldsymbol{x}_k) = \mathbf{0}^n,$$

or, equivalently,

$$\nabla \boldsymbol{f}(\boldsymbol{x}_k)\boldsymbol{x} = \nabla \boldsymbol{f}(\boldsymbol{x}_k)\boldsymbol{x}_k - \boldsymbol{f}(\boldsymbol{x}_k),$$

where

$$\nabla \boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} \nabla f_1(\boldsymbol{x})^{\mathrm{T}} \\ \nabla f_2(\boldsymbol{x})^{\mathrm{T}} \\ \vdots \\ \nabla f_n(\boldsymbol{x})^{\mathrm{T}} \end{pmatrix}$$

is the *Jacobian* of $\boldsymbol{f}$ at $\boldsymbol{x}$. Assuming that the Jacobian is non-singular, the above linear system has a unique solution, which defines the new iterate, $\boldsymbol{x}_{k+1}$, that is,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \nabla \boldsymbol{f}(\boldsymbol{x}_k)^{-1} \boldsymbol{f}(\boldsymbol{x}_k).$$

(One can show that if $\boldsymbol{f}$ satisfies some additional requirements, this sequence of iterates will converge to a solution to the original nonlinear system, either from any starting point—global convergence—or from a point sufficiently closed to a solution—local convergence.)

(a) Consider the nonlinear system

$$\boldsymbol{f}(x_1, x_2) = \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix} = \begin{pmatrix} 2(x_1 - 2)^3 + x_1 - 2x_2 \\ 4x_2 - 2x_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Perform one iteration of the above algorithm, starting from $\boldsymbol{x}_0 = (1, 0)^{\mathrm{T}}$. Calculate the value of

$$\|\boldsymbol{f}(x_1, x_2)\| = \sqrt{f_1(x_1, x_2)^2 + f_2(x_1, x_2)^2}$$

both at $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$. (Observe that $\|\boldsymbol{f}(\boldsymbol{x})\| = 0$ if and only if $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}^n$, whence the values of $\|\boldsymbol{f}(\boldsymbol{x}_k)\|$, $k = 1, 2, \dots$ can be used as a measure of convergence of the iterates.)

(b) Explain why the above method generalizes Newton's method for unconstrained optimization to a larger class of problems. ∎

Exercise 11.9 (over-determined linear equations) Consider the problem to

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2,$$

where $\boldsymbol{A}$ is an $m \times n$ matrix and $\boldsymbol{b} \in \mathbb{R}^m$. Assume that $m \geq n$ and that the rank of $\boldsymbol{A}$ is n.

(a) Write down the necessary optimality conditions for this problem. Are they also necessary for global optimality? Why/why not?

(b) Write down the globally optimal solution in closed form. ∎

Exercise 11.10 (sufficient descent conditions) Consider the sufficient descent condition (11.5). Why does it have that form, and why is the alternative form

$$-\nabla f(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{p}_k \geq s_1$$

not acceptable? ∎

Exercise 11.11 (Newton's method under affine transformations) Suppose that we make the following change of variables: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$, where

$\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is invertible. Show that if we apply Newton's method to the problem in $\boldsymbol{y}$, we obtain exactly the same sequence as when applying the method in the original space. In other words, show that Newton's method is invariant to such changes of variables. ■

Exercise 11.12 (Levenberg–Marquardt, exam 990308) Consider the unconstrained optimization problem to

$$\text{minimize } f(\boldsymbol{x}) := \boldsymbol{q}^{\mathrm{T}}\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x}, \tag{11.29a}$$

$$\text{subject to } \boldsymbol{x} \in \mathbb{R}^n, \tag{11.29b}$$

where $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ is positive semi-definite but not positive definite. We attack the problem through a Levenberg–Marquardt strategy, that is, we utilize a Newton-type method where a multiple $\gamma > 0$ of the unit matrix is added to the Hessian of f (that is, to the matrix $\boldsymbol{Q}$) in order to guarantee that the (modified) Newton equation is uniquely solvable. (See Section 11.2.2.) This implies that, given an iteration point $\boldsymbol{x}_k$, the search direction $\boldsymbol{p}_k$ is determined by solving the linear system

$$[\nabla^2 f(\boldsymbol{x}_k) + \gamma \boldsymbol{I}^n]\boldsymbol{p} = -\nabla f(\boldsymbol{x}_k), \tag{11.30}$$

that is,

$$[\boldsymbol{Q} + \gamma \boldsymbol{I}^n]\boldsymbol{p} = -(\boldsymbol{Q}\boldsymbol{x}_k + \boldsymbol{q}).$$

(a) Consider the formula

$$\boldsymbol{x}_{k+1} := \boldsymbol{x}_k + \boldsymbol{p}_k, \qquad k = 0, 1, \ldots, \tag{11.31}$$

that is, the algorithm that is obtained by utilizing the Newton-like search direction $\boldsymbol{p}_k$ from (11.30) and the step length 1 in every iteration. Show that this iterative step is the same as that to let $\boldsymbol{x}_{k+1}$ be given by the solution to the problem to

$$\text{minimize } f(\boldsymbol{y}) + \frac{\gamma}{2}\|\boldsymbol{y} - \boldsymbol{x}_k\|^2, \tag{11.32a}$$

$$\text{subject to } \boldsymbol{y} \in \mathbb{R}^n. \tag{11.32b}$$

(b) Suppose that an optimal solution to (11.29) exists. Suppose also that the sequence $\{\boldsymbol{x}_k\}$ generated by the algorithm (11.31) converges to a point $\boldsymbol{x}_\infty$. (This can actually be shown to hold.) Show that $\boldsymbol{x}_\infty$ is an optimal solution to (11.29).

[*Note:* This algorithm is in fact a special case of the *proximal point algorithm*. Suppose that f is a convex function on $\mathbb{R}^n$ and the variables are constrained to a non-empty, closed and convex set $S \subseteq \mathbb{R}^n$.

We extend the iteration formula (11.32) to the following:

$$\begin{aligned} &\text{minimize } f(\boldsymbol{y}) + \frac{\gamma_k}{2}\|\boldsymbol{y} - \boldsymbol{x}_k\|^2, && \text{(11.33a)} \\ &\text{subject to } \boldsymbol{y} \in S, && \text{(11.33b)} \end{aligned}$$

where $\{\gamma_k\} \subset (0, 2)$ is a sequence of positive numbers that is bounded away from zero, and where $\boldsymbol{x}_{k+1}$ is taken as the unique vector $\boldsymbol{y}$ solving (11.33). If an optimal solution exists, it is possible to show that the sequence given by (11.33) converges to a solution. See [Pat98, Ber99] for overviews of this class of methods. (It is called "proximal point" because of the above interpretation: that the next iterate is close, proximal, to the previous one.)] ∎

Exercise 11.13 (unconstrained optimization algorithms, exam 980819) Consider the unconstrained optimization problem to

$$\begin{aligned} &\text{minimize } f(\boldsymbol{x}), \\ &\text{subject to } \boldsymbol{x} \in \mathbb{R}^n, \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is in C^1.

Let $\{\boldsymbol{x}_k\}$ be a sequence of iteration points generated by some algorithm for solving this problem, and suppose that it holds that $\{\nabla f(\boldsymbol{x}_k)\} \to \boldsymbol{0}^n$, that is, the gradient value tends to zero (which of course is a favourable behaviour of the algorithm). The question is what this means in terms of the convergence of the more important sequence $\{\boldsymbol{x}_k\}$.

Consider therefore the sequence $\{\boldsymbol{x}_k\}$, and also the sequence $\{f(\boldsymbol{x}_k)\}$ of function values. Given the assumption that $\{\nabla f(\boldsymbol{x}_k)\} \to \boldsymbol{0}^n$, is it true that $\{\boldsymbol{x}_k\}$ and/or $\{f(\boldsymbol{x}_k)\}$ converges or are even bounded? Provide every possible case in terms of the convergence of these two sequences, and give examples, preferably simple ones for $n = 1$. ∎

Exercise 11.14 (conjugate directions) Prove Proposition 11.9. ∎

Exercise 11.15 (conjugate gradient method) Apply the conjugate gradient method to the system $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{q}$, where

$$\boldsymbol{Q} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{q} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

∎

Exercise 11.16 (convergence of the conjugate gradient method, I) In the conjugate gradient method, prove that the vector $\boldsymbol{p}_i$ can be written as a linear combination of the set of vectors $\{\boldsymbol{q}, \boldsymbol{Q}\boldsymbol{q}, \boldsymbol{Q}^2\boldsymbol{q}, \dots, \boldsymbol{Q}^i\boldsymbol{q}\}$. Also prove that $\boldsymbol{x}_{i+1}$ minimizes the quadratic function $\mathbb{R}^n \ni \boldsymbol{x} \mapsto f(\boldsymbol{x}) := \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{q}^{\mathrm{T}}\boldsymbol{x}$ over all the linear combinations of these vectors. ■

Exercise 11.17 (convergence of the conjugate gradient method, II) Use the result of the previous problem to establish that the conjugate gradient method converges in a number of iterations equal to the number of distinct eigenvalues of the matrix $\boldsymbol{Q}$. ■

Answers to the exercises

II

Chapter 1: Modelling and classification

Exercise 1.1
Variables:

$$\begin{aligned} x_j &= \text{number of units produced in process } j,\ j = 1, 2; \\ y &= \text{number of half hours hiring the model.} \end{aligned}$$

Optimization model:

$$\begin{aligned} \text{maximize} \quad & f(x, y) := 50(3x_1 + 5x_2) - 3(x_1 + 2x_2) - 2(2x_1 + 3x_2) - 5000y, \\ \text{subject to} \quad & x_1 + 2x_2 \leq 20,000, \\ & 2x_1 + 3x_2 \leq 35,000, \\ & 3x_1 + 5x_2 \leq 1,000 + 200y \\ & x_1 \geq 0, \\ & x_2 \geq 0, \\ & 0 \leq y \leq 6. \end{aligned}$$

■

Exercise 1.2
Variables:

$$\begin{aligned} x_j &= \text{number of trainees trained during month } j,\ j = 1, \dots, 5; \\ y_j &= \text{number of technicians available at the beginning of month } j,\ j = 1, \dots, 5. \end{aligned}$$

Optimization model:

$$
\begin{array}{lrcl}
\text{minimize } z = & \displaystyle\sum_{j=1}^{5}(15000y_j + 7500x_j) & & \\
\text{subject to} & 160y_1 - 50x_1 \geq & & 6000 \\
& 160y_2 - 50x_2 \geq & & 7000 \\
& 160y_3 - 50x_3 \geq & & 8000 \\
& 160y_4 - 50x_4 \geq & & 9500 \\
& 160y_5 - 50x_5 \geq & & 11,500 \\
& 0.95y_1 + \quad x_1 = & & y_2 \\
& 0.95y_2 + \quad x_2 = & & y_3 \\
& 0.95y_3 + \quad x_3 = & & y_4 \\
& 0.95y_4 + \quad x_4 = & & y_5 \\
& y_1 = & & 50 \\
& y_j, x_j \geq 0, \; j = 1, \dots, 5. & &
\end{array}
$$

■

Exercise 1.3 We declare the following indices:

- i, $i = 1, \dots, 3$: Work place,
- k, $k = 1, \dots, 2$: Connection point,

and variables

- (x_i, y_i): Coordinates for work place i;
- $t_{i,k}$: Indicator variable; its value is defined as 1 if work place i is connected to the connection point k, and as 0 otherwise;
- z: The longest distance to the window.

The problem to minimize the maximum distance to the window is that to

$$\text{minimize} \quad z, \tag{B.1}$$

subject to the work spaces being inside the rectangle:

$$\frac{d}{2} \leq x_i \leq l - \frac{d}{2}, \quad i = 1, \dots, 3, \tag{B.2}$$

$$\frac{d}{2} \leq y_i \leq b - \frac{d}{2}, \quad i = 1, \dots, 3, \tag{B.3}$$

that the work spaces do not overlap:

$$(x_i - x_j)^2 + (y_i - y_j)^2 \geq d^2, \quad i = 1, \dots, 3, \; j = 1, \dots, 3, \; i \neq j, \tag{B.4}$$

that the cables are long enough:

$$t_{1,k}\left[(x_i-\frac{l}{2})^2+(y_i-0)^2\right]\le a_i^2,\quad i=1,\dots,3, \tag{B.5}$$
$$t_{2,k}\left[(x_i-l)^2+(y_i-\frac{b}{2})^2\right]\le a_i^2,\quad i=1,\dots,3, \tag{B.6}$$

that each work space must be connected to a connection point:

$$t_{i,1}+t_{i,2}=1,\quad i=1,\dots,3, \tag{B.7}$$
$$t_{i,k}\in\{0,1\},\quad i=1,\dots,3,\ k=1,2, \tag{B.8}$$

and finally that the value of z is at least as high as the longest distance to the window:

$$b-y_i\ge z,\quad i=1,\dots,3. \tag{B.9}$$

The problem hence is to minimize the objective function in (B.1) under the constraints (B.2)–(B.9). ■

Exercise 1.4 We declare the following indices:

- i: Warehouses ($i=1,\dots,10$),
- j: Department stores ($j=1,\dots,30$),

and variables:

- x_{ij}: portion (between 0 and 1) of the total demand at department store j which is served from warehouse i,
- y_i: Indicator variable; its value is defined as 1 if warehouse i is built, and 0 otherwise.

We also need the following constants, describing the department stores that are within the specified maximum distance from a warehouse:

$$a_{ij}:=\begin{cases}1, & \text{if } d_{ij}\le D,\\ 0, & \text{otherwise,}\end{cases}\qquad i=1,\dots,10,\ j=1,\dots,30.$$

(a) The problem becomes:

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{10} c_i y_i, \\
\text{subject to} \quad & x_{ij} \leq a_{ij} y_i, \qquad i = 1, \dots, 10, \quad j = 1, \dots, 30, \\
& \sum_{j=1}^{30} e_j x_{ij} \leq k_i y_i, \qquad i = 1, \dots, 10, \\
& \sum_{i=1}^{10} x_{ij} = 1, \qquad j = 1, \dots, 30, \\
& x_{ij} \geq 0, \qquad j = 1, \dots, 30, \\
& y_i \in \{0, 1\}, \qquad i = 1, \dots, 10.
\end{aligned}
$$

The first constraint makes sure that only warehouses that are built and which lie sufficiently close to a department store can supply any goods to it.

The second constraint describes the capacity of each warehouse, and the demand at the various department stores.

The third and fourth constraints describe that the total demand at a department store must be a non-negative (in fact, convex) combination of the contributions from the different warehouses.

(b) Additional constraints: $x_{ij} \in \{0, 1\}$ for all i and j. ∎

Chapter 3: Convexity

Exercise 3.1 Use the definition of convexity (Definition 3.1). ∎

Exercise 3.2 a) S is a polyhedron. It is the parallelogram with the corners $a_1 + a_2, a_1 - a_2, -a_1 + a_2, -a_1 - a_2$, that is, $S = \operatorname{conv}\{a_1 + a_2, a_1 - a_2, -a_1 + a_2, -a_1 - a_2\}$ which is a polytope and hence a polyhedron.

b) S is a polyhedron.

c) S is not a polyhedron. Note that although S is defined as an intersection of halfspaces it is not a polyhedron, since we need infinitely many halfspaces.

d) $S = \{\boldsymbol{x} \in \mathbb{R}^n \mid -\mathbf{1}^n \leq \boldsymbol{x} \leq \mathbf{1}^n\}$, that is, a polyhedron.

e) S is a polyhedron. By squaring both sides of the inequality, it follows that $-2(\boldsymbol{x}^0 - \boldsymbol{x}^1)^{\mathrm{T}} \boldsymbol{x} \leq \|\boldsymbol{x}^1\|_2^2 - \|\boldsymbol{x}^0\|_2^2$, so S is in fact a halfspace.

f) S is a polyhedron. Similarly as in e) above it follows that S is the intersection of the halfspaces

$$-2(\boldsymbol{x}^0 - \boldsymbol{x}^i)^{\mathrm{T}}\boldsymbol{x} \leq \|\boldsymbol{x}^i\|_2^2 - \|\boldsymbol{x}^0\|_2^2, \quad i = 1, \dots, k.$$

■

Exercise 3.3 a) $\boldsymbol{x}^1$ is not an extreme point. b) $\boldsymbol{x}^2$ is an extreme point. This follows by checking the rank of the equality subsystem and then using Theorem 3.17. ■

Exercise 3.4 Let

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{A} \\ -\boldsymbol{A} \\ -\boldsymbol{I}^n \end{pmatrix}, \quad \boldsymbol{d} = \begin{pmatrix} \boldsymbol{b} \\ -\boldsymbol{b} \\ \boldsymbol{0}^n \end{pmatrix}.$$

Then P is defined by $\boldsymbol{Dx} \leq \boldsymbol{d}$. Further, P is nonempty, so let $\tilde{\boldsymbol{x}} \in P$. Now, if $\tilde{\boldsymbol{x}}$ is not an extreme point of P, then the rank of equality subsystem i lower than n. By using this it is possible to construct an $\boldsymbol{x}' \in P$ such that the rank of the equality subsystem of $\boldsymbol{x}'$ is at least one larger than the rank of the equality subsystem of $\tilde{\boldsymbol{x}}$. If this argument is used repeatedly we end up with an extreme point of P. ■

Exercise 3.5 We have that

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 0.5 \begin{pmatrix} 0 \\ 1 \end{pmatrix} + 0.5 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 0.5 \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

and since $(0,1)^{\mathrm{T}}, (1,0)^{\mathrm{T}} \in Q$ and $(1,1)^{\mathrm{T}} \in C$ we are done. ■

Exercise 3.6 Assume that $a_1, a_2, a_3, b \in \mathbb{R}$ satisfy

$$a_1x_1 + a_2x_2 + a_3x_3 \leq b, \quad \forall \boldsymbol{x} \in A, \tag{B.10}$$

$$a_1x_1 + a_2x_2 + a_3x_3 \geq b, \quad \forall \boldsymbol{x} \in B. \tag{B.11}$$

From (B.10) it follows that $a_2 = 0$ and that $a_3 \leq b$. Further, since $(1/n, n, 1)^{\mathrm{T}} \in B$ for all $n > 0$, from (B.11) we have that $a_3 \geq b$. Hence, it holds that $a_3 = b$. Since $(0,0,0)^{\mathrm{T}}, (1, n^2, n)^{\mathrm{T}} \in B$ for all $n \geq 0$, inequality (B.11) shows that $b \leq 0$ and $a_3 \geq 0$. Hence $a_2 = a_3 = b = 0$, and it follows that $H = \{\boldsymbol{x} \in \mathbb{R}^3\} \mid x_1 = 0\}$ is the only hyperplane that separates A and B. Finally, $A \subseteq H$ and $(0,0,0)^{\mathrm{T}} \in H \cap B$, so H meets both A and B. ■

Exercise 3.7 Let B be the intersection of all closed halfspaces in $\mathbb{R}^n$ containing A. It follows easely that $A \subseteq B$. In order to show that

$B \subseteq A$, show that $A^c \subseteq B^c$ by using the Separation Theorem 3.24. ■

Exercise 3.8 Assume that $P \neq \emptyset$. Then, by using Farkas' Lemma (Theorem 3.30), show that there exists a $\boldsymbol{p} \neq \mathbf{0}^m$ such that $\boldsymbol{p} \geq \mathbf{0}^m$ and $\boldsymbol{B}\boldsymbol{p} \geq \mathbf{0}^m$. From this it follows that P is unbounded and hence not compact. ■

Exercise 3.10 The function is strictly convex on $\mathbb{R}^2$. ■

Exercise 3.11 a) Not convex; b)–f) strictly convex. ■

Exercise 3.12 a)–f) Strictly convex. ■

Exercise 3.13 a)

$$f(x, y) = \frac{1}{2}(x, y)\begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + (3, -1)\begin{bmatrix} x \\ y \end{bmatrix}.$$

b) Yes. c) Yes. ■

Exercise 3.14 a) Non-convex; b) convex; c) non-convex; d) convex; e) convex. ■

Exercise 3.15 Yes. ■

Exercise 3.16 Yes. ■

Exercise 3.17 We will try to apply Definition 3.45. It is clear that the objective function can be written as the minimization of a (strictly) convex function. The constraints are analyzed thus: the first and third, taken together and applying also Example 3.37(c), describe a closed and convex set; the second and fourth constraint describes a (convex) polyhedron. By Proposition 3.3 we therefore are done. The answer is Yes. ■

Exercise 3.18 The first constraint is redundant; the feasible set hence is a nonempty polyedron. Regarding the objective function, it is defined only for positive x_1; the objective function is strictly convex on $\mathbb{R}_{++}$, since its second derivative there equals $1/x_1 > 0$ [cf. Theorem 3.41(b)]. We may extend the definition of $x_1 \ln x_1$ to a continuous (in fact convex) function, on the whole of $\mathbb{R}_+$ by defining $0 \ln 0 = 0$. With this classic extension, together with the constraint, we see that it is the problem of maximizing a convex function over a closed convex set. This is not a convex problem. The answer is No. ■

Chapter 4: An introduction to optimality conditions

Exercise 4.1 The claim is False. ■

Exercise 4.2 Investigating the Hessian matrix yields that $a \in (-4, 2)$ and $b \in \mathbb{R}$ implies that the objective function is strictly convex (in fact, strongly convex, because it is quadratic).

[*Note:* It would be a mistake to here perform a classic transformation, namely to observe that the problem is symmetric in x_1 and x_2 and utilize this to eliminate one of the variables through the identification $x_1^* = x_2^*$. Suppose we do so. We then reduce the problem to that of minimizing the one-dimensional function $x \mapsto (4+a)x^2 - 2x + b$ over $\mathbb{R}$. The condition for this function to be strictly convex, and therefore have a unique solution (see the above remark on strong convexity), is that $a > -4$, which is a milder condition than the above. However, if the value of a is larger than 2 the *original* problem has no solution! Indeed, suppose we look at the direction $\boldsymbol{x} \in \mathbb{R}^2$ in which $x_1 = -x_2 = p$. Then, the function $f(\boldsymbol{x})$ behaves like $(2-a)p^2 - 2p + b$ which clearly tends to minus infinity whenever $|p|$ tends to infinity, whenever $a > 2$. It is important to notice that the transformation works *when the problem has a solution*; otherwise, it is not.] ■

Exercise 4.3 Let $\rho(\boldsymbol{x}) := \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}$. Stationarity for ρ at $\boldsymbol{x}$ means that

$$\frac{2}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}(\boldsymbol{A}\boldsymbol{x} - \rho(\boldsymbol{x}) \cdot \boldsymbol{x}) = \boldsymbol{0}^n.$$

If $\boldsymbol{x}_i \neq \boldsymbol{0}^n$ is an eigenvector of $\boldsymbol{A}$, corresponding to the eigenvalue λ_i, then $\rho(\boldsymbol{x}_i) = \lambda_i$ holds. From the above two equations follow that for $\boldsymbol{x} \neq \boldsymbol{0}^n$ to be stationary it is both necessary and sufficient that $\boldsymbol{x}$ is an eigenvector.

The global minimum is therefore an arbitrary nonzero eigenvector, corresponding to the minimal eigenvalue λ_i of $\boldsymbol{A}$. ■

Exercise 4.4 (a) The proof is by contradiction, so suppose that $\bar{\boldsymbol{x}}$ is a local optimum, $\boldsymbol{x}^*$ is a global optimum, and that $f(\bar{\boldsymbol{x}}) < f(\boldsymbol{x}^*)$ holds. We first note that by the local optimality of $\bar{\boldsymbol{x}}$ and the affine nature of the constraints, it must hold that

$$\nabla f(\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{p} = \boldsymbol{0}^m, \qquad \text{for all vectors } \boldsymbol{p} \text{ with } \boldsymbol{A}\boldsymbol{p} = \boldsymbol{0}^m.$$

We will especially look at the vector $\boldsymbol{p} := \boldsymbol{x}^* - \bar{\boldsymbol{x}}$.

Next, by assumption, $f(\bar{\boldsymbol{x}}) < f(\boldsymbol{x}^*)$, which implies that $(\bar{\boldsymbol{x}}-\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{Q}(\bar{\boldsymbol{x}}-\boldsymbol{x}^*) < 0$ holds. We utilize this strict inequality together with the above to last establish that, for every $\gamma > 0$,

$$f(\bar{\boldsymbol{x}} + \gamma(\bar{\boldsymbol{x}} - \boldsymbol{x}^*)) < f(\bar{\boldsymbol{x}}),$$

which contradicts the local optimality of $\bar{\boldsymbol{x}}$. We are done. ∎

Exercise 4.5 Utilize the variational inequality characterization of the projection operation. ∎

Exercise 4.6 Utilize Proposition 4.22(b) for this special case of feasible set. We obtain the following necessary conditions for $\boldsymbol{x}^* \geq \mathbf{0}^n$ to be local minimum:

$$0 \leq x_j^* \perp \frac{\partial f(\boldsymbol{x}^*)}{\partial x_j} \geq 0, \qquad j = 1, 2, \dots, n,$$

where (for real values a and b) $a \perp b$ means the condition that $a \cdot b = 0$ holds. In other words, if $x_j^* > 0$ then the partial derivative of f at $\boldsymbol{x}^*$ with respect to x_j must be zero; conversely, if this partial derivative if non-zero then the value of x_j^* must be zero. (This is called complementarity.) ∎

Exercise 4.7 By a logarithmic transformation, we may instead maximize the function $f(\boldsymbol{x}) = \sum_{j=1}^n a_j \log x_j$. The optimal solution is

$$x_j^* = \frac{a_j}{\sum_{i=1}^n a_i}, \qquad j = 1, \dots, n.$$

(Check the optimality conditions for a problem defined over a simplex.)

We confirm that it is a unique optimal solution by checking that the objective function is strictly concave where it is defined. ∎

Chapter 5: Optimality conditions

Exercise 5.1 $(2, 1)^{\mathrm{T}}$ is a KKT point for this problem with KKT multipliers $(1, 0)^{\mathrm{T}}$. Since the problem is convex, this is also a globally optimal solution (cf. Theorem 5.45). Slater's CQ (and, in fact, LICQ as well) is verified. ∎

Exercise 5.2 (a) Feasible set of the problem consits of countably many isolated points $x_k = -\pi/2 + 2\pi k$, $k = 1, 2, \dots$, each of which is thus a locally optimal solution. The globally optimal solution is $x^* = -\pi/2$. KKT conditions are not satisfied at the points of local minimum and therefore they are not necessary for optimality in this problem. (The reason is of course that CQs are not verified.)

(b) It is easy to verify that FJ conditions are satisfied (as they should be, cf. Theorems 5.8 and 5.15).

(c) The point $(x, y) = (0, 0)$ is a FJ point, but it has nothing to do with points of local minimum. ■

Exercise 5.3 KKT system:

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} &\geq \boldsymbol{b}, \\ \boldsymbol{\lambda} &\geq \boldsymbol{0}, \\ \boldsymbol{c} - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} &= \boldsymbol{0}, \\ \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) &= 0. \end{aligned}$$

Combining the last two equations we obtain $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\lambda}$. ■

Exercise 5.4 (a) Clearly, two problems are equivelent. On the other hand, $\nabla\{\sum_{i=1}^m [h_i(\boldsymbol{x})]^2\} = 2\sum_{i=1}^m h_i(\boldsymbol{x})\nabla h_i(\boldsymbol{x}) = \boldsymbol{0}$ at every feasible solution. Therefore, MFCQ is violated at every feasible point of the problem (5.22) (even though Slater's CQ, LICQ, or at least MFCQ might hold for the original problem).

(b) The objective function is non-differentiable (well, only directionally differentiable). Therefore, we rewrite the problem as

$$\begin{aligned} &\text{minimize} \;\; z, \\ &\text{subject to} \;\; \begin{cases} f_1(\boldsymbol{x}) - z \leq 0, \\ f_2(\boldsymbol{x}) - z \leq 0, \end{cases} \end{aligned}$$

The problem verifies MFCQ (e.g., the direction $(\boldsymbol{0}, 1)^{\mathrm{T}} \in \overset{\circ}{G}(\boldsymbol{x}, z)$ for all feasible points $(\boldsymbol{x}, z)$. Therefore, KKT conditions are necessary for local optimality; these conditions are exaclty what we need. ■

Exercise 5.5 Problem is convex + CQ $\Longrightarrow$ need to find an arbitrary KKT-point. KKT system:

$$\begin{aligned} \boldsymbol{x} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} &= \boldsymbol{0}, \\ \boldsymbol{A}\boldsymbol{x} &= \boldsymbol{b} \end{aligned}$$

Therefore, $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = \boldsymbol{0}$, and $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = -\boldsymbol{b}$. Finally, $\boldsymbol{x} = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{b}$. ■

Exercise 5.6 (b) Show that KKT-multiplier λ is positive at every optimal solution. It means that $\sum_{j=1}^n x_j^2 = 1$ is satisfied at every optimal solution; use convexity to conclude that there may be only one optimal solution. ■

Exercise 5.7 (a) Locally and globally optimal solutions may be found using geometrical considerations; $(x, y) = (2, 0)$ gives us a local min, $(x, y) = (3/2, 3/2)$ is a globally optimal solution. KKT system incidently has two [in the space (x, y)] solutions, but at every point there are infinitely many KKT multipliers. Therefore, in this particular problem KKT-conditions are both necessary and sufficient for local optimality.

(b) The gradients of constraints are linearly dependent at every feasible point; thus LICQ is violated.

The fiasible set is a union of two convex sets $\mathcal{F}_1 = \{\, (x, y)^{\mathrm{T}} \mid y = 0,\, x - y \geq 0 \,\}$ and $\mathcal{F}_2 = \{\, (x, y)^{\mathrm{T}} \mid y \geq 0,\, x - y = 0 \,\}$. Thus we can solve two convex optimization problems to minimize f over $\mathcal{F}_1$, and to minimize f over $\mathcal{F}_2$; then simply choose the best solution.

(c) The feasible set may be split into 2^n convex parts $\mathcal{F}_I$, $I \subseteq \{\, 1, \ldots, n \,\}$, where

$$\begin{aligned} \boldsymbol{a}_i^{\mathrm{T}} &= b_i, \text{and } x_i \geq 0, \quad i \in I, \\ \boldsymbol{a}_i^{\mathrm{T}} &\geq b_i, \text{and } x_i = 0, \quad i \notin I. \end{aligned}$$

Thus we (in principle) have reduced the original non-convex problem that violates LICQ to 2^n convex problems. ■

Exercise 5.8 Use KKT-conditions (convex problem+Slater's CQ). $c \leq -1$. ■

Exercise 5.9 Slater's CQ $\Longrightarrow$ KKT conditions are necessary for optimality. Prove that $x_j^* > 0$; then

$$x_j^* = \frac{Dc_j}{\sum_{j=1}^n c_j}, \quad j = 1, \ldots, n.$$

■

Chapter 6: Lagrangian duality

Exercise 6.7

$$\begin{array}{llll} \lambda = 1 \implies & x_1 = 1, \; x_2 = 2, & \text{infeasible}, & q(1) = 6; \\ \lambda = 2 \implies & x_1 = 1, \; x_2 = 5/2, & \text{infeasible}, & q(2) = 43/4; \\ \lambda = 3 \implies & x_1 = 3, \; x_2 = 3, & \text{feasible}, & q(3) = 9. \end{array}$$

Further, $f(3,3) = 21$, so $43/4 \le f^* \le 21$. ∎

Chapter 8: Linear programming models

Exercise 8.1 (a) Introduce the new variables $\boldsymbol{y} \in \mathbb{R}^m$. Then the problem is equivalent to the linear program

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{m} y_i \\ \text{subject to} \quad & -\boldsymbol{y} \le \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \le \boldsymbol{y}, \\ & -\boldsymbol{1}^n \le \boldsymbol{x} \le \boldsymbol{1}^n. \end{aligned}$$

(b) Introduce the new variables $\boldsymbol{y} \in \mathbb{R}^m$ and $t \in \mathbb{R}$. Then the problem is equivalent to the linear program

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{m} y_i + t \\ \text{subject to} \quad & -\boldsymbol{y} \le \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \le \boldsymbol{y}, \\ & -t\boldsymbol{1}^n \le \boldsymbol{x} \le t\boldsymbol{1}^n. \end{aligned}$$

∎

Exercise 8.2 (a) Let

$$\boldsymbol{B} = \begin{pmatrix} -(\boldsymbol{v}^1)^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ -(\boldsymbol{v}^k)^{\mathrm{T}} & 1 \\ (\boldsymbol{w}^1)^{\mathrm{T}} & -1 \\ \vdots & \vdots \\ (\boldsymbol{w}^l)^{\mathrm{T}} & -1 \end{pmatrix}, \qquad \boldsymbol{x} = \begin{pmatrix} \boldsymbol{a} \\ b \end{pmatrix}.$$

Then from the rank assumption it follows that $\operatorname{rank} \boldsymbol{B} = n + 1$, which means that $\boldsymbol{x} \neq \boldsymbol{0}^{n+1}$ implies that $\boldsymbol{B}\boldsymbol{x} \neq \boldsymbol{0}^{k+l}$. Hence the problem can be solved by solving the linear program

$$\begin{aligned} \text{minimize} \quad & (\boldsymbol{0}^{n+1})^{\mathrm{T}} \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{B}\boldsymbol{x} \ge \boldsymbol{0}^{k+l}, \\ & (\boldsymbol{1}^{k+l})^{\mathrm{T}} \boldsymbol{B}\boldsymbol{x} = 1. \end{aligned}$$

(b) Let $\alpha = R^2 - \|\boldsymbol{x}^c\|_2^2$. Then the problem can be solved by solving the linear program

$$\begin{array}{lll} \text{minimize} & (\mathbf{0}^n)^{\mathrm{T}}\boldsymbol{x}^c + 0\alpha & \\ \text{subject to} & \|\boldsymbol{v}^i\|_2^2 - 2(\boldsymbol{v}^i)^{\mathrm{T}}\boldsymbol{x}^c \le \alpha, & i = 1,\dots,k, \\ & \|\boldsymbol{w}^i\|_2^2 - 2(\boldsymbol{w}^i)^{\mathrm{T}}\boldsymbol{x}^c \ge \alpha, & i = 1,\dots,l, \end{array}$$

and compute R as $R = \sqrt{\alpha + \|\boldsymbol{x}^c\|_2^2}$ (from the first set of inequalities in the LP above it follows that $\alpha + \|\boldsymbol{x}^c\|_2^2 \ge 0$ so this is well defined). ∎

Exercise 8.3 Since P is bounded there exists no $\boldsymbol{y} \neq \mathbf{0}^n$ such that $\boldsymbol{A}\boldsymbol{y} \le \mathbf{0}^m$. Hence there exist no feasible solution to the system

$$\begin{aligned} \boldsymbol{A}\boldsymbol{y} &\le \mathbf{0}^m, \\ \boldsymbol{d}^{\mathrm{T}}\boldsymbol{y} &= 1, \end{aligned}$$

which implies that $z > 0$ in every feasible solution to (8.9).

Further, let $(\boldsymbol{y}^*, z^*)$ be a feasible solution to (8.9). Then $z^* > 0$ and $\boldsymbol{x}^* = \boldsymbol{y}^*/z^*$ is feasible to (8.8), and $f(\boldsymbol{x}^*) = g(\boldsymbol{y}^*, z^*)$. Conversely, let $\boldsymbol{x}^*$ be a feasible solution to (8.8). Then by the hypothesis $\boldsymbol{d}^{\mathrm{T}}\boldsymbol{x}^* + \beta > 0$. Let $z^* = 1/(\boldsymbol{d}^{\mathrm{T}}\boldsymbol{x}^* + \beta)$ and $\boldsymbol{y}^* = z^*\boldsymbol{x}^*$. Then $(\boldsymbol{y}^*, z^*)$ is a feasible solution to (8.9) and $g(\boldsymbol{y}^*, z^*) = f(\boldsymbol{x}^*)$. These fact together imply the assertion. ∎

Exercise 8.4 The problem can be transformed into the standard form:

$$\begin{array}{llrrrrrrrl} \text{minimize} & z' = & x_1' & -5x_2^+ & +5x_2^- & -7x_3^+ & +7x_3^- & & & \\ \text{subject to} & & 5x_1' & -2x_2^+ & +2x_2^- & +6x_3^+ & -6x_3^- & -s_1 & & = 15, \\ & & 3x_1' & +4x_2^+ & -4x_2^- & -9x_3^+ & +9x_3^- & & & = 9, \\ & & 7x_1' & +3x_2^+ & -3x_2^- & +5x_3^+ & -5x_3^- & & +s_2 & = 23, \\ & & x_1', & x_2^+, & x_2^-, & x_3^+, & x_3^-, & s_1, & s_2 & \ge 0, \end{array}$$

where $x_1' = x_1 + 2$, $x_2 = x_2^+ - x_2^-$, $x_3 = x_3^+ - x_3^-$, and $z' = z - 2$. ∎

Exercise 8.5 (a) The first equality constraint gives that

$$x_3 = \frac{1}{6}(11 - 2x_1 - 4x_2).$$

Now, by substituting x_3 with this exression in the objective function and the second equality constraint the problem is in standard form and x_3 is eliminated.

(b) If $x_3 \geq 0$, then we must add the constraint $(11-2x_1-4x_2)/6 \geq 0$ to the problem. But this is an inequality, so in order to transform the problem into standard form we must add a slackvarible. ■

Exercise 8.6 Assume that the column in the constraint matrix corresponding to the variable x_j^+ is $\boldsymbol{a}_j$. Then the column in the constraint matrix corresponding to the variable x_j^- is $-\boldsymbol{a}_j$. The statement then follows from the definition of basic feasible solution, since $\boldsymbol{a}_j$ and $-\boldsymbol{a}_j$ are linearly dependent. ■

Exercise 8.7 Let P be the set of feasible solutions to (8.10) and Q be the set of feasible solutions to (8.11). Obviously $P \subseteq Q$. In order to show that $Q \subseteq P$ assume that there exists an $\boldsymbol{x} \in Q$ such that $\boldsymbol{x} \notin P$ and derive a contradiction. ■

Chapter 9: The simplex method

Exercise 9.1 The phase I problem becomes

$$\begin{aligned}
\text{minimize} \quad & w = a_1 + a_1 \\
\text{subject to} \quad & -3x_1 - 2x_2 + x_3 - s_1 + a_1 = 3, \\
& x_1 + x_2 - 2x_3 - s_2 + a_2 = 1, \\
& x_1, x_2, x_3, s_1, s_2, a_1, a_2 \geq 0.
\end{aligned}$$

From the equality constraints it follows that $a_1+a_2 \geq 4$ for all $x_1, x_2, x_3, s_1, s_2 \geq 0$. Hence, in particular, it follows that $w \geq 4$ for all feasible solutions to the phase I problem, which means that the original problem is infeasible.
■

Exercise 9.2 (a) The standard form is given by

$$\begin{aligned}
\text{minimize} \quad & 3x_1 + 2x_2 + x_3 \\
\text{subject to} \quad & 2x_1 + x_3 - s_1 = 3, \\
& 2x_1 + 2x_2 + x_3 = 5, \\
& x_1, x_2, x_3, s_1 \geq 0.
\end{aligned}$$

By solving the phase I problem with the Simplex algorithm we get the feasible basis $\boldsymbol{x}_B = (x_1, x_2)^{\mathrm{T}}$. Then by solving the phase II problem with the Simplex algorithm we get the optimal solution $\boldsymbol{x}^* = (x_1, x_2, x_3)^{\mathrm{T}} = (0, 1, 3)^{\mathrm{T}}$.

(b) No, the set of all optimal solution is given by the set

$$\{\boldsymbol{x} \in \mathbb{R}^3 \mid \lambda(0,1,3)^{\mathrm{T}} + (1-\lambda)(0,0,5)^{\mathrm{T}}; \quad \lambda \in [0,1]\}.$$

■

Exercise 9.3 The reduced cost for all the variables except for x_j must be greater than or equal to 0. Hence it follows that the current basis is optimal to the problem that arises if x_j is fixed to zero. The assertion then follows from the fact that the current basis is non-degenerate. ■

Chapter 10: LP duality and sensitivity analysis

Exercise 10.1 The linear programming dual is given by

$$\begin{array}{llrrrcr}
\text{minimize} & 11y_1 & +23y_2 & +12y_3 & & \\
\text{subject to} & 4y_1 & +3y_2 & +7y_3 & \geq & 6, \\
& 3y_1 & +2y_2 & +4y_3 & \geq & -3, \\
& -8y_1 & +7y_2 & +3y_3 & \leq & -2, \\
& 7y_1 & +6y_2 & +2y_3 & = & 5, \\
& & y_2 & & \leq & 0, \\
& & & y_3 & \geq & 0.
\end{array}$$

■

Exercise 10.2 (a) The linear programming dual is given by

$$\begin{array}{llllcl}
\text{maximize} & \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}^1 & +\boldsymbol{l}^{\mathrm{T}}\boldsymbol{y}^2 & +\boldsymbol{u}^{\mathrm{T}}\boldsymbol{y}^3 & & \\
\text{subject to} & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}^1 & +\boldsymbol{I}^n\boldsymbol{y}^2 & +\boldsymbol{I}^n\boldsymbol{y}^3 & = & \boldsymbol{c}, \\
& & \boldsymbol{y}^2 & & \geq & \boldsymbol{0}^n, \\
& & & \boldsymbol{y}^3 & \leq & \boldsymbol{0}^n.
\end{array}$$

(b) A feasible solution to the linear programming dual is given by

$$\begin{aligned}
\boldsymbol{y}^1 &= \boldsymbol{0}^m, \\
\boldsymbol{y}^2 &= (\max\{0, c_1\}, \dots, \max\{0, c_n\})^{\mathrm{T}}, \\
\boldsymbol{y}^3 &= (\min\{0, c_1\}, \dots, \min\{0, c_n\})^{\mathrm{T}}.
\end{aligned}$$

■

Exercise 10.3 First, check that $\boldsymbol{y} = (\boldsymbol{B}^{-1})^{\mathrm{T}}\boldsymbol{c}_B$ is feasible to the LP dual problem. Then show that $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}$ equals the optimal objective function value for the primal problem. The assertion then follows from the Weak Duality Theorem. ■

Exercise 10.4 Use the Weak and Stong Duality Theorems. ■

Exercise 10.5 The LP dual is infeasible. Hence, from the Weak and Strong Duality Theorems it follows that the primal problem is either infesible or unbounded. ■

Exercise 10.6 By using the Strong Duality Theorem we get the following polyhedron:

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{x} &\geq \boldsymbol{b}, \\
\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} &\leq \boldsymbol{c}, \\
\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} &= \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}, \\
\boldsymbol{x} &\geq \boldsymbol{0}^n, \\
\boldsymbol{y} &\leq \boldsymbol{0}^m.
\end{aligned}$$

■

Exercise 10.7 From the Strong Duality Theorem it follows that $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}^* = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}^*$. Use this to establish the statement. ■

Exercise 10.8 The dual problem only contains two varibles and hence can be solved graphically. We get the optimal solution $\boldsymbol{y}^* = (-2, 0)^{\mathrm{T}}$. The complementary slackness conditions then implies that $x_1 = x_2 = x_3 = x_5 = 0$. Hence, let $\boldsymbol{x}_B = (x_4, x_6)^{\mathrm{T}}$. The optimal solution is $\boldsymbol{x}^* = (x_1, x_2, x_3, x_4, x_5, x_6)^{\mathrm{T}} = (0, 0, 0, 3, 0, 1)^{\mathrm{T}}$. ■

Exercise 10.9 From the complementary slackness conditions and the fact that $c_1/a_1 \geq \cdots \geq c_n/a_n$ it follows that

$$\begin{aligned}
u &= \frac{c_r}{a_r}, \\
y_j &= c_j - \frac{c_r}{a_r}a_j, \quad j = 1, \ldots, r-1, \\
y_j &= 0, \quad j = r, \ldots, n,
\end{aligned}$$

is a dual feasible solution which together with the given primal solution fulfil the LP primal-dual optimality conditions. ■

Exercise 10.14 The basis $\boldsymbol{x}_B = (x_1, x_2)^{\mathrm{T}}$ is optimal as long as $c_3 \leq 5$ and $c_4 \geq 8$. ■

Exercise 10.15 b) The basis $\boldsymbol{x}_B = (x_1, x_3)^{\mathrm{T}}$ is optimal for all $\delta \geq -6.5$. c) The basis $\boldsymbol{x}_B = (x_1, x_3)$ is not primal feasible for $\delta = -7$, but it is dual feasible, so by using the Dual Simplex method it follows that $\boldsymbol{x}_B = (x_1, x_5)^{\mathrm{T}}$ is an optimal basis. ■

Chapter 11: Unconstrained optimization

Exercise 11.2 The directional derivative is $13 > 0$; the answer is No. ■

Exercise 11.3 (a) The search direction is not a descent direction, for example because the Hessian matrix is indefinite or negative definite. (b) The linear system is unsolvable, for example because the Hessian matrix is indefinite. [*Note:* Even for indefinite Hessians, the search direction might exist for *some* right-hand sides.] (c) Use the Levenberg–Marquardt modification. ■

Exercise 11.4 Let $y_1 := x_1 - 2$ and $y_2 := \sqrt{5}(x_2 + 6)$. We then get $f(\boldsymbol{x}) = g(\boldsymbol{y}) = y_1^2 + y_2^2$. At every $\boldsymbol{y} \in \mathbb{R}^2$ the negative gradient points towards the optimum! ■

Exercise 11.5 (a) $\boldsymbol{x}_1 = (1/2, 1)^{\mathrm{T}}$. (b) The Hessian matrix is

$$\nabla^2 f(\boldsymbol{x}_1) = \begin{pmatrix} 10 & -4 \\ -4 & 2 \end{pmatrix}.$$

The answer is Yes. (c) The answer is Yes. ■

Exercise 11.6 (a) $\boldsymbol{x}_1 = (2, 1/2)^{\mathrm{T}}$. (b) The answer is No. The gradient is zero. (c) The answer is Yes. ■

Exercise 11.7 (a) (b) $\mu \in (0, 0.6)$. ■

Exercise 11.8 (a) $\boldsymbol{f}(\boldsymbol{x}_0) = (-1, -2)^{\mathrm{T}} \Longrightarrow \|\boldsymbol{f}(\boldsymbol{x}_0)\| = \sqrt{5}$; $x_1 = (4/3, 2/3)^{\mathrm{T}} \Longrightarrow \|\boldsymbol{f}(\boldsymbol{x}_1)\| = 16/27$. (b) If $\boldsymbol{f}$ is the gradient of a C^2 function $f : \mathbb{R}^n \mapsto \mathbb{R}$ we obtain that $\nabla \boldsymbol{f} = \nabla^2 f$, that is, Newton's method for unconstrained optimization is obtained. ∎

Exercise 11.9 (a) $\boldsymbol{x}^* = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{b}$. (b) The objective function is convex, since the Hessian is $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ (which is *always* positive semi-definite; check!). Therefore, the normal solution in (a) is globally optimal. ∎

Exercise 11.12 (a) We have that

$$\nabla f(\boldsymbol{y}) + \gamma(\boldsymbol{y} - \boldsymbol{x}_k) = \boldsymbol{0}^n \iff \boldsymbol{Q}\boldsymbol{y} + \boldsymbol{q} + \gamma(\boldsymbol{y} - \boldsymbol{x}_k) = \boldsymbol{0}^n \iff$$
$$(\boldsymbol{Q} + \gamma\boldsymbol{I}^n)\boldsymbol{y} = \gamma\boldsymbol{x}_k - \boldsymbol{q} \iff (\boldsymbol{Q} + \gamma\boldsymbol{I}^n)(\boldsymbol{y} - \boldsymbol{x}_k).$$

Further,

$$(\boldsymbol{Q} + \boldsymbol{\gamma}\boldsymbol{I}^n)(\boldsymbol{y} - \boldsymbol{x}_k) = \gamma\boldsymbol{x}_k - \boldsymbol{q} - (\boldsymbol{Q} + \gamma\boldsymbol{I}^n)\boldsymbol{x}_k = -(\boldsymbol{Q}\boldsymbol{x}_k + \boldsymbol{q}).$$

(b) If $\{\boldsymbol{x}_k\}$ converges to $\boldsymbol{x}_\infty$ then $\{\boldsymbol{p}_k\} = \{\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\}$ must converge to zero. From the updating formula we obtain that $\boldsymbol{p}_k = (\boldsymbol{Q} + \gamma\boldsymbol{I}^n)^{-1}\nabla f(\boldsymbol{x}_k)$ for every k. The sequence $\{\nabla f(\boldsymbol{x}_k)\}$ converges to $\nabla f(\boldsymbol{x}_\infty)$, since $f \in C^1$. If $\nabla f(\boldsymbol{x}_\infty) \neq \boldsymbol{0}^n$ it would hold that $\{\boldsymbol{p}_k\}$ would converge to $(\boldsymbol{Q} + \gamma\boldsymbol{I}^n)^{-1}\nabla f(\boldsymbol{x}_\infty) \neq \boldsymbol{0}^n$, since $(\boldsymbol{Q} + \gamma\boldsymbol{I}^n)^{-1}$ is positive definite when $\boldsymbol{Q} + \gamma\boldsymbol{I}^n$ is. This leads to a contradiction. Hence, $\nabla f(\boldsymbol{x}_\infty) = \boldsymbol{0}^n$. Since f is convex $\boldsymbol{x}_\infty$ is a global minimum of f over $\mathbb{R}^n$. ∎

Exercise 11.13 Case I: $\{\nabla f(\boldsymbol{x}_k)\} \to \boldsymbol{0}^n$; $\{\boldsymbol{x}_k\}$ and $\{f(\boldsymbol{x}_k)\}$ diverge.
Example: $f(x) = -\log x$; $\{x_k\} \to \infty$; $\{f(x_k)\} \to -\infty$; $\{f'(x_k)\} \to 0$.

Case II: $\{\nabla f(\boldsymbol{x}_k)\} \to \boldsymbol{0}^n$; $\{\boldsymbol{x}_k\}$ diverges; $\{f(\boldsymbol{x}_k)\}$ converges.
Example: $f(x) = 1/x$; $\{x_k\} \to \infty$; $\{f(x_k)\} \to 0$; $\{f(x_k)\} \to 0$.

Case III: $\{\nabla f(\boldsymbol{x}_k)\} \to \boldsymbol{0}^n$; $\{\boldsymbol{x}_k\}$ is bounded; $\{f(\boldsymbol{x}_k)\}$ is bounded.
Example: $f(x) = \frac{1}{3}x^3 - x$; $x_k = \begin{cases} 1 + 1/k, & k \text{ even} \\ -1 - 1/k & k \text{ odd} \end{cases}$
$\{x_k\}$ has two limit points: ± 1; $\{f(x_k)\}$ has two limit points: $\pm 2/3$.

Case IV: $\{\nabla f(\boldsymbol{x}_k)\} \to \boldsymbol{0}^n$; $\{\boldsymbol{x}_k\}$ is bounded; $\{f(\boldsymbol{x}_k)\}$ converges.
Example: $f(x) = x^2 - 1$; x_k as above; $\{f(x_k)\} \to 0$.

Case V: $\{\nabla f(\boldsymbol{x}_k)\} \to \boldsymbol{0}^n$; $\{\boldsymbol{x}_k\}$ and $\{f(\boldsymbol{x}_k)\}$ converge.
Example: f as in Case IV; $x_k = 1 + 1/k$. ∎

Chapter 12: Optimization over convex sets

Exercise 12.2 (b) $\boldsymbol{x}_1 = (12/5, 4/5)^{\mathrm{T}}$; UBD $= f(\boldsymbol{x}_1) = 8$. The LP problem defined at $\boldsymbol{x}_0$ gives LBD $= 0$. Hence, $f^* \in [0, 8]$. ∎

References

[Aba67] J. Abadie, *On the Kuhn–Tucker theorem*, in Nonlinear Programming (NATO Summer School, Menton, 1964), North-Holland, Amsterdam, 1967, pp. 19–36.

[Arm66] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific Journal of Mathematics, 16 (1966), pp. 1–3.

[AHU58] K. J. Arrow, L. Hurwicz, and H. Uzawa, eds., *Studies in Linear and Non-Linear Programming*, Stanford University Press, Stanford, CA, 1958.

[AHU61] K. J. Arrow, L. Hurwicz, and H. Uzawa, *Constraint qualifications in maximization problems*, Naval Research Logistics Quarterly, 8 (1961), pp. 175–191.

[Avr76] M. Avriel, *Nonlinear Programming: Analysis and Methods*, Prentice Hall Series in Automatic Computation, Prentice Hall, Englewood Cliffs, NJ, 1976.

[Ban22] S. Banach, *Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales*, Fundamenta Mathematicae, 3 (1922), pp. 133–181.

[Bar71] R. H. Bartels, *A stabilization of the simplex method*, Numerische Mathematik, 16 (1971), pp. 414–434.

[BaG69] R. H. Bartels and G. H. Golub, *The simplex method of linear programming using LU-decomposition*, Communications of the ACM, 12 (1969), pp. 266–268 and 275–278.

[BSS93] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, New York, NY, second ed., 1993.

[Ben62] J. F. Benders, *Partitioning procedures for solving mixed variables programming problems*, Numerische Mathematik, 4 (1962), pp. 238–252.

[Ber99] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Bellmont, MA, second ed., 1999.

[Ber04] ———, *Lagrange multipliers with optimal sensitivity properties in constrained optimization*, Report LIDS 2632, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 2004.

[BNO03] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, Belmont, MA, 2003.

[BeT89] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice Hall, London, U.K., 1989.

[BeT00] D. P. Bertsekas and J. N. Tsitsiklis, *Gradient convergence in gradient methods with errors*, SIAM Journal on Optimization, 10 (2000), pp. 627–642.

[Bla77] R. G. Bland, *New finite pivoting rules for the simplex method*, Mathematics of Operations Research, 2 (1977), pp. 103–107.

[BlO72] E. Blum and W. Oettli, *Direct proof of the existence theorem in quadratic programming*, Operations Research, 20 (1972), pp. 165–167.

[BGLS03] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical Optimization: Theoretical and Practical Aspects*, Universitext, Springer-Verlag, Berlin, 2003. Translated from the original French edition, published by Springer-Verlag 1997.

[BoS00] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer Series in Operations Research, Springer-Verlag, New York, NY, 2000.

[BoL00] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, CMS Books in Mathematics, Springer-Verlag, New York, NY, 2000.

[BHM77] S. P. Bradley, A. C. Hax, and T. L. Magnanti, *Applied Mathematical Programming*, Addison-Wesley, Reading, MA, 1977.

[Bre73] R. P. Brent, *Algorithms for Minimization Without Derivatives*, Prentice Hall Series in Automatic Computation, Prentice Hall, Englewood Cliffs, NJ, 1973. Reprinted by Dover Publications, Inc., Mineola, NY, 2002.

[Bro09] L. E. J. Brouwer, *On continuous vector distributions on surfaces*, Amsterdam Proceedings, 11 (1909).

[Bro12] ———, *Über Abbildung von Mannigfaltigkeiten*, Mathematische Annalen, 71 (1912), pp. 97–115.

[Bro70] C. G. Broyden, *The convergence of single-rank quasi-Newton methods*, Mathematics of Computation, 24 (1970), pp. 365–382.

[BGIS95] R. Burachik, L. M. G. Drummond, A. N. Iusem, and B. F. Svaiter, *Full convergence of the steepest descent method with inexact line searches*, Optimization, 32 (1995), pp. 137–146.

[BuF91] J. V. Burke and M. C. Ferris, *Characterization of solution sets of convex programs*, Operations Research Letters, 10 (1991), pp. 57–60.

[Car07] C. Carathéodory, *Über den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehmen*, Mathematische Annalen, 64 (1907), pp. 95–115.

[Car11] ——, *Über den Variabilitätsbereich der Fourier'schen Konstanten von positiven harmonischen Funktionen*, Rendiconti del Circolo Matematico di Palermo, 32 (1911), pp. 193–217.

[Cau1847] A. Cauchy, *Méthode générale pour la résolution des systèmes d'équations simultanées*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 25 (1847), pp. 536–538.

[Cha52] A. Charnes, *Optimality and degeneracy in linear programming*, Econometrica, 20 (1952), pp. 160–170.

[Chv83] V. Chvátal, *Linear Programming*, Freeman, New York, 1983.

[CGT00] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust-Region Methods*, vol. 1 of MPS/SIAM Series on Optimization, SIAM and Mathematical Programming Society, Philadelphia, PA, 2000.

[Dan51] G. B. Dantzig, *Maximization of a linear function of variables subject to linear inequalities*, in Activity Analysis of Production and Allocation, T. C. Koopmans, ed., New York, NY, 1951, John Wiley & Sons, pp. 339–347.

[Dan53] ——, *Computational algorithm of the revised simplex method*, Report RM 1266, The Rand Corporation, Santa Monica, CA, 1953.

[Dan57] ——, *Concepts, origins, and use of linear programming*, in Proceedings of the First International Conference on Operational Research, Oxford, 1957, M. Davies, R. T. Eddison, and T. Page, eds., London, U.K., 1957, The English Universities Press, pp. 100–108.

[Dan63] ——, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.

[DOW55] G. B. Dantzig, A. Orden, and P. Wolfe, *The generalized simplex method for minimizing a linear form under linear inequality restraints*, Pacific Journal of Mathematics, 5 (1955), pp. 183–195.

[DaT97] G. B. Dantzig and M. N. Thapa, *Linear programming 1: Introduction*, Springer-Verlag, New York, NY, 1997.

[DaT03] ——, *Linear programming 2: Theory and Extensions*, Springer-Verlag, New York, NY, 2003.

[DaW60] G. B. Dantzig and P. Wolfe, *Decomposition principle for linear programs*, Operations Research, 8 (1960), pp. 101–111.

[Dav59] W. C. Davidon, *Variable metric method for minimization*, Report ANL-5990 Rev, Argonne National Laboratories, Argonne, IL, 1959. Also published in SIAM Journal on Optimization, 1 (1991), pp. 1–17.

[DeF49] B. De Finetti, *Sulla stratificazioni convesse*, Annali di Matematica Pura ed Applicata, 30 (1949), pp. 173–183.

[DeS83] J. E. Dennis and R. E. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice Hall, Englewood Cliffs, NJ, 1983.

[DiJ79] Y. M. I. Dirickx and L. P. Jennergren, *Systems Analysis by Multilevel Methods: With Applications to Economics and Management*, vol. 6 of International Series on Applied Systems Analysis, John Wiley & Sons, Chichester, U.K., 1979.

[DuH78] J. C. Dunn and S. Harshbarger, *Conditional gradient algorithms with open loop step size rules*, Journal of Mathematical Analysis and Applications, 62 (1978), pp. 432–444.

[Eav71] B. C. Eaves, *On quadratic programming*, Management Science, 17 (1971), pp. 698–711.

[EHL01] T. F. Edgar, D. M. Himmelblau, and L. S. Lasdon, *Optimization of Chemical Processes*, McGraw-Hill, New York, NY, second ed., 2001.

[Erm66] Y. M. Ermol'ev, *Methods for solving nonlinear extremal problems*, Kibernetika, 2 (1966), pp. 1–17. Translated in Cybernetics, 2 (1966), pp. 1–14.

[Eva70] J. P. Evans, *On constraint qualifications in nonlinear programming*, Naval Research Logistics Quarterly, 17 (1970), pp. 281–286.

[Eve63] H. Everett, III, *Generalized Lagrange multiplier method for solving problems of optimum allocation of resources*, Operations Research, 11 (1963), pp. 399–417.

[Fac95] F. Facchinei, *Minimization of SC^1 functions and the Maratos effect*, Operations Research Letters, 17 (1995), pp. 131–137.

[Fal67] J. E. Falk, *Lagrange multipliers and nonlinear programming*, Journal of Mathematical Analysis and Applications, 19 (1967), pp. 141–159.

[Far1902] J. Farkas, *Über die Theorie der einfachen Ungleichungen*, Journal für die Reine und Angewandte Mathematik, 124 (1902), pp. 1–24.

[Fen51] W. Fenchel, *Convex cones, sets and functions*, mimeographed lecture notes, Princeton University, Princeton, NY, 1951.

[Fia83] A. V. Fiacco, *Introduction to sensitivity and stability analysis in nonlinear programming*, vol. 165 of Mathematics in Science and Engineering, Academic Press Inc., Orlando, FL, 1983.

[FiM68] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley & Sons,, New York, NY, 1968. Also published as volume 4 in the Classics in Applied Mathematics Series, SIAM, Philadelphia, PA, 1990.

[Fis81] M. L. Fisher, *The Lagrangian relaxation method for solving integer programming problems*, Management Science, 27 (1981), pp. 1–18.

[Fis85] ———, *An applications oriented guide to Lagrangian relaxation*, Interfaces, 15 (1985), pp. 10–21.

[Fle70] R. Fletcher, *A new approach to variable metric algorithms*, Computer Journal, 13 (1970), pp. 317–322.

[Fle87] ———, *Practical Methods of Optimization*, John Wiley & Sons, Chichester, U.K., second ed., 1987.

[FLT02] R. Fletcher, S. Leyffer, and P. L. Toint, *On the global convergence of a filter-SQP algorithm*, SIAM Journal on Optimization, 13 (2002), pp. 44–59.

[FlP63] R. Fletcher and M. J. D. Powell, *A rapidly convergent descent method for minimization*, Computer Journal, 6 (1963), pp. 163–168.

[FlR64] R. Fletcher and C. M. Reeves, *Function minimization by conjugate gradients*, Computer Journal, 7 (1964), pp. 149–154.

[FrW56] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Research Logistics Quarterly, 3 (1956), pp. 95–110.

[Geo74] A. M. Geoffrion, *Lagrangean relaxation for integer programming. Approaches to integer programming*, Mathematical Programming Study, 2 (1974), pp. 82–114.

[Gil66] E. G. Gilbert, *An iterative procedure for computing the minimum of a quadratic form on a convex set*, SIAM Journal on Control, 4 (1966), pp. 61–80.

[GiM73] P. E. Gill and W. Murray, *A numerically stable form of the simplex algorithm*, Linear Algebra and Its Applications, 7 (1973), pp. 99–138.

[GMSW89] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright, *A practical anti-cycling procedure for linearly constrained optimization*, Mathematical Programming, 45 (1989), pp. 437–474.

[Gol70] D. Goldfarb, *A family of variable-metric methods derived by variational means*, Mathematics of Computation, 24 (1970), pp. 23–26.

[Gol64] A. A. Goldstein, *Convex programming in Hilbert space*, Bulletin of the American Mathematical Society, 70 (1964), pp. 709–710.

[GrD03] A. Granas and J. Dugundji, *Fixed Point Theory*, Springer Monographs in Mathematics, Springer-Verlag, New York, NY, 1969.

[Gui69] M. Guignard, *Generalized Kuhn–Tucker conditions for mathematical programming problems in a Banach space*, SIAM Journal on Control, 7 (1969), pp. 232–241.

[Had10] J. Hadamard, *Sur quelques applications de l'indice de Kronecker*, in Introduction à la théorie des fonctions d'une variable, J. Tannary, ed., vol. 2, Hermann, Paris, 1910, pp. 875–915.

[Han75] S. P. Han, *Penalty Lagrangian methods in a quasi-Newton approach*, Report TR 75-252, Computer Science, Cornell University, Ithaca, NY, 1975.

[HaH96] G. K. Hauer and H. M. Hoganson, *Tailoring a decomposition method to a large forest management scheduling problem in northern Ontario*, INFOR, 34 (1996), pp. 209–231.

[HLV87] D. W. Hearn, S. Lawphongpanich, and J. A. Ventura, *Restricted simplicial decomposition: Computation and extensions*, Mathematical Programming Study, 31 (1987), pp. 99–118.

[HWC74] M. Held, P. Wolfe, and H. P. Crowder, *Validation of subgradient optimization*, Mathematical Programming, 6 (1974), pp. 62–88.

[HiL93] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, vol. 305–306 of Grundlehren der mathematischen Wissenschaften, Springer-Verlag, Berlin, 1993.

[Hof53] A. Hoffman, *Cycling in the simplex algorithm*, Report 2974, National Bureau of Standards, Gaithersburg, MD, 1953.

[Ius03] A. N. Iusem, *On the convergence properties of the projected gradient method for convex optimization*, Computational and Applied Mathematics, 22 (2003), pp. 37–52.

[Joh48] F. John, *Extremum problems with inequalities as subsidiary conditions*, in Studies and Essays Presented to R. Courant on his 60th Birthday, January 8, 1948, Interscience Publishers, Inc., New York, NY, 1948, pp. 187–204.

[JoM74] L. A. Johnson and D. S. Montgomery, *Operations Research in Production Planning, Scheduling and Inventory Control*, John Wiley & Sons, New York, NY, 1974.

[Kar84a] N. Karmarkar, *A new polynomial-time algorithm for linear programming*, in Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing, New York, 1984, The Association for Computing Machinery, pp. 302–311.

[Kar84b] ———, *A new polynomial-time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.

[Kha79] L. G. Khachiyan, *A polynomial algorithm in linear programming*, Doklady Akademii Nauk SSSR, 244 (1979), pp. 1093–1096.

[Kha80] ———, *Polynomial algorithms in linear programming*, Akademiya Nauk SSSR. Zhurnal Vychislitel'noĭ Matematiki i Matematicheskoĭ Fiziki, 20 (1980), pp. 51–68.

[KlM72] V. Klee and G. J. Minty, *How good is the simplex algorithm?*, in Inequalities, III, New York, NY, 1972, Academic Press, pp. 159–175.

[KLT03] T. G. Kolda, R. M. Lewis, and V. Torczon, *Optimization by direct search: New perspectives on some classical and modern methods*, SIAM Review, 45 (2003), pp. 385–482.

[Kre78] E. Kreyszig, *Introductory Functional Analysis with Applications*, John Wiley & Sons, New York, NY, 1978.

[KuT51] H. W. Kuhn and A. W. Tucker, *Nonlinear programming*, in Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950, Berkeley and Los Angeles, CA, 1951, University of California Press, pp. 481–492.

[LaP05] T. Larsson and M. Patriksson, *Global optimality conditions for discrete and nonconvex optimization—with applications to Lagrangian heuristics and column generation*, tech. rep., Department of Mathematics, Chalmers University of Technology, Gothenburg, Sweden, 2005. To appear in *Operations Research*.

[LPS96] T. Larsson, M. Patriksson, and A.-B. Strömberg, *Conditional subgradient optimization—theory and applications*, European Journal of Operational Research, 88 (1996), pp. 382–403.

[LPS99] ———, *Ergodic, primal convergence in dual subgradient schemes for convex programming*, Mathematical Programming, 86 (1999), pp. 283–312.

[Las70] L. S. Lasdon, *Optimization Theory for Large Systems*, Macmillan, New York, NY, 1970.

[Law76] E. Lawler, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart and Winston, New York, NY, 1976.

[LRS91] J. K. Lenstra, A. H. G. Rinnooy Kan, and A. Schrijver, eds., *History of Mathematical Programming. A Collection of Personal Reminiscences*, North-Holland, Amsterdam, 1991.

[LeP66] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, USSR Computational Mathematics and Mathematical Physics, 6 (1966), pp. 1–50.

[Lip1877] R. Lipschitz, *Lehrbuch der Analysis*, Cohn & Sohn, Leipzig, 1877.

[Lue84] D. G. Luenberger, *Linear and Nonlinear Programming*, Addison Wesley, Reading, MA, second ed., 1984.

[Man65] O. L. Mangasarian, *Pseudo-convex functions*, SIAM Journal on Control, 3 (1965), pp. 281–290.

[Man69] ———, *Nonlinear Programming*, McGraw-Hill, New York, NY, 1969. Also published as volume 10 in the Classics in Applied Mathematics Series, SIAM, Philadelphia, PA, 1994.

[Man88] ———, *A simple characterization of solution sets of convex programs*, Operations Research Letters, 7 (1988), pp. 21–26.

[MaF67] O. L. Mangasarian and S. Fromovitz, *The Fritz John necessary optimality conditions in the presence of equality and inequality constraints*, Journal of Mathematical Analysis and Applications, 17 (1967), pp. 37–47.

[Mar78] N. Maratos, *Exact penalty function algorithms for finite dimensional and control optimization problems*, PhD thesis, Imperial College of Science and Technology, University of London, London, U.K., 1978.

[Min10] H. Minkowski, *Geometrie der Zahlen*, Teubner, Leipzig, 1910.

[Min11] ———, *Gesammelte Abhandlungen*, vol. II, Teubner, Leipzig, 1911, ch. Theorie der knovexen Körper, Insbesondere Begründung ihres Ober flächenbegriffs.

[Mot36] T. Motzkin, *Beiträge zur Theorie del linearen Ungleichungen*, Azriel, Israel, 1936.

[Mur83] K. G. Murty, *Linear Programming*, John Wiley & Sons, New York, NY, 1983.

[Mur95] ———, *Operations Research: Deterministic Optimization Models*, Prentice Hall, Englewood Cliffs, NJ, 1995.

[Nas50] J. F. Nash, Jr., *Equilibrium points in n-person games*, Proceedings of the National Academy of Sciences of the United States of America, 36 (1950), pp. 48–49.

[Nas51] ———, *Non-cooperative games*, Annals of Mathematics, 54 (1951), pp. 286–295.

[NaS96] S. G. Nash and A. Sofer, *Linear and Nonlinear Programming*, MacGraw-Hill, Singapore, 1996.

[NeW88] G. L. Nemhauser and L. Wolsey, *Integer and Combinatorial Optimization*, Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, New York, NY, 1988.

[New1687] I. S. Newton, *Philosophiae Naturalis Principia Mathematica*, London, U.K., 1687.

[NoW99] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research, Springer-Verlag, New York, NY, 1999.

[Orc54] W. Orchard-Hays, *Background, development and extensions of the revised simplex method*, Report RM 1433, The Rand Corporation, Santa Monica, CA, 1954.

[Pad99] M. Padberg, *Linear Optimization and Extensions*, no. 12 in Algorithms and Combinatorics, Springer-Verlag, Berlin, second ed., 1999.

[PaT91] E. R. Panier and A. L. Tits, *Avoiding the Maratos effect by means of a nonmonotone line search, I. General constrained problems*, SIAM Journal on Numerical Analysis, 28 (1991), pp. 1183–1195.

[PaS82] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Prentice Hall, Englewood Cliffs, 1982.

[Pat98] M. Patriksson, *Nonlinear Programming and Variational Inequalities: A Unified Approach*, vol. 23 of Applied Optimization, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[PoR69] E. Polak and G. Ribière, *Note sur la convergence de méthodes de directions conjuguées*, Revue Française d'Information et Recherche Opérationnelle, 3 (1969), pp. 35–43.

[Pol69] B. T. Polyak, *Minimization of unsmooth functionals*, USSR Computational Mathematics and Mathematical Physics, 9 (1969), pp. 14–29.

[Pow78] M. J. D. Powell, *A fast algorithm for nonlinearly constrained optimization calculations*, in Numerical Analysis, Proceedings of the Seventh Biennial Conference held at the University of Dundee, Dundee, June 28–July 1, 1977, G. A. Watson, ed., vol. 630 of Lecture Notes in Mathematics, Berlin, 1978, Springer-Verlag, pp. 144–157.

[PsD78] B. N. Pshenichnyj and Y. M. Danilin, *Numerical Methods in Extremal Problems*, MIR Publishers, Moscow, 1978.

[Rad19] H. Rademacher, *Über partielle und totale Differenzierbarkeit von Funktionen mehrerer Variabeln under über die Transformation der Doppelintegrale*, Mathematische Annalen, 79 (1919), pp. 340–359.

[Rar98] R. L. Rardin, *Optimization in Operations Research*, Prentice Hall, Englewood Cliffs, NJ, 1998.

[Roc70] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[RoW97] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317 of Grundlehren der mathematischen Wissenschaften, Springer-Verlag, Berlin, 1997.

[Sau72] M. A. Saunders, *Large-scale linear programming using the Cholesky factorization*, Tech. Rep. Stan-cs-72-252, Computer Sciences Department, Stanford University, Stanford, 1972.

[Sch86] A. Schrijver, *Theory of Linear and Integer Programming*, Wiley, Chichester, 1986.

[Sch03] ———, *Combinatorial optimization*, vol. 24 of Algorithms and Combinatorics, Springer-Verlag, Berlin, 2003.

[Sha70] D. F. Shanno, *Conditioning of quasi-Newton methods for function minimization*, Mathematics of Computation, 24 (1970), pp. 647–656.

[She76] M. A. Shepilov, *Method of the generalized gradient for finding the absolute minimum of a convex function*, Cybernetics, 12 (1976), pp. 547–553.

[Sho70a] N. Z. Shor, *Convergence rate of the gradient descent method with dilatation of the space*, Cybernetics, 6 (1972), pp. 102–108.

[Sho70b] ———, *Utilization of the operation of space dilatation in the minimization of convex functions*, Cybernetics, 6 (1972), pp. 7–15.

[Sho77] ———, *Cut-off method with space extension in convex programming problems*, Cybernetics, 13 (1977), pp. 94–96.

[Sho85] ———, *Minimization Methods for Non-Differentiable Functions*, Springer-Verlag, Berlin, 1985. Translated from the Russian by K. C. Kiwiel and A. Ruszczyński.

[StW70] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimensions I*, Springer-Verlag, Berlin, 1970.

[Tah03] H. A. Taha, *Operations Research: An Introduction*, Prentice Hall, Englewood Cliffs, NJ, seventh ed., 2003.

[UUV04] M. Ulbrich, S. Ulbrich, and L. N. Vicente, *A globally convergent primal-dual interior-point filter method for nonlinear programming*, Mathematical Programming, 100 (2004), pp. 379–410.

[Van01] R. J. Vanderbei, *Linear Programming. Foundations and Extensions*, vol. 37 of International Series in Operations Research & Management Science, Kluwer Academic Publishers, Boston, MA, second ed., 2001.

[vHo77] B. von Hohenbalken, *Simplicial decomposition in nonlinear programming algorithms*, Mathematical Programming, 13 (1977), pp. 49–68.

[vNe28] J. von Neumann, *Zur Theorie der Gesellschaftsspiele*, Mathematische Annalen, 100 (1928), pp. 295–320.

[vNM43] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ, 1943.

[Wag75] H. M. Wagner, *Principles of Operations Research: With Applications to Managerial Decisions*, Prentice Hall, Englewood Cliffs, NJ, second ed., 1975.

[Wil99] H. P. Williams, *Model Building in Mathematical Programming*, John Wiley & Sons, Chichester, UK, fourth ed., 1999.

[Wil63] R. B. Wilson, *A simplicial algorithm for concave programming*, PhD thesis, Graduate School of Business Administration, Harvard University, Cambridge, MA, 1963.

[Wol69] P. Wolfe, *Convergence conditions for ascent methods*, SIAM Review, 11 (1969), pp. 226–235.

[Wol75] ———, *A method of conjugate subgradients for minimizing nondifferentiable functions*, Mathematical Programming Study, 3 (1975), pp. 145–173.

[Wol98] L. A. Wolsey, *Integer Programming*, Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, New York, NY, 1998.

[YuN77] D. B. Yudin and A. S. Nemirovskii, *Informational complexity and efficient methods for the solution of convex extremal problems*, Matekon, 13 (1977), pp. 25–45.

[Zan69] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*, Prentice Hall, Englewood Cliffs, NJ, 1969.

Index

Abadie's CQ, 125
active constraint ($\mathcal{I}(\boldsymbol{x})$), 89
adjacent extreme points, 216
affine hull, 43
affine combination, 43
affine function, 12, 58
affine independence, 34
affine subspace, 34
affine transformation, 303
algebraic characterization of adjacency, 217
approximate line search, 278
Armijo step, 279, 300, 316
artificial variables, 229
augmented Lagrangian function, 352
augmented Lagrangian method, 352

Banach's Theorem, 101
barrier function, 332
barrier problem, 332
basic feasible solution, 211
basic solution, 211
basic variables, 211
basis, 35
BFGS method, 275
Bland's rule, 234
boundary, 38
bounded set, 37
Brouwer's Theorem, 101
bundle method, 169

calculus rules, 39
canonical form, 241
Carathéodory's Theorem, 45
Cartesian product set, 147
central difference formula, 299
characterization of nonsingular matrices, 36
classification of optimization models, 11
closed mapping, 163
closed sets, 37
closure, 37
coercive function, 78
column dropping, 312
column generation, 7
combinatorial optimization, 157
complementarity, 147
Complementary Slackness Theorem, 249
composite function, 59
composite function, 104
composite operator, 104
concave function, 57
cone, 49
cone of feasible directions, 115
conjugate direction, 288, 296
conjugate gradient, 291
conjugate gradient method, 291
constraints, 4
continuity, 96
continuous function, 38
continuous relaxation, 157
continuously differentiable function, 39
contractive operator, 101
convergence rate, 298
    geometric, 101, 173
    linear, 298
    quadratic, 298

superlinear, 298
convex analysis, 41
convex combination, 43
convex function, 57, 96, 163
convex hull, 43
convex programming, 12
convex set, 41
coordinates, 35
CQ, 124

Danskin's Theorem, 164
decision science, 10
decision variable, 6
degenerate solution, 211
descent direction, 85, 169
DFP method, 295
differentiability, 167
differentiable, 38
direction of unboundedness, 222
directional derivative, 38, 163
distance function, 67
divergent series step length rule, 168, 311
domination, 348
dual feasible basis, 252
dual infeasible basis, 252
dual linear program, 240
dual simplex algorithm, 254
dual Simplex method, 252
duality gap, 145

effective domain, 143
efficient frontier, 348
eigenvalues and eigenvectors, 36
epigraph, 60, 80
ε-optimal solution, 96
equality constraint, 11
equivalent systems, 220
Euclidean projection, 66
Everett's Theorem, 179
exact penalty function, 342
existence of optimal solution, 214
extreme direction, 214
extreme point, 45

Farkas' Lemma, 56, 248
feasibility heuristic, 160
feasible direction, 88
feasible solution, 5, 14
feasible-direction methods, 307
filter, 348
filter-SQP methods, 348
finite termination, 283
finitely generated cone, 55
fixed point, 100
Fletcher–Reeves formula, 294
forward difference formula, 299
Frank–Wolfe algorithm, 309
Frank–Wolfe Theorem, 82
Fritz–John conditions, 121

Gauss–Seidel method, 105
geometric convergence rate, 101, 173
global minimum, 76
global optimality conditions, 146
global optimum, 76
necessary and sufficient conditions, 87, 91
Golden section, 278
gradient, 38
gradient projection algorithm, 315
gradient related, 272
gradient related method, 280, 282
Gram–Schmidt procedure, 290

hard constraint, 17
Hessian, 38

$\mathcal{I}(\boldsymbol{x})$, 89
identity matrix $\boldsymbol{I}^n$, 36
ill-conditioning, 348
implicit function, 39, 298
Implicit Function Theorem, 167
indicator function (χ_S), 170, 327
inequality constraint, 11
infimum, 14
integer programming, 12
integrality property, 13
interior, 38
interior penalty function, 96
interior point algorithm, 235
interpolation, 278

Jacobi method, 105

Karmarkar's algorithm, 235
KKT conditions, 125

Lagrange function, 142
Lagrange multiplier, 142, 182
Lagrange multipliers, 122
Lagrangian dual function, 143
Lagrangian dual problem, 143
Lagrangian relaxation, 18, 142
least-squares data fitting, 267
level set, 65, 78, 80
Levenberg–Marquardt, 274, 303
LICQ, 131
limit, 37
limit points, 37
line search, 277
  approximate, 278
  Armijo step length rule, 279, 300, 316
  Golden section, 278
  interpolation, 278
  Newton's method, 278
linear convergence rate, 298
linear function, 39
linear independence, 34
linear space, 34
linear-fractional programming, 219
Lipschitz continuity, 281
local convergence, 341
local minimum, 76
local optimum, 76
  necessary conditions, 85, 86, 89
  sufficient conditions, 87
logarithmic barrier, 333
logical constraint, 5
lower semi-continuity, 79

Maratos effect, 346
mathematical model, 4
mathematical programming, 9
matrices, 35
matrix inverse, 36
matrix game, 105
matrix norm, 35
matrix product, 35
matrix transpose, 35
max function, 164
mean-value theorem, 39
merit function, 343
method of successive averages (MSA), 321
MFCQ, 130
minimax theorem, 105
minimum, 14
minimum distance (dist_S), 170
multi-objective optimization, 14, 348

near-optimality, 95
negative curvature, 271
neighbourhood, 38
Newton's method, 273, 278, 301
Newton–Raphson method, 105, 273
Nobel laureates, 11
non-basic variables, 211
non-coordinability, 178
non-differentiable function, 284
non-expansive operator, 99
nonsingular matrix, 36
norm, 34
normal cone (N_X), 93

objective function, 4
open ball, 37
open set, 37
operations research, 10
optimal BFS, 222
optimal solution, 5
optimal value, 5
optimality, 10
optimization under uncertainty, 14
optimize, 3
orthogonality, 34, 147
orthonormal basis, 35

partial pricing, 226
pattern search methods, 299
penalty, 18
penalty function, 18
penalty parameter, 328
perturbation function ($p(\boldsymbol{b})$), 245
perturbation function ($p(\boldsymbol{u})$), 181
Phase I, 319
Phase I problem, 229

physical constraint, 5
piece-wise linear function, 285
Polak–Ribière formula, 294
Polyak step, 168
polyhedral cone, 49
polyhedron, 47
polytope, 45
positive (semi-)definite matrices, 37
pre-conditioning, 294
primal infeasibility criterion, 253
primal Simplex method, 252
projection, 66
projection arc, 316
projection operator, 66, 91, 99
projection problem, 319
proof by contradiction, 33
proper function, 15, 170
proximal point algorithm, 303
pseudo-convex function, 108

$\boldsymbol{Q}$-orthogonal, 288
quadratic convergence rate, 298
quadratic function, 40, 64
quadratic programming, 319
quasi-convex function, 109
quasi-Newton methods, 275, 295, 343

Rademacher's Theorem, 285
rank-two update, 295
recession cone, 81
reduced cost, 223
redundant constraint, 106
relaxation, 141
Relaxation Theorem, 141
Representation Theorem, 50, 214, 312
restricted master problem, 312
restricted simplicial decomposition, 313
restrification, 352
revised simplex method, 235

saddle point, 105, 147
scalar product, 34
secant method, 275
sensitivity analysis, 181, 182
sensitivity analysis for LP, 256
separation of convex sets, 99
Separation Theorem, 52, 98, 166
sequential linear programming (SLP), 351
sequential quadratic programming (SQP), 343
shadow price, 248
simplex, 314
Simplex method, 221
simplex method, 10
simplicial decomposition algorithm, 312
slack variable, 7
Slater CQ, 131
SLP algorithm, 351
soft constraint, 18, 180
SQP algorithm, 343
stalling, 235
standard basis, 35
stationary point, 85, 90
steepest descent, 270
steepest-edge rule, 226
stochastic programming, 14
strict inequality, 81
strict local minimum, 76
strictly convex function, 58
strictly quasi-convex function, 278
strong duality, 148
Strong Duality Theorem, 149, 151–153, 155, 246
sub-differentiability, 165
subdifferential, 162
subgradient, 162, 286
subgradient optimization, 169
subgradient projection method, 168
superlinear convergence rate, 298
symmetric matrices, 36

tangent cone, 115
traveling salesman problem, 158
triangle inequality, 36
trust region methods, 286
twice differentiable, 38

unimodal function, 278
unique optimum, 83

upper semi-continuity, 79

variable, 4
variational inequality, 90, 104
vector, 34
vector-valued functions, 38
von Neumann's Minimax Theorem, 105

Weak Duality Theorem, 144, 245
weak sharp minimum, 174
weak Wolfe condition, 280
weakly coercive function, 78
Weierstrass' Theorem, 80, 166
Wolfe condition, 280