Fundamentals of Optimization Theory With Applications to Machine Learning

Jean Gallier and Jocelyn Quaintance
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
e-mail: jean@cis.upenn.edu

November 10, 2025

Contents

Preface **9**

1 Introduction **13**

I Preliminaries for Optimization Theory **23**

2 Topology **25**
2.1 Metric Spaces and Normed Vector Spaces 25
2.2 Topological Spaces 31
2.3 Subspace and Product Topologies 36
2.4 Continuous Functions 41
2.5 Limits and Continuity; Uniform Continuity 45
2.6 Continuous Linear and Multilinear Maps 50
2.7 Complete Metric Spaces and Banach Spaces 56
2.8 Completion of a Metric Space 57
2.9 Completion of a Normed Vector Space 65
2.10 The Contraction Mapping Theorem 66
2.11 Further Readings 67
2.12 Summary 67
2.13 Problems 68

3 Differential Calculus **71**
3.1 Directional Derivatives, Total Derivatives 71
3.2 Properties of Derivatives 80
3.3 Jacobian Matrices 87
3.4 The Implicit and The Inverse Function Theorems 95
3.5 Second-Order and Higher-Order Derivatives 101
3.6 Taylor's Formula, Faà di Bruno's Formula 109
3.7 Further Readings 113
3.8 Summary 114
3.9 Problems 115

4 Extrema of Real-Valued Functions **117**

4.1 Local Extrema and Lagrange Multipliers . . . 118
4.2 Using Second Derivatives to Find Extrema . . . 130
4.3 Using Convexity to Find Extrema . . . 133
4.4 Summary . . . 143
4.5 Problems . . . 144

5 Newton's Method and Its Generalizations 147
5.1 Newton's Method for Real Functions of a Real Argument . . . 147
5.2 Generalizations of Newton's Method . . . 149
5.3 Summary . . . 158
5.4 Problems . . . 158

6 Quadratic Optimization Problems 167
6.1 Quadratic Optimization: The Positive Definite Case . . . 167
6.2 Quadratic Optimization: The General Case . . . 177
6.3 Maximizing a Quadratic Function on the Unit Sphere . . . 182
6.4 Summary . . . 187
6.5 Problems . . . 188

7 Schur Complements and Applications 189
7.1 Schur Complements . . . 189
7.2 SPD Matrices and Schur Complements . . . 192
7.3 SP Semidefinite Matrices and Schur Complements . . . 193
7.4 Summary . . . 195
7.5 Problems . . . 195

II Linear Optimization 197

8 Convex Sets, Cones, $\mathcal{H}$-Polyhedra 199
8.1 What is Linear Programming? . . . 199
8.2 Affine Subsets, Convex Sets, Hyperplanes, Half-Spaces . . . 201
8.3 Cones, Polyhedral Cones, and $\mathcal{H}$-Polyhedra . . . 204
8.4 Summary . . . 209
8.5 Problems . . . 210

9 Linear Programs 211
9.1 Linear Programs, Feasible Solutions, Optimal Solutions . . . 211
9.2 Basic Feasible Solutions and Vertices . . . 218
9.3 Summary . . . 225
9.4 Problems . . . 225

10 The Simplex Algorithm 229

10.1 The Idea Behind the Simplex Algorithm . . . 229
10.2 The Simplex Algorithm in General . . . 238
10.3 How to Perform a Pivoting Step Efficiently . . . 245
10.4 The Simplex Algorithm Using Tableaux . . . 249
10.5 Computational Efficiency of the Simplex Method . . . 258
10.6 Summary . . . 259
10.7 Problems . . . 260

11 Linear Programming and Duality 263
11.1 Variants of the Farkas Lemma . . . 263
11.2 The Duality Theorem in Linear Programming . . . 269
11.3 Complementary Slackness Conditions . . . 277
11.4 Duality for Linear Programs in Standard Form . . . 278
11.5 The Dual Simplex Algorithm . . . 281
11.6 The Primal-Dual Algorithm . . . 288
11.7 Summary . . . 298
11.8 Problems . . . 298

III NonLinear Optimization 303

12 Basics of Hilbert Spaces 305
12.1 The Projection Lemma . . . 305
12.2 Duality and the Riesz Representation Theorem . . . 318
12.3 Farkas–Minkowski Lemma in Hilbert Spaces . . . 323
12.4 Summary . . . 324
12.5 Problems . . . 325

13 General Results of Optimization Theory 327
13.1 Optimization Problems; Basic Terminology . . . 327
13.2 Existence of Solutions of an Optimization Problem . . . 331
13.3 Minima of Quadratic Functionals . . . 336
13.4 Elliptic Functionals . . . 342
13.5 Iterative Methods for Unconstrained Problems . . . 345
13.6 Gradient Descent Methods for Unconstrained Problems . . . 348
13.7 Convergence of Gradient Descent with Variable Stepsize . . . 356
13.8 Steepest Descent for an Arbitrary Norm . . . 359
13.9 Newton's Method For Finding a Minimum . . . 361
13.10 Conjugate Gradient Methods; Unconstrained Problems . . . 365
13.11 Gradient Projection for Constrained Optimization . . . 377
13.12 Penalty Methods for Constrained Optimization . . . 379
13.13 Summary . . . 381
13.14 Problems . . . 383

14 Introduction to Nonlinear Optimization 387
14.1 The Cone of Feasible Directions . . . 389
14.2 Active Constraints and Qualified Constraints . . . 395
14.3 The Karush–Kuhn–Tucker Conditions . . . 402
14.4 Equality Constrained Minimization . . . 413
14.5 Hard Margin Support Vector Machine; Version I . . . 418
14.6 Hard Margin Support Vector Machine; Version II . . . 423
14.7 Lagrangian Duality and Saddle Points . . . 431
14.8 Weak and Strong Duality . . . 440
14.9 Handling Equality Constraints Explicitly . . . 449
14.10 Dual of the Hard Margin Support Vector Machine . . . 452
14.11 Conjugate Function and Legendre Dual Function . . . 457
14.12 Some Techniques to Obtain a More Useful Dual Program . . . 467
14.13 Uzawa's Method . . . 471
14.14 Summary . . . 476
14.15 Problems . . . 478

15 Subgradients and Subdifferentials ⊛ 481
15.1 Extended Real-Valued Convex Functions . . . 483
15.2 Subgradients and Subdifferentials . . . 492
15.3 Basic Properties of Subgradients and Subdifferentials . . . 504
15.4 Additional Properties of Subdifferentials . . . 511
15.5 The Minimum of a Proper Convex Function . . . 515
15.6 Generalization of the Lagrangian Framework . . . 521
15.7 Summary . . . 525
15.8 Problems . . . 526

16 Dual Ascent Methods; ADMM 529
16.1 Dual Ascent . . . 531
16.2 Augmented Lagrangians and the Method of Multipliers . . . 535
16.3 ADMM: Alternating Direction Method of Multipliers . . . 540
16.4 Convergence of ADMM ⊛ . . . 543
16.5 Stopping Criteria . . . 552
16.6 Some Applications of ADMM . . . 553
16.7 Solving Hard Margin (SVM_{h2}) Using ADMM . . . 559
16.8 Applications of ADMM to ℓ^1-Norm Problems . . . 560
16.9 Summary . . . 566
16.10 Problems . . . 567

IV Applications to Machine Learning 569

17 Positive Definite Kernels 571

17.1 Feature Maps and Kernel Functions . . . 571
17.2 Basic Properties of Positive Definite Kernels . . . 577
17.3 Hilbert Space Representation of a Positive Kernel . . . 584
17.4 Kernel PCA . . . 587
17.5 Summary . . . 590
17.6 Problems . . . 591

18 Soft Margin Support Vector Machines 593
18.1 Soft Margin Support Vector Machines; (SVM_{s1}) . . . 596
18.2 Solving SVM (SVM_{s1}) Using ADMM . . . 611
18.3 Soft Margin Support Vector Machines; (SVM_{s2}) . . . 612
18.4 Solving SVM (SVM_{s2}) Using ADMM . . . 619
18.5 Soft Margin Support Vector Machines; ($SVM_{s2'}$) . . . 620
18.6 Classification of the Data Points in Terms of ν ($SVM_{s2'}$) . . . 631
18.7 Existence of Support Vectors for ($SVM_{s2'}$) . . . 634
18.8 Solving SVM ($SVM_{s2'}$) Using ADMM . . . 644
18.9 Soft Margin Support Vector Machines; (SVM_{s3}) . . . 648
18.10 Classification of the Data Points in Terms of ν (SVM_{s3}) . . . 655
18.11 Existence of Support Vectors for ($SVM_{\ni}$ . . . 657
18.12 Solving SVM (SVM_{s3}) Using ADMM . . . 659
18.13 Soft Margin SVM; (SVM_{s4}) . . . 663
18.14 Solving SVM (SVM_{s4}) Using ADMM . . . 671
18.15 Soft Margin SVM; (SVM_{s5}) . . . 673
18.16 Solving SVM (SVM_{s5}) Using ADMM . . . 677
18.17 Summary and Comparison of the SVM Methods . . . 679
18.18 Problems . . . 692

19 Ridge Regression, Lasso, Elastic Net 697
19.1 Ridge Regression . . . 698
19.2 Ridge Regression; Learning an Affine Function . . . 701
19.3 Kernel Ridge Regression . . . 710
19.4 Lasso Regression (ℓ^1-Regularized Regression) . . . 714
19.5 Lasso Regression; Learning an Affine Function . . . 718
19.6 Elastic Net Regression . . . 724
19.7 Summary . . . 730
19.8 Problems . . . 730

20 ν-SV Regression 733
20.1 ν-SV Regression; Derivation of the Dual . . . 733
20.2 Existence of Support Vectors . . . 744
20.3 Solving ν-Regression Using ADMM . . . 754
20.4 Kernel ν-SV Regression . . . 760
20.5 ν-Regression Version 2; Penalizing b . . . 763

20.6 Summary 770
20.7 Problems 771

V Appendix 773

A Total Orthogonal Families in Hilbert Spaces 775
A.1 Total Orthogonal Families, Fourier Coefficients 775
A.2 The Hilbert Space $\ell^2(K)$ and the Riesz–Fischer Theorem 784
A.3 Summary 793
A.4 Problems 794

B Matlab Programs 795
B.1 Hard Margin (SVM_{h2}) 795
B.2 Soft Margin SVM ($\mathrm{SVM}_{s2'}$) 799
B.3 Soft Margin SVM (SVM_{s3}) 807
B.4 ν-SV Regression 812

Bibliography 819

Index 825

Preface

In recent years, computer vision, robotics, machine learning, and data science have been some of the key areas that have contributed to major advances in technology. Anyone who looks at papers or books in the above areas will be baffled by a strange jargon involving exotic terms such as kernel PCA, ridge regression, lasso regression, support vector machines (SVM), Lagrange multipliers, KKT conditions, *etc.* Do support vector machines chase cattle to catch them with some kind of super lasso? No! But one will quickly discover that behind the jargon which always comes with a new field (perhaps to keep the outsiders out of the club), lies a lot of "classical" linear algebra and techniques from optimization theory. And there comes the main challenge: in order to understand and use tools from machine learning, computer vision, and so on, one needs to have a firm background in linear algebra and optimization theory. To be honest, some probability theory and statistics should also be included, but we already have enough to contend with.

Many books on machine learning struggle with the above problem. How can one understand what are the dual variables of a ridge regression problem if one doesn't know about the Lagrangian duality framework? Similarly, how is it possible to discuss the dual formulation of SVM without a firm understanding of the Lagrangian framework?

The easy way out is to sweep these difficulties under the rug. If one is just a consumer of the techniques we mentioned above, the cookbook recipe approach is probably adequate. But this approach doesn't work for someone who really wants to do serious research and make significant contributions. To do so, we believe that one must have a solid background in linear algebra and optimization theory.

This is a problem because it means investing a great deal of time and energy studying these fields, but we believe that perseverance will be amply rewarded.

This second volume covers some elements of optimization theory and applications, especially to machine learning. This volume is divided in five parts:

(1) Preliminaries of Optimization Theory.

(2) Linear Optimization.

(3) Nonlinear Optimization.

(4) Applications to Machine Learning.

(5) An appendix consisting of two chapers; one on Hilbert bases and the Riesz–Fischer theorem, the other one containing `Matlab` code.

Part I is devoted to some preliminaries of optimization theory. The goal of most optimization problems is to minimize (or maximize) some objective function J subject to equality or inequality constraints. Therefore it is important to understand when a function J has a minimum or a maximum (an optimum). In most optimization problems, we need to find necessary conditions for a function $J\colon \Omega \to \mathbb{R}$ to have a local extremum with respect to a subset U of Ω (where Ω is open). This can be done in two cases:

(1) The set U is defined by a set of equations,

$$U = \{x \in \Omega \mid \varphi_i(x) = 0,\ \ 1 \leq i \leq m\},$$

where the functions $\varphi_i\colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

(2) The set U is defined by a set of inequalities,

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ \ 1 \leq i \leq m\},$$

where the functions $\varphi_i\colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

The case of equality constraints is much easier to deal with and is treated in Chapter 4.

In the case of equality constraints, a necessary condition for a local extremum with respect to U can be given in terms of *Lagrange multipliers*.

Part II deals with the special case where the objective function is a linear form and the constraints are affine inequality and equality constraints. This subject is known as *linear programming*, and the next four chapters give an introduction to the subject.

Part III is devoted to nonlinear optimization, which is the case where the objective function J is not linear and the constaints are inequality constraints. Since it is practically impossible to say anything interesting if the constraints are not convex, we quickly consider the convex case.

Chapter 13 is devoted to some general results of optimization theory. A main theme is to find sufficient conditions that ensure that an objective function has a minimum which is achieved. We define gradient descent methods (including Newton's method), and discuss their convergence.

Chapter 14 contains the most important results of nonlinear optimization theory. Theorem 14.6 gives necessary conditions for a function J to have a minimum on a subset U defined by convex inequality constraints in terms of the Karush–Kuhn–Tucker conditions. Furthermore, if J is also convex and if the KKT conditions hold, then J has a global minimum.

We illustrate the KKT conditions on an interesting example from machine learning the so-called *hard margin support vector machine*; see Sections 14.5 and 14.6. The problem is to separate two disjoint sets of points, $\{u_i\}_{i=1}^p$ and $\{v_j\}_{j=1}^q$, using a hyperplane satisfying some optimality property (to maximize the margin).

Section 14.7 contains the most important results of the chapter. The notion of Lagrangian duality is presented and we discuss *weak duality* and *strong duality*.

In Chapter 15, we consider some deeper aspects of the theory of convex functions that are not necessarily differentiable at every point of their domain. Some substitute for the gradient is needed. Fortunately, for convex functions, there is such a notion, namely *subgradients*. A major motivation for developing this more sophisticated theory of differentiation of convex functions is to extend the Lagrangian framework to convex functions that are not necessarily differentiable.

Chapter 16 is devoted to the presentation of one of the best methods known at the present for solving optimization problems involving equality constraints, called ADMM (alternating direction method of multipliers). In fact, this method can also handle more general constraints, namely, membership in a convex set. It can also be used to solve *lasso minimization*.

In Section 16.4, we prove the convergence of ADMM under exactly the same assumptions as in Boyd et al. [17]. It turns out that Assumption (2) in Boyd et al. [17] implies that the matrices $A^\top A$ and $B^\top B$ are invertible (as we show after the proof of Theorem 16.1). This allows us to prove a convergence result stronger than the convergence result proven in Boyd et al. [17].

The next four chapters constitute Part IV, which covers some applications of optimization theory (in particular Lagrangian duality) to machine learning.

Chapter 17 is an introduction to positive definite kernels and the use of kernel functions in machine learning called a *kernel function*.

We illustrate the kernel methods on kernel PCA.

In Chapter 18 we return to the problem of separating two disjoint sets of points, $\{u_i\}_{i=1}^p$ and $\{v_j\}_{j=1}^q$, but this time we do not assume that these two sets are separable. To cope with nonseparability, we allow points to invade the safety zone around the separating hyperplane, and even points on the wrong side of the hyperplane. Such a method is called *soft margin support vector machine (SVM)*. We discuss variations of this method, including ν-SV classification. In each case we present a careful derivation of the dual. We prove rigorous results about the existence of support vectors.

In Chapter 19, we discuss *linear regression*, *ridge regression*, *lasso regression* and *elastic net regression*.

In Chapter 20 we present *ν-SV Regression*. This method is designed in the same spirit as soft margin SVM, in the sense that it allows a margin of error. Here the errors are penalized

in the ℓ^1-sense. We present a careful derivation of the dual and discuss the existence of support vectors.

The methods presented in Chapters 18, 19 and 20 have all been implemented in `Matlab`, and much of this code is given in Appendix B. Remarkably, ADMM emerges as the main engine for solving most of these optimization problems. Thus it is nice to see the continuum spanning from theoretical considerations of convergence and correctness to practical matters of implementation. It is fun to see how these abstract Lagrange multipliers yield concrete results such as the weight vector w defining the desired hyperplane in regression or SVM.

Except for a few exceptions we provide complete proofs. We did so to make this book self-contained, but also because we believe that no deep knowledge of this material can be acquired without working out some proofs. However, our advice is to skip some of the proofs upon first reading, especially if they are long and intricate.

The chapters or sections marked with the symbol $\circledast$ contain material that is typically more specialized or more advanced, and they can be omitted upon first (or second) reading.

Acknowledgement: We would like to thank Christine Allen-Blanchette, Kostas Daniilidis, Carlos Esteves, Spyridon Leonardos, Stephen Phillips, João Sedoc, Stephen Shatz, Jianbo Shi, and Marcelo Siqueira, for reporting typos and for helpful comments. Mary Pugh and William Yu (at the University of Toronto) taught a course using our book and reported a number of typos and errors. We warmly thank them as well as their students, not only for finding errors, but also for very hepful comments and suggestions for simplifying some proofs. Thanks to Gilbert Strang. We learned much from his books which have been a major source of inspiration. Special thanks to Steven Boyd. We learned a lot from his remarkable book on convex optimization and his papers, and Part III of our book is significantly inspired by his writings. The first author also wishes to express his deepest gratitute to Philippe G. Ciarlet who was his teacher and mentor in 1970-1972 while he was a student at ENPC in Paris. Professor Ciarlet was by far his best teacher. He also knew how to instill in his students the importance of intellectual rigor, honesty, and modesty. He still has his typewritten notes on measure theory and integration, and on numerical linear algebra. The latter became his wonderful book Ciarlet [25], from which we have borrowed heavily.

Chapter 1

Introduction

This second volume covers some elements of optimization theory and applications, especially to machine learning. This volume is divided in five parts:

(1) Preliminaries of Optimization Theory.

(2) Linear Optimization.

(3) Nonlinear Optimization.

(4) Applications to Machine Learning.

(5) An appendix consisting of two chapers; one on Hilbert bases and the Riesz–Fischer theorem, the other one containing `Matlab` code.

Part I is devoted to some preliminaries of optimization theory. The goal of most optimization problems is to minimize (or maximize) some objective function J subject to equality or inequality constraints. Therefore it is important to understand when a function J has a minimum or a maximum (an optimum). If the function J is sufficiently differentiable, then a necessary condition for a function to have an optimum typically involves the derivative of the function J, and if J is real-valued, its gradient ∇J.

Thus it is desirable to review some basic notions of topology and calculus, in particular, to have a firm grasp of the notion of derivative of a function between normed vector spaces. Partial derivatives $\partial f/\partial A$ of functions whose range and domain are spaces of matrices tend to be used casually, even though in most cases a correct definition is never provided. It is possible, and simple, to define rigorously derivatives, gradients, and directional derivatives of functions defined on matrices and to avoid these nonsensical partial derivatives.

Chapter 2 contains a review of basic topological notions used in analysis. We pay particular attention to complete metric spaces and complete normed vector spaces. In fact, we provide a detailed construction of the completion of a metric space (and of a normed vector space) using equivalence classes of Cauchy sequences. Chapter 3 is devoted to some notions

of differential calculus, in particular, directional derivatives, total derivatives, gradients, Hessians, and the inverse function theorem.

Chapter 4 deals with extrema of real-valued functions. In most optimization problems, we need to find necessary conditions for a function $J\colon \Omega \to \mathbb{R}$ to have a local extremum with respect to a subset U of Ω (where Ω is open). This can be done in two cases:

(1) The set U is defined by a set of equations,

$$U = \{x \in \Omega \mid \varphi_i(x) = 0,\ \ 1 \leq i \leq m\},$$

where the functions $\varphi_i\colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

(2) The set U is defined by a set of inequalities,

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ \ 1 \leq i \leq m\},$$

where the functions $\varphi_i\colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

In (1), the equations $\varphi_i(x) = 0$ are called *equality constraints*, and in (2), the inequalities $\varphi_i(x) \leq 0$ are called *inequality constraints*. The case of equality constraints is much easier to deal with and is treated in Chapter 4.

If the functions φ_i are convex and Ω is convex, then U is convex. This is a very important case that we will discuss later. In particular, if the functions φ_i are affine, then the equality constraints can be written as $Ax = b$, and the inequality constraints as $Ax \leq b$, for some $m \times n$ matrix A and some vector $b \in \mathbb{R}^m$. We will also discuss the case of affine constraints later.

In the case of equality constraints, a necessary condition for a local extremum with respect to U can be given in terms of *Lagrange multipliers*. In the case of inequality constraints, there is also a necessary condition for a local extremum with respect to U in terms of generalized Lagrange multipliers and the *Karush–Kuhn–Tucker* conditions. This will be discussed in Chapter 14.

In Chapter 5 we discuss Newton's method and some of its generalizations (the Newton–Kantorovich theorem). These are methods to find the zeros of a function.

Chapter 6 covers the special case of determining when a quadratic function has a minimum, subject to affine equality constraints. A complete answer is provided in terms of the notion of symmetric positive semidefinite matrices.

The Schur complement is introduced in Chapter 7. We give a complete proof of a criterion for a matrix to be positive definite (or positive semidefinite) stated in Boyd and Vandenberghe [18] (Appendix B).

Part II deals with the special case where the objective function is a linear form and the constraints are affine inequality and equality constraints. This subject is known as linear

programming, and the next four chapters give an introduction to the subject. Although linear programming has been supplanted by convex programming and its variants, it is still a great workhorse. It is also a great warm up for the general treatment of Lagrangian duality. We pay particular attention to versions of Farkas' lemma, which is at the heart of duality in linear programming.

Part III is devoted to nonlinear optimization, which is the case where the objective function J is not linear and the constaints are inequality constraints. Since it is practically impossible to say anything interesting if the constraints are not convex, we quickly consider the convex case.

In optimization theory one often deals with function spaces of infinite dimension. Typically, these spaces either are Hilbert spaces or can be completed as Hilbert spaces. Thus it is important to have some minimum knowledge about Hilbert spaces, and we feel that this minimum knowledge includes the projection lemma, the fact that a closed subset has an orthogonal complement, the Riesz representation theorem, and a version of the Farkas–Minkowski lemma. Chapter 12 covers these topics. A more detailed introduction to Hilbert spaces is given in Appendix A.

Chapter 13 is devoted to some general results of optimization theory. A main theme is to find sufficient conditions that ensure that an objective function has a minimum which is achieved. We define the notion of a coercive function. The most general result is Theorem 13.2, which applies to a coercive convex function on a convex subset of a separable Hilbert space. In the special case of a coercive quadratic functional, we obtain the Lions–Stampacchia theorem (Theorem 13.6), and the Lax–Milgram theorem (Theorem 13.7). We define elliptic functionals, which generalize quadratic functions defined by symmetric positive definite matrices. We define gradient descent methods, and discuss their convergence. A gradient descent method looks for a descent direction and a stepsize parameter, which is obtained either using an exact line search or a backtracking line search. A popular technique to find the search direction is steepest descent. In addition to steepest descent for the Euclidean norm, we discuss steepest descent for an arbitrary norm. We also consider a special case of steepest descent, Newton's method. This method converges faster than the other gradient descent methods, but it is quite expensive since it requires computing and storing Hessians. We also present the method of conjugate gradients and prove its correctness. We briefly discuss the method of gradient projection and the penalty method in the case of constrained optima.

Chapter 14 contains the most important results of nonlinear optimization theory. We begin by defining the cone of feasible directions and then state a necessary condition for a function to have local minimum on a set U that is not necessarily convex in terms of the cone of feasible directions. The cone of feasible directions is not always convex, but it is if the constraints are inequality constraints. An inequality constraint $\varphi(u) \leq 0$ is said to be *active* is $\varphi(u) = 0$. One can also define the notion of *qualified constraint*. Theorem 14.5 gives necessary conditions for a function J to have a minimum on a subset U defined by qualified inequality constraints in terms of the Karush–Kuhn–Tucker conditions (for short

KKT conditions), which involve nonnegative Lagrange multipliers. The proof relies on a version of the Farkas–Minkowski lemma. Some of the KTT conditions assert that $\lambda_i \varphi_i(u) = 0$, where $\lambda_i \geq 0$ is the Lagrange multiplier associated with the constraint $\varphi_i \leq 0$. To some extent, this implies that active constaints are more important than inactive constraints, since if $\varphi_i(u) < 0$ is an inactive constraint, then $\lambda_i = 0$. In general, the KKT conditions are useless, unless the constraints are convex. In this case, there is a manageable notion of qualified constraint given by Slater's conditions. Theorem 14.6 gives necessary conditions for a function J to have a minimum on a subset U defined by convex inequality constraints in terms of the Karush–Kuhn–Tucker conditions. Furthermore, if J is also convex and if the KKT conditions hold, then J has a global minimum.

In Section 14.4, we apply Theorem 14.6 to the special case where the constraints are equality constraints, which can be expressed as $Ax = b$. In the special case where the convex objective function J is a convex quadratic functional of the form

$$J(x) = \frac{1}{2} x^\top P x + q^\top x + r,$$

where P is a $n \times n$ symmetric positive semidefinite matrix, the necessary and sufficient conditions for having a minimum are expressed by a linear system involving a matrix called the KKT matrix. We discuss conditions that guarantee that the KKT matrix is invertible, and how to solve the KKT system. We also briefly discuss variants of Newton's method dealing with equality constraints.

We illustrate the KKT conditions on an interesting example, the so-called hard margin support vector machine; see Sections 14.5 and 14.6. The problem is a classification problem, or more accurately a separation problem. Suppose we have two nonempty disjoint finite sets of p *blue* points $\{u_i\}_{i=1}^p$ and q *red* points $\{v_j\}_{j=1}^q$ in $\mathbb{R}^n$. Our goal is to find a hyperplane H of equation $w^\top x - b = 0$ (where $w \in \mathbb{R}^n$ is a nonzero vector and $b \in \mathbb{R}$), such that all the blue points u_i are in one of the two open half-spaces determined by H, and all the red points v_j are in the other open half-space determined by H.

If the two sets are indeed separable, then in general there are infinitely many hyperplanes separating them. Vapnik had the idea to find a hyperplane that maximizes the smallest distance between the points and the hyperplane. Such a hyperplane is indeed unique and is called a maximal hard margin hyperplane, or hard margin support vector machine. The support vectors are those for which the constraints are active.

Section 14.7 contains the most important results of the chapter. The notion of Lagrangian duality is presented. Given a primal optimization problem (P) consisting in minimizing an objective function $J(v)$ with respect to some inequality constraints $\varphi_i(v) \leq 0$, $i = 1, \ldots, m$, we define the *dual function* $G(\mu)$ as the result of minimizing the Lagrangian

$$L(v, \mu) = J(v) + \sum_{i=1}^{m} \mu_i \varphi_i(v)$$

with respect to v, with $\mu \in \mathbb{R}^m_+$. The dual program (D) is then to maximize $G(\mu)$ with respect to $\mu \in \mathbb{R}^m_+$. It turns out that G is a concave function, and the dual program is an unconstrained maximization. This is actually a misleading statement because G is generally a partial function, so maximizing $G(\mu)$ is equivalent to a constrained maximization problem in which the constraints specify the domain of G, but in many cases, we obtain a dual program simpler than the primal program. If d^* is the optimal value of the dual program and if p^* is the optimal value of the primal program, we always have

$$d^* \leq p^*,$$

which is known as *weak duality*. Under certain conditions, $d^* = p^*$, that is, the duality gap is zero, in which case we say that *strong duality* holds. Also, under certain conditions, a solution of the dual yields a solution of the primal, and if the primal has an optimal solution, then the dual has an optimal solution, but beware that the converse is generally false (see Theorem 14.17). We also show how to deal with equality constraints, and discuss the use of conjugate functions to find the dual function. Our coverage of Lagrangian duality is quite thorough, but we do not discuss more general orderings such as the semidefinite ordering. For these topics which belong to convex optimization, the reader is referred to Boyd and Vandenberghe [18].

In Chapter 15, we consider some deeper aspects of the theory of convex functions that are not necessarily differentiable at every point of their domain. Some substitute for the gradient is needed. Fortunately, for convex functions, there is such a notion, namely *subgradients*. Geometrically, given a (proper) convex function f, the subgradients at x are vectors normal to supporting hyperplanes to the epigraph of the function at $(x, f(x))$. The *subdifferential* $\partial f(x)$ to f at x is the set of all subgradients at x. A crucial property is that f is differentiable at x iff $\partial f(x) = \{\nabla f_x\}$, where ∇f_x is the gradient of f at x. Another important property is that a (proper) convex function f attains its minimum at x iff $0 \in \partial f(x)$. A major motivation for developing this more sophisticated theory of "differentiation" of convex functions is to extend the Lagrangian framework to convex functions that are not necessarily differentiable.

Experience shows that the applicability of convex optimization is significantly increased by considering extended real-valued functions, namely functions $f\colon S \to \mathbb{R} \cup \{-\infty, +\infty\}$, where S is some subset of $\mathbb{R}^n$ (usually convex). This is reminiscent of what happens in measure theory, where it is natural to consider functions that take the value $+\infty$.

In Section 15.1, we introduce extended real-valued functions, which are functions that may also take the values $\pm\infty$. In particular, we define proper convex functions, and the closure of a convex function. Subgradients and subdifferentials are defined in Section 15.2. We discuss some properties of subgradients in Section 15.3 and Section 15.4. In particular, we relate subgradients to one-sided directional derivatives. In Section 15.5, we discuss the problem of finding the minimum of a proper convex function and give some criteria in terms of subdifferentials. In Section 15.6, we sketch the generalization of the results presented in Chapter 14 about the Lagrangian framework to programs allowing an objective function and inequality constraints which are convex but not necessarily differentiable.

This chapter relies heavily on Rockafellar [61]. We tried to distill the body of results needed to generalize the Lagrangian framework to convex but not necessarily differentiable functions. Some of the results in this chapter are also discussed in Bertsekas [9, 12, 10].

Chapter 16 is devoted to the presentation of one of the best methods known at the present for solving optimization problems involving equality constraints, called ADMM (alternating direction method of multipliers). In fact, this method can also handle more general constraints, namely, membership in a convex set. It can also be used to solve *lasso minimization*.

In this chapter, we consider the problem of minimizing a convex function J (not necessarily differentiable) under the equality constraints $Ax = b$. In Section 16.1, we discuss the dual ascent method. It is essentially gradient descent applied to the dual function G, but since G is maximized, gradient descent becomes gradient ascent.

In order to make the minimization step of the dual ascent method more robust, one can use the trick of adding the penalty term $(\rho/2)\,\|Au - b\|_2^2$ to the Lagrangian. We obtain the *augmented Lagrangian*

$$L_\rho(u, \lambda) = J(u) + \lambda^\top (Au - b) + (\rho/2)\,\|Au - b\|_2^2,$$

with $\lambda \in \mathbb{R}^m$, and where $\rho > 0$ is called the *penalty parameter*. We obtain the minimization Problem (P_ρ),

$$\begin{array}{ll} \text{minimize} & J(u) + (\rho/2)\,\|Au - b\|_2^2 \\ \text{subject to} & Au = b, \end{array}$$

which is equivalent to the original problem.

The benefit of adding the penalty term $(\rho/2)\,\|Au - b\|_2^2$ is that by Proposition 15.37, Problem (P_ρ) has a unique optimal solution under mild conditions on A. Dual ascent applied to the dual of (P_ρ) is called the *method of multipliers* and is discussed in Section 16.2.

The new twist in ADMM is to split the function J into two independent parts, as $J(x, z) = f(x) + g(z)$, and to consider the Minimization Problem (P_{admm}),

$$\begin{array}{ll} \text{minimize} & f(x) + g(z) \\ \text{subject to} & Ax + Bz = c, \end{array}$$

for some $p \times n$ matrix A, some $p \times m$ matrix B, and with $x \in \mathbb{R}^n$, $z \in \mathbb{R}^m$, and $c \in \mathbb{R}^p$. We also assume that f and g are convex.

As in the method of multipliers, we form the augmented Lagrangian

$$L_\rho(x, z, \lambda) = f(x) + g(z) + \lambda^\top (Ax + Bz - c) + (\rho/2)\,\|Ax + Bz - c\|_2^2,$$

with $\lambda \in \mathbb{R}^p$ and for some $\rho > 0$. The major difference with the method of multipliers is that instead of performing a minimization step jointly over x and z, ADMM first performs an

x-minimization step and then a z-minimization step. Thus x and z are updated in an alternating or sequential fashion, which accounts for the term *alternating direction*. Because the Lagrangian is augmented, some mild conditions on A and B imply that these minimization steps are guaranteed to terminate. ADMM is presented in Section 16.3.

In Section 16.4, we prove the convergence of ADMM under exactly the same assumptions as in Boyd et al. [17]. It turns out that Assumption (2) in Boyd et al. [17] implies that the matrices $A^\top A$ and $B^\top B$ are invertible (as we show after the proof of Theorem 16.1). This allows us to prove a convergence result stronger than the convergence result proven in Boyd et al. [17]. In particular, we prove that *all* of the sequences (x^k), (z^k), and (λ^k) converge to optimal solutions $(\widetilde{x}, \widetilde{z})$, and $\widetilde{\lambda}$.

In Section 16.5, we discuss stopping criteria. In Section 16.6, we present some applications of ADMM, in particular, minimization of a proper closed convex function f over a closed convex set C in $\mathbb{R}^n$ and quadratic programming. The second example provides one of the best methods for solving quadratic problems, in particular, the SVM problems discussed in Chapter 18. Section 16.8 gives applications of ADMM to ℓ^1-norm problems, in particular, lasso regularization which plays an important role in machine learning.

The next four chapters constitute Part IV, which covers some applications of optimization theory (in particular Lagrangian duality) to machine learning.

Chapter 17 is an introduction to positive definite kernels and the use of kernel functions in machine learning.

Let X be a nonempty set. If the set X represents a set of highly nonlinear data, it may be advantageous to map X into a space F of much higher dimension called the *feature space*, using a function $\varphi\colon X \to F$ called a *feature map*. This idea is that φ "unwinds" the description of the objects in F in an attempt to make it linear. The space F is usually a vector space equipped with an inner product $\langle -, - \rangle$. If F is infinite dimensional, then we assume that it is a Hilbert space.

Many algorithms that analyze or classify data make use of the inner products $\langle \varphi(x), \varphi(y) \rangle$, where $x, y \in X$. These algorithms make use of the function $\kappa\colon X \times X \to \mathbb{C}$ given by

$$\kappa(x, y) = \langle \varphi(x), \varphi(y) \rangle, \qquad x, y \in X,$$

called a *kernel function*.

The kernel trick is to pretend that we have a feature embedding $\varphi\colon X \to F$ (actuallly unknown), but to only use inner products $\langle \varphi(x), \varphi(y) \rangle$ that can be evaluated using the original data through the known kernel function κ. It turns out that the functions of the form κ as above can be defined in terms of a condition which is reminiscent of positive semidefinite matrices (see Definition 17.2). Furthermore, every function satisfying Definition 17.2 arises from a suitable feature map into a Hilbert space; see Theorem 17.8.

We illustrate the kernel methods on kernel PCA (see Section 17.4).

In Chapter 18 we return to the problem of separating two disjoint sets of points, $\{u_i\}_{i=1}^p$ and $\{v_j\}_{j=1}^q$, but this time we do not assume that these two sets are separable. To cope with nonseparability, we allow points to invade the safety zone around the separating hyperplane, and even points on the wrong side of the hyperplane. Such a method is called soft margin support vector machine. We discuss variations of this method, including ν-SV classification. In each case we present a careful derivation of the dual and we explain how to solve it using ADMM. We prove rigorous results about the existence of support vectors.

In Chapter 19 we discuss linear regression. This problem can be cast as a learning problem. We observe a sequence of (distinct) pairs $((x_1, y_1), \ldots, (x_m, y_m))$ called a *set of training data*, where $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$, viewed as input-output pairs of some unknown function f that we are trying to infer. The simplest kind of function is a linear function $f(x) = x^\top w$, where $w \in \mathbb{R}^n$ is a vector of coefficients usually called a *weight vector*. Since the problem is overdetermined and since our observations may be subject to errors, we can't solve for w exactly as the solution of the system $Xw = y$, so instead we solve the least-squares problem of minimizing $\|Xw - y\|_2^2$, where X is the $m \times n$ matrix whose *rows* are the row vectors $x_i^\top$. In general there are still infinitely many solutions so we add a regularizing term. If we add the term $K\|w\|_2^2$ to the objective function $J(w) = \|Xw - y\|_2^2$, then we have *ridge regression*. This problem is discussed in Section 19.1.

We derive the dual program. The dual has a unique solution which yields a solution of the primal. However, the solution of the dual is given in terms of the matrix $XX^\top$ (whereas the solution of the primal is given in terms of $X^\top X$), and since our data points x_i are represented by the rows of the matrix X, we see that this solution only involves inner products of the x_i. This observation is the core of the idea of kernel functions, which we introduce. We also explain how to solve the problem of learning an affine function $f(x) = x^\top w + b$.

In general the vectors w produced by ridge regression have few zero entries. In practice it is highly desirable to obtain sparse solutions, that is, vectors w with many components equal to zero. This can be achieved by replacing the regularizing term $K\|w\|_2^2$ by the regularizing term $K\|w\|_1$; that is, to use the ℓ^1-norm instead of the ℓ^2-norm; see Section 19.4. This method has the exotic name of *lasso regression*. This time there is no closed-form solution, but this is a convex optimization problem and there are efficient iterative methods to solve it. We show that ADMM provides an efficient solution.

Lasso has some undesirable properties, in particular when the dimension of the data is much larger than the number of data. In order to alleviate these problems, *elastic net regression* penalizes w with *both* an ℓ^2 regularizing term $K\|w\|_2^2$ and an ℓ^1 regularizing term $\tau\|w\|_1$. The method of elastic net blends ridge regression and lasso and attempts to retain their best properties; see Section 19.6. It can also be solved using ADMM but it appears to be much slower than lasso when K is small and the dimension of the data is much larger than the number of data.

In Chapter 20 we present ν-SV Regression. This method is designed in the same spirit as soft margin SVM, in the sense that it allows a margin of error. Here the errors are

penalized in the ℓ^1-sense. We discuss several variations of the method and show how to solve them using ADMM. We present a careful derivation of the dual and discuss the existence of support vectors.

Part I

Preliminaries for Optimization Theory

Chapter 2

Topology

This chapter contains a review of basic topological concepts. First metric spaces are defined. Next normed vector spaces are defined. Closed and open sets are defined, and their basic properties are stated. The general concept of a topological space is defined. The closure and the interior of a subset are defined. The subspace topology and the product topology are defined. Continuous maps and homeomorphisms are defined. Limits of sequences are defined. Continuous linear maps and multilinear maps are defined and studied briefly. Cauchy sequences and complete metric spaces are defined. We prove that every metric space can be embedded in a complete metric space called its completion. A complete normed vector space is called a *Banach space*. We prove that every normed vector space can be embedded in a complete normed vector space. We conclude with the contraction mapping theorem in a complete metric space.

2.1 Metric Spaces and Normed Vector Spaces

Most spaces considered in this book have a topological structure given by a metric or a norm, and we first review these notions. We begin with metric spaces. Recall that $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$.

Definition 2.1. A *metric space* is a set E together with a function $d\colon E \times E \to \mathbb{R}_+$, called a *metric, or distance*, assigning a nonnegative real number $d(x, y)$ to any two points $x, y \in E$, and satisfying the following conditions for all $x, y, z \in E$:

(D1) $d(x, y) = d(y, x)$. (symmetry)

(D2) $d(x, y) \geq 0$, and $d(x, y) = 0$ iff $x = y$. (positivity)

(D3) $d(x, z) \leq d(x, y) + d(y, z)$. (triangle inequality)

Geometrically, Condition (D3) expresses the fact that in a triangle with vertices x, y, z, the length of any side is bounded by the sum of the lengths of the other two sides. From

(D3), we immediately get

$$|d(x, y) - d(y, z)| \leq d(x, z).$$

Let us give some examples of metric spaces. Recall that the *absolute value* $|x|$ of a real number $x \in \mathbb{R}$ is defined such that $|x| = x$ if $x \geq 0$, $|x| = -x$ if $x < 0$, and for a complex number $x = a + ib$, by $|x| = \sqrt{a^2 + b^2}$.

Example 2.1.

1. Let $E = \mathbb{R}$, and $d(x, y) = |x - y|$, the absolute value of $x - y$. This is the so-called natural metric on $\mathbb{R}$.

2. Let $E = \mathbb{R}^n$ (or $E = \mathbb{C}^n$). We have the *Euclidean metric*

$$d_2(x, y) = \left(|x_1 - y_1|^2 + \cdots + |x_n - y_n|^2\right)^{\frac{1}{2}},$$

   the distance between the points $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$.

3. For every set E, we can define the *discrete metric*, defined such that $d(x, y) = 1$ iff $x \neq y$, and $d(x, x) = 0$.

4. For any $a, b \in \mathbb{R}$ such that $a < b$, we define the following sets:

   $[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$, (closed interval)

   $(a, b) = \{x \in \mathbb{R} \mid a < x < b\}$, (open interval)

   $[a, b) = \{x \in \mathbb{R} \mid a \leq x < b\}$, (interval closed on the left, open on the right)

   $(a, b] = \{x \in \mathbb{R} \mid a < x \leq b\}$, (interval open on the left, closed on the right)

   Let $E = [a, b]$, and $d(x, y) = |x - y|$. Then $([a, b], d)$ is a metric space.

We will need to define the notion of proximity in order to define convergence of limits and continuity of functions. For this we introduce some standard "small neighborhoods."

Definition 2.2. Given a metric space E with metric d, for every $a \in E$, for every $\rho \in \mathbb{R}$, with $\rho > 0$, the set

$$B(a, \rho) = \{x \in E \mid d(a, x) \leq \rho\}$$

is called the *closed ball of center a and radius ρ*, the set

$$B_0(a, \rho) = \{x \in E \mid d(a, x) < \rho\}$$

is called the *open ball of center a and radius ρ*, and the set

$$S(a, \rho) = \{x \in E \mid d(a, x) = \rho\}$$

is called the *sphere of center a and radius ρ*. It should be noted that ρ is finite (i.e., not $+\infty$). A subset X of a metric space E is *bounded* if there is a closed ball $B(a, \rho)$ such that $X \subseteq B(a, \rho)$.

Clearly, $B(a, \rho) = B_0(a, \rho) \cup S(a, \rho)$.

Example 2.2.

1. In $E = \mathbb{R}$ with the distance $|x - y|$, an open ball of center a and radius ρ is the open interval $(a - \rho, a + \rho)$.

2. In $E = \mathbb{R}^2$ with the Euclidean metric, an open ball of center a and radius ρ is the set of points inside the disk of center a and radius ρ, excluding the boundary points on the circle.

3. In $E = \mathbb{R}^3$ with the Euclidean metric, an open ball of center a and radius ρ is the set of points inside the sphere of center a and radius ρ, excluding the boundary points on the sphere.

One should be aware that intuition can be misleading in forming a geometric image of a closed (or open) ball. For example, if d is the discrete metric, a closed ball of center a and radius $\rho < 1$ consists only of its center a, and a closed ball of center a and radius $\rho \geq 1$ consists of the entire space!

If $E = [a, b]$, and $d(x, y) = |x - y|$, as in Example 2.1, an open ball $B_0(a, \rho)$, with $\rho < b - a$, is in fact the interval $[a, a + \rho)$, which is closed on the left.

We now consider a very important special case of metric spaces, normed vector spaces. Normed vector spaces have already been defined in Chapter 8 (Vol. I) (Definition 8.1 (Vol. I)), but for the reader's convenience we repeat the definition.

Definition 2.3. Let E be a vector space over a field K, where K is either the field $\mathbb{R}$ of reals, or the field $\mathbb{C}$ of complex numbers. A *norm on* E is a function $\| \| \colon E \to \mathbb{R}_+$, assigning a nonnegative real number $\|u\|$ to any vector $u \in E$, and satisfying the following conditions for all $x, y \in E$:

(N1) $\|x\| \geq 0$, and $\|x\| = 0$ iff $x = 0$. (positivity)

(N2) $\|\lambda x\| = |\lambda| \, \|x\|$. (homogeneity (or scaling))

(N3) $\|x + y\| \leq \|x\| + \|y\|$. (triangle inequality)

A vector space E together with a norm $\| \|$ is called a *normed vector space*.

We showed in Chapter 8 (Vol. I), that

$$\|-x\| = \|x\|,$$

and from (N3), we get

$$|\|x\| - \|y\|| \leq \|x - y\|.$$

Given a normed vector space E, if we define d such that

$$d(x,\, y) = \|x - y\|,$$

it is easily seen that d is a metric. *Thus, every normed vector space is immediately a metric space.* Note that the metric associated with a norm is invariant under translation, that is,

$$d(x + u,\, y + u) = d(x,\, y).$$

For this reason we can restrict ourselves to open or closed balls of center 0.

Examples of normed vector spaces were given in Example 8.1 (Vol. I). We repeat the most important examples.

Example 2.3. Let $E = \mathbb{R}^n$ (or $E = \mathbb{C}^n$). There are three standard norms. For every $(x_1, \ldots, x_n) \in E$, we have the norm $\|x\|_1$, defined such that,

$$\|x\|_1 = |x_1| + \cdots + |x_n|,$$

we have the *Euclidean norm* $\|x\|_2$, defined such that,

$$\|x\|_2 = \left(|x_1|^2 + \cdots + |x_n|^2\right)^{\frac{1}{2}},$$

and the *sup*-norm $\|x\|_\infty$, defined such that,

$$\|x\|_\infty = \max\{|x_i| \mid 1 \leq i \leq n\}.$$

More generally, we define the *ℓ^p-norm* (for $p \geq 1$) by

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}.$$

We proved in Proposition 8.1 (Vol. I) that the ℓ^p-norms are indeed norms. The closed unit balls centered at $(0, 0)$ for $\| \|_1$, $\| \|_2$, and $\| \|_\infty$, along with the containment relationships, are shown in Figures 2.1 and 2.2. Figures 2.3 and 2.4 illustrate the situation in $\mathbb{R}^3$.

Remark: In a normed vector space we define a closed ball or an open ball of radius ρ as a closed ball or an open ball of center 0. We may use the notation $B(\rho)$ for $B(0, \rho)$ and $B_0(\rho)$ for $B_0(0, \rho)$.

We will now define the crucial notions of open sets and closed sets within a metric space

Definition 2.4. Let E be a metric space with metric d. A subset $U \subseteq E$ is an *open set* in E if either $U = \emptyset$, or for every $a \in U$, there is some open ball $B_0(a, \rho)$ such that, $B_0(a, \rho) \subseteq U$.[1] A subset $F \subseteq E$ is a *closed set* in E if its complement $E - F$ is open in E. See Figure 2.5.

[1] Recall that $\rho > 0$.

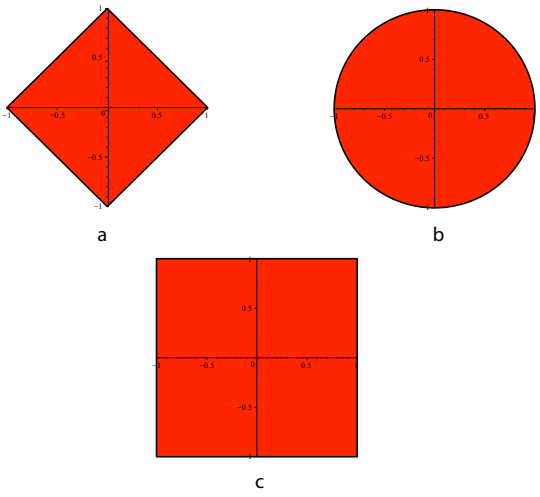


Figure 2.1: Figure a shows the diamond shaped closed ball associated with $\| \|_1$. Figure b shows the closed unit disk associated with $\| \|_2$, while Figure c illustrates the closed unit ball associated with $\| \|_\infty$.

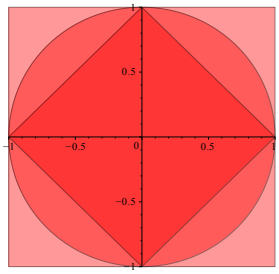


Figure 2.2: The relationship between the closed unit balls centered at $(0, 0)$.

The set E itself is open, since for every $a \in E$, every open ball of center a is contained in E. In $E = \mathbb{R}^n$, given n intervals $[a_i, b_i]$, with $a_i < b_i$, it is easy to show that the open n-cube

$$\{(x_1, \ldots, x_n) \in E \mid a_i < x_i < b_i,\ 1 \leq i \leq n\}$$

is an open set. In fact, it is possible to find a metric for which such open n-cubes are open balls! Similarly, we can define the closed n-cube

$$\{(x_1, \ldots, x_n) \in E \mid a_i \leq x_i \leq b_i,\ 1 \leq i \leq n\},$$

which is a closed set.

The open sets satisfy some important properties that lead to the definition of a topological space.

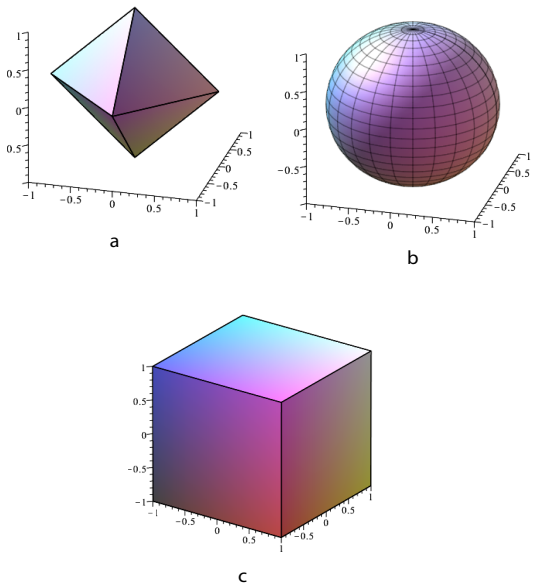


Figure 2.3: Figure a shows the octahedral shaped closed ball associated with $\| \ \|_1$. Figure b shows the closed spherical associated with $\| \ \|_2$, while Figure c illustrates the closed unit ball associated with $\| \ \|_\infty$.

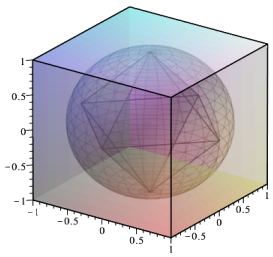


Figure 2.4: The relationship between the closed unit balls centered at $(0, 0, 0)$.

Proposition 2.1. *Given a metric space E with metric d, the family $\mathcal{O}$ of all open sets defined in Definition 2.4 satisfies the following properties:*

(O1) For every finite family $(U_i)_{1 \leq i \leq n}$ of sets $U_i \in \mathcal{O}$, we have $U_1 \cap \cdots \cap U_n \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under finite intersections.

(O2) For every arbitrary family $(U_i)_{i \in I}$ of sets $U_i \in \mathcal{O}$, we have $\bigcup_{i \in I} U_i \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under arbitrary unions.

(O3) $\emptyset \in \mathcal{O}$, and $E \in \mathcal{O}$, i.e., $\emptyset$ and E belong to $\mathcal{O}$.

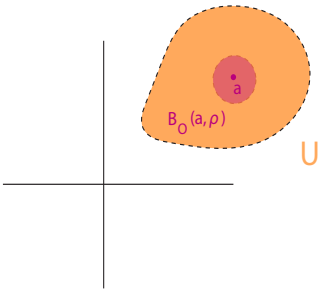


Figure 2.5: An open set U in $E = \mathbb{R}^2$ under the standard Euclidean metric. Any point in the peach set U is surrounded by a small raspberry open set which lies within U.

Furthermore, for any two distinct points $a \neq b$ in E, there exist two open sets U_a and U_b such that, $a \in U_a$, $b \in U_b$, and $U_a \cap U_b = \emptyset$.

Proof. It is straightforward. For the last point, letting $\rho = d(a, b)/3$ (in fact $\rho = d(a, b)/2$ works too), we can pick $U_a = B_0(a, \rho)$ and $U_b = B_0(b, \rho)$. By the triangle inequality, we must have $U_a \cap U_b = \emptyset$. □

The above proposition leads to the very general concept of a topological space.

One should be careful that, in general, the family of open sets is not closed under infinite intersections. For example, in $\mathbb{R}$ under the metric $|x - y|$, letting $U_n = (-1/n, +1/n)$, each U_n is open, but $\bigcap_n U_n = \{0\}$, which is not open.

2.2 Topological Spaces

Motivated by Proposition 2.1, a topological space is defined in terms of a family of sets satisfying the properties of open sets stated in that proposition.

Definition 2.5. Given a set E, a *topology on E (or a topological structure on E)*, is defined as a family $\mathcal{O}$ of subsets of E called *open sets*, and satisfying the following three properties:

(1) For every finite family $(U_i)_{1 \leq i \leq n}$ of sets $U_i \in \mathcal{O}$, we have $U_1 \cap \cdots \cap U_n \in \mathcal{O}$, *i.e.*, $\mathcal{O}$ is closed under finite intersections.

(2) For every arbitrary family $(U_i)_{i \in I}$ of sets $U_i \in \mathcal{O}$, we have $\bigcup_{i \in I} U_i \in \mathcal{O}$, *i.e.*, $\mathcal{O}$ is closed under arbitrary unions.

(3) $\emptyset \in \mathcal{O}$, and $E \in \mathcal{O}$, i.e., $\emptyset$ and E belong to $\mathcal{O}$.

A set E together with a topology $\mathcal{O}$ on E is called a *topological space*. Given a topological space $(E, \mathcal{O})$, a subset F of E is a *closed set* if $F = E - U$ for some open set $U \in \mathcal{O}$, i.e., F is the complement of some open set.

It is possible that an open set is also a closed set. For example, $\emptyset$ and E are both open and closed.

Definition 2.6. When a topological space contains a proper nonempty subset U which is both open and closed, the space E is said to be *disconnected*.

By taking complements, we can state properties of the closed sets dual to those of Definition 2.5. If we denote the family of closed sets of E as $\mathcal{F} = \{F \subseteq E \mid E - F \in \mathcal{O}\}$, then the closed sets satisfy the following properties:

(1) For every finite family $(F_i)_{1 \leq i \leq n} \in \mathcal{F}$, we have $F_1 \cup \cdots \cup F_n \in \mathcal{F}$, *i.e.*, $\mathcal{F}$ is closed under finite unions.

(2) For every arbitrary family $(F_i)_{i \in I}$ of sets $F_i \in \mathcal{F}$, we have $\bigcap_{i \in I} F_i \in \mathcal{F}$, *i.e.*, $\mathcal{F}$ is closed under arbitrary intersections.

(3) $\emptyset \in \mathcal{F}$, and $E \in \mathcal{F}$, i.e., $\emptyset$ and E belong to $\mathcal{F}$.

One of the reasons why topological spaces are important is that the definition of a topology only involves a certain family $\mathcal{O}$ of sets, and not **how** such family is generated from a metric or a norm. For example, different metrics or different norms can define the same family of open sets. Many topological properties only depend on the family $\mathcal{O}$ and not on the specific metric or norm. But the fact that a topology is definable from a metric or a norm is important, because it usually implies nice properties of a space. *All our examples will be spaces whose topology is defined by a metric or a norm.*

Definition 2.7. A topological space $(E, \mathcal{O})$ is said to satisfy the *Hausdorff separation axiom (or T_2-separation axiom)* if for any two distinct points $a \neq b$ in E, there exist two open sets U_a and U_b such that, $a \in U_a$, $b \in U_b$, and $U_a \cap U_b = \emptyset$. When the T_2-separation axiom is satisfied, we also say that $(E, \mathcal{O})$ is a *Hausdorff space*.

As shown by Proposition 2.1, any metric space is a topological Hausdorff space, the family of open sets being in fact the family of arbitrary unions of open balls. Similarly, any normed vector space is a topological Hausdorff space, the family of open sets being the family of arbitrary unions of open balls. The topology $\mathcal{O}$ consisting of all subsets of E is called the *discrete topology*.

Remark: Most (if not all) spaces used in analysis are Hausdorff spaces. Intuitively, the Hausdorff separation axiom says that there are enough "small" open sets. Without this axiom, some counter-intuitive behaviors may arise. For example, a sequence may have more than one limit point (or a compact set may not be closed). Nevertheless, non-Hausdorff topological spaces arise naturally in algebraic geometry. But even there, some substitute for separation is used.

It is also worth noting that the Hausdorff separation axiom implies the following property.

Proposition 2.2. *If a topological space $(E, \mathcal{O})$ is Hausdorff, then for every $a \in E$, the set $\{a\}$ is closed.*

Proof. If $x \in E - \{a\}$, then $x \neq a$, and so there exist open sets U_a and U_x such that $a \in U_a$, $x \in U_x$, and $U_a \cap U_x = \emptyset$. See Figure 2.6. Thus, for every $x \in E - \{a\}$, there is an open set U_x containing x and contained in $E - \{a\}$, showing by (O3) that $E - \{a\}$ is open, and thus that the set $\{a\}$ is closed. □

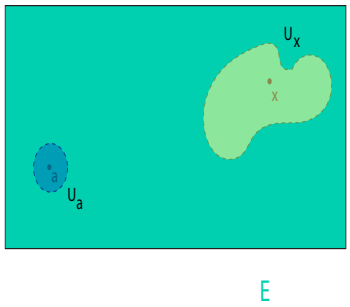


Figure 2.6: A schematic illustration of the Hausdorff separation property.

Given a topological space $(E, \mathcal{O})$, given any subset A of E, since $E \in \mathcal{O}$ and E is a closed set, the family $\mathcal{C}_A = \{F \mid A \subseteq F, F \text{ a closed set}\}$ of closed sets containing A is nonempty, and since any arbitrary intersection of closed sets is a closed set, the intersection $\bigcap \mathcal{C}_A$ of the sets in the family $\mathcal{C}_A$ is the smallest closed set containing A. By a similar reasoning, the union of all the open subsets contained in A is the largest open set contained in A.

Definition 2.8. Given a topological space $(E, \mathcal{O})$, given any subset A of E, the smallest closed set containing A is denoted by $\overline{A}$, and is called the *closure, or adherence of* A. See Figure 2.7. A subset A of E is *dense in* E if $\overline{A} = E$. The largest open set contained in A is denoted by $\overset{\circ}{A}$, and is called the *interior of* A. See Figure 2.8. The set $\mathrm{Fr}\, A = \overline{A} \cap \overline{E - A}$ is called the *boundary (or frontier) of* A. We also denote the boundary of A by ∂A. See Figure 2.9.

Remark: The notation $\overline{A}$ for the closure of a subset A of E is somewhat unfortunate, since $\overline{A}$ is often used to denote the set complement of A in E. Still, we prefer it to more cumbersome notations such as $\mathrm{clo}(A)$, and we denote the complement of A in E by $E - A$ (or sometimes, A^c).

By definition, it is clear that a subset A of E is closed iff $A = \overline{A}$. The set $\mathbb{Q}$ of rationals is dense in $\mathbb{R}$. It is easily shown that $\overline{A} = \overset{\circ}{A} \cup \partial A$ and $\overset{\circ}{A} \cap \partial A = \emptyset$.

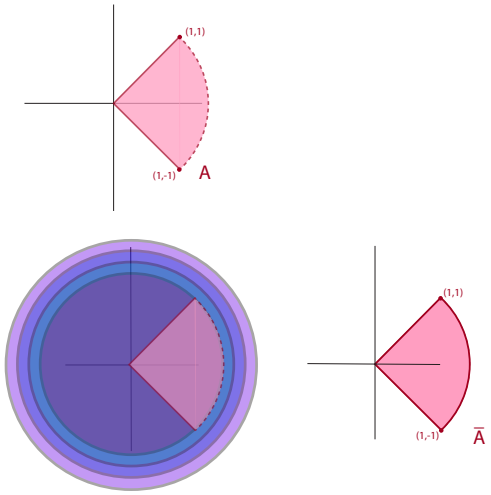


Figure 2.7: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The subset A is the section $B_0(1)$ in the first and fourth quadrants bound by the lines $y = x$ and $y = -x$. The closure of A is obtained by the intersection of A with the closed unit ball.

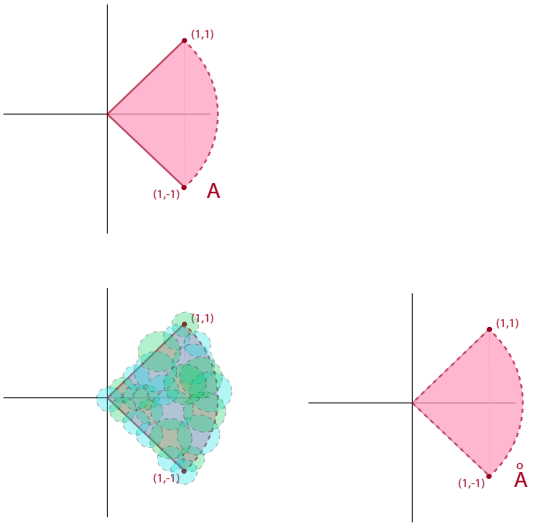


Figure 2.8: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The subset A is the section $B_0(1)$ in the first and fourth quadrants bound by the lines $y = x$ and $y = -x$. The interior of A is obtained by the covering A with small open balls.

Another useful characterization of $\overline{A}$ is given by the following proposition.

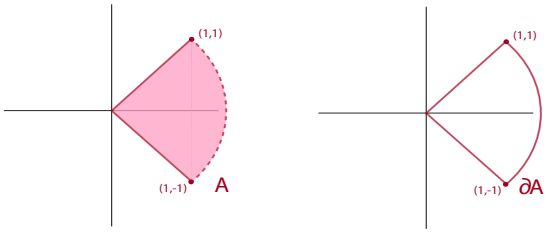


Figure 2.9: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The subset A is the section $B_0(1)$ in the first and fourth quadrants bound by the lines $y = x$ and $y = -x$. The boundary of A is $\overline{A} - \overset{\circ}{A}$.

Proposition 2.3. *Given a topological space $(E, \mathcal{O})$, given any subset A of E, the closure $\overline{A}$ of A is the set of all points $x \in E$ such that for every open set U containing x, then $U \cap A \neq \emptyset$. See Figure 2.10.*

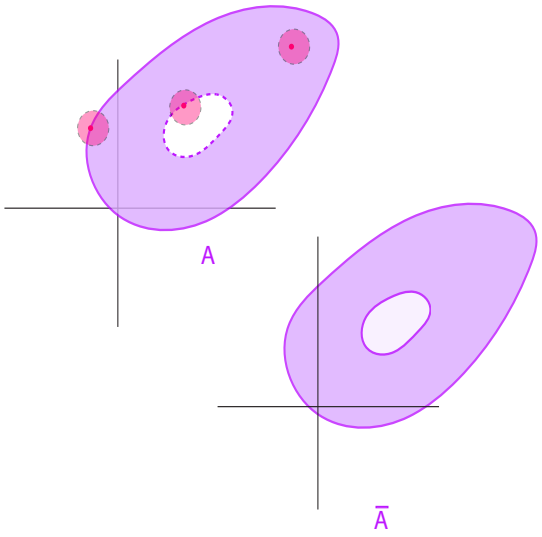


Figure 2.10: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The purple subset A is illustrated with three red points, each in its closure since the open ball centered at each point has nontrivial intersection with A.

Proof. If $A = \emptyset$, since $\emptyset$ is closed, the proposition holds trivially. Thus assume that $A \neq \emptyset$. First assume that $x \in \overline{A}$. Let U be any open set such that $x \in U$. If $U \cap A = \emptyset$, since U is open, then $E - U$ is a closed set containing A, and since $\overline{A}$ is the intersection of all closed sets containing A, we must have $x \in E - U$, which is impossible. Conversely, assume that

$x \in E$ is a point such that for every open set U containing x, $U \cap A \neq \emptyset$. Let F be any closed subset containing A. If $x \notin F$, since F is closed, then $U = E - F$ is an open set such that $x \in U$, and $U \cap A = \emptyset$, a contradiction. Thus, we have $x \in F$ for every closed set containing A, that is, $x \in \overline{A}$. □

Often it is necessary to consider a subset A of a topological space E, and to view the subset A as a topological space.

2.3 Subspace and Product Topologies

The following proposition shows how to define a topology on a subset.

Proposition 2.4. *Given a topological space $(E, \mathcal{O})$, given any subset A of E, let*

$$\mathcal{U} = \{U \cap A \mid U \in \mathcal{O}\}$$

be the family of all subsets of A obtained as the intersection of any open set in $\mathcal{O}$ with A. The following properties hold.

(1) The space $(A, \mathcal{U})$ is a topological space.

(2) If E is a metric space with metric d, then the restriction $d_A \colon A \times A \to \mathbb{R}_+$ of the metric d to A defines a metric space. Furthermore, the topology induced by the metric d_A agrees with the topology defined by $\mathcal{U}$, as above.

Proof. Left as an exercise. □

Proposition 2.4 suggests the following definition.

Definition 2.9. Given a topological space $(E, \mathcal{O})$, given any subset A of E, the *subspace topology on A induced by $\mathcal{O}$* is the family $\mathcal{U}$ of open sets defined such that

$$\mathcal{U} = \{U \cap A \mid U \in \mathcal{O}\}$$

is the family of all subsets of A obtained as the intersection of any open set in $\mathcal{O}$ with A. We say that $(A, \mathcal{U})$ has the *subspace topology*. If (E, d) is a metric space, the restriction $d_A \colon A \times A \to \mathbb{R}_+$ of the metric d to A is called the *subspace metric*.

For example, if $E = \mathbb{R}^n$ and d is the Euclidean metric, we obtain the subspace topology on the closed n-cube

$$\{(x_1, \ldots, x_n) \in E \mid a_i \leq x_i \leq b_i,\ 1 \leq i \leq n\}.$$

See Figure 2.11.

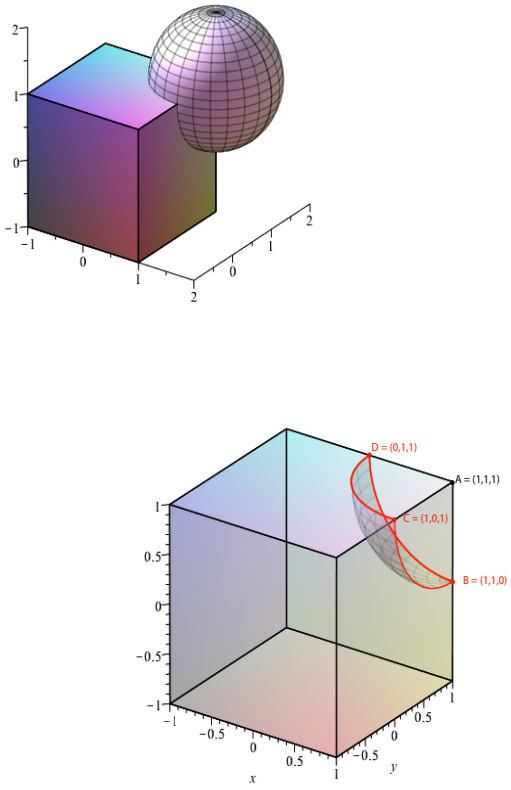


Figure 2.11: An example of an open set in the subspace topology for $\{(x, y, z) \in \mathbb{R}^3 \mid -1 \leq x \leq 1, -1 \leq y \leq 1, -1 \leq z \leq 1\}$. The open set is the corner region $ABCD$ and is obtained by intersection the cube $B_0((1, 1, 1), 1)$.

One should realize that every open set $U \in \mathcal{O}$ which is entirely contained in A is also in the family $\mathcal{U}$, but $\mathcal{U}$ may contain open sets that are not in $\mathcal{O}$. For example, if $E = \mathbb{R}$ with $|x - y|$, and $A = [a, b]$, then sets of the form $[a, c)$, with $a < c < b$ belong to $\mathcal{U}$, but they are not open sets for $\mathbb{R}$ under $|x - y|$. However, there is agreement in the following situation.

Proposition 2.5. *Given a topological space $(E, \mathcal{O})$, given any subset A of E, if $\mathcal{U}$ is the subspace topology, then the following properties hold.*

(1) If A is an open set $A \in \mathcal{O}$, then every open set $U \in \mathcal{U}$ is an open set $U \in \mathcal{O}$.

(2) If A is a closed set in E, then every closed set w.r.t. the subspace topology is a closed set w.r.t. $\mathcal{O}$.

Proof. Left as an exercise. □

The concept of product topology is also useful. We have the following proposition.

Proposition 2.6. *Given n topological spaces $(E_i, \mathcal{O}_i)$, let $\mathcal{B}$ be the family of subsets of $E_1 \times \cdots \times E_n$ defined as follows:*

$$\mathcal{B} = \{U_1 \times \cdots \times U_n \mid U_i \in \mathcal{O}_i,\ 1 \leq i \leq n\},$$

and let $\mathcal{P}$ be the family consisting of arbitrary unions of sets in $\mathcal{B}$, including $\emptyset$. Then $\mathcal{P}$ is a topology on $E_1 \times \cdots \times E_n$.

Proof. Left as an exercise. □

Definition 2.10. Given n topological spaces $(E_i, \mathcal{O}_i)$, the *product topology on* $E_1 \times \cdots \times E_n$ is the family $\mathcal{P}$ of subsets of $E_1 \times \cdots \times E_n$ defined as follows: if

$$\mathcal{B} = \{U_1 \times \cdots \times U_n \mid U_i \in \mathcal{O}_i,\ 1 \leq i \leq n\},$$

then $\mathcal{P}$ is the family consisting of arbitrary unions of sets in $\mathcal{B}$, including $\emptyset$. See Figure 2.12.

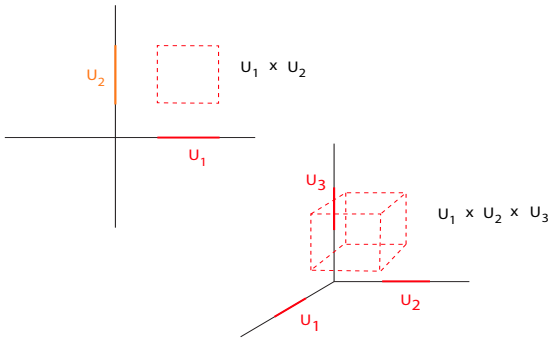


Figure 2.12: Examples of open sets in the product topology for $\mathbb{R}^2$ and $\mathbb{R}^3$ induced by the Euclidean metric.

If each (E_i, d_{E_i}) is a metric space, there are three natural metrics that can be defined on $E_1 \times \cdots \times E_n$:

$$\begin{aligned}
d_1((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &= d_{E_1}(x_1, y_1) + \cdots + d_{E_n}(x_n, y_n),\\
d_2((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &= \left((d_{E_1}(x_1, y_1))^2 + \cdots + (d_{E_n}(x_n, y_n))^2\right)^{\frac{1}{2}},\\
d_\infty((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &= \max\{d_{E_1}(x_1, y_1), \ldots, d_{E_n}(x_n, y_n)\}.
\end{aligned}$$

Proposition 2.7. *The following inequalities hold:*

$$\begin{aligned}
d_\infty((x_1, \ldots, x_n), (y_1, \ldots, y_n)) \leq d_2((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &\leq d_1((x_1, \ldots, x_n), (y_1, \ldots, y_n))\\
&\leq n d_\infty((x_1, \ldots, x_n), (y_1, \ldots, y_n)),
\end{aligned}$$

so these distances define the same topology, which is the product topology.

If each $(E_i, \| \|_{E_i})$ is a normed vector space, there are three natural norms that can be defined on $E_1 \times \cdots \times E_n$:

$$\|(x_1, \ldots, x_n)\|_1 = \|x_1\|_{E_1} + \cdots + \|x_n\|_{E_n},$$
$$\|(x_1, \ldots, x_n)\|_2 = \left(\|x_1\|_{E_1}^2 + \cdots + \|x_n\|_{E_n}^2\right)^{\frac{1}{2}},$$
$$\|(x_1, \ldots, x_n)\|_\infty = \max\left\{\|x_1\|_{E_1}, \ldots, \|x_n\|_{E_n}\right\}.$$

Proposition 2.8. *The following inequalities hold:*

$$\|(x_1, \ldots, x_n)\|_\infty \leq \|(x_1, \ldots, x_n)\|_2 \leq \|(x_1, \ldots, x_n)\|_1 \leq n\|(x_1, \ldots, x_n)\|_\infty,$$

so these norms define the same topology, which is the product topology.

It can also be verified that when $E_i = \mathbb{R}$, with the standard topology induced by $|x - y|$, the topology product on $\mathbb{R}^n$ is the standard topology induced by the Euclidean norm.

Definition 2.11. Two metrics d_1 and d_2 on a space E are *equivalent* if they induce the same topology $\mathcal{O}$ on E (i.e., they define the same family $\mathcal{O}$ of open sets). Similarly, two norms $\| \|_1$ and $\| \|_2$ on a space E are *equivalent* if they induce the same topology $\mathcal{O}$ on E.

Given a topological space $(E, \mathcal{O})$, it is often useful, as in Proposition 2.6, to define the topology $\mathcal{O}$ in terms of a subfamily $\mathcal{B}$ of subsets of E.

Definition 2.12. Given a topological space $(E, \mathcal{O})$, we say that a family $\mathcal{B}$ of subsets of E is a *basis for the topology* $\mathcal{O}$, if $\mathcal{B}$ is a subset of $\mathcal{O}$, and if every open set U in $\mathcal{O}$ can be obtained as some union (possibly infinite) of sets in $\mathcal{B}$ (agreeing that the empty union is the empty set).

For example, given any metric space (E, d), $\mathcal{B} = \{B_0(a, \rho) \mid a \in E, \rho > 0\}$ is a basis for the topology. In particular, if $d = \| \|_2$, the open intervals form a basis for $\mathbb{R}$, while the open disks form a basis for $\mathbb{R}^2$. The open rectangles also form a basis for $\mathbb{R}^2$ with the standard topology.

It is immediately verified that if a family $\mathcal{B} = (U_i)_{i \in I}$ is a basis for the topology of $(E, \mathcal{O})$, then $E = \bigcup_{i \in I} U_i$, and the intersection of any two sets $U_i, U_j \in \mathcal{B}$ is the union of some sets in the family $\mathcal{B}$ (again, agreeing that the empty union is the empty set). Conversely, a family $\mathcal{B}$ with these properties is the basis of the topology obtained by forming arbitrary unions of sets in $\mathcal{B}$.

Definition 2.13. Given a topological space $(E, \mathcal{O})$, a *subbasis for* $\mathcal{O}$ is a family $\mathcal{S}$ of subsets of E, such that the family $\mathcal{B}$ of all finite intersections of sets in $\mathcal{S}$ (including E itself, in case of the empty intersection) is a basis of $\mathcal{O}$. See Figure 2.13.

The following proposition gives useful criteria for determining whether a family of open subsets is a basis of a topological space.

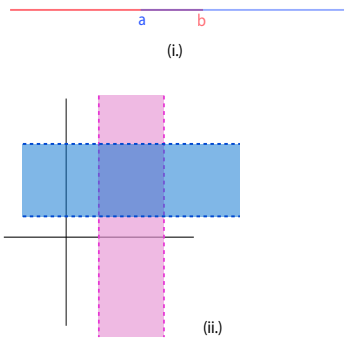


Figure 2.13: Figure $(i.)$ shows that the set of infinite open intervals forms a subbasis for $\mathbb{R}$. Figure $(ii.)$ shows that the infinite open strips form a subbasis for $\mathbb{R}^2$.

Proposition 2.9. *Given a topological space $(E, \mathcal{O})$ and a family $\mathcal{B}$ of open subsets in $\mathcal{O}$ the following properties hold:*

(1) The family $\mathcal{B}$ is a basis for the topology $\mathcal{O}$ iff for every open set $U \in \mathcal{O}$ and every $x \in U$, there is some $B \in \mathcal{B}$ such that $x \in B$ and $B \subseteq U$. See Figure 2.14.

(2) The family $\mathcal{B}$ is a basis for the topology $\mathcal{O}$ iff

(a) For every $x \in E$, there is some $B \in \mathcal{B}$ such that $x \in B$.

(b) For any two open subsets, $B_1, B_2 \in \mathcal{B}$, for every $x \in E$, if $x \in B_1 \cap B_2$, then there is some $B_3 \in \mathcal{B}$ such that $x \in B_3$ and $B_3 \subseteq B_1 \cap B_2$. See Figure 2.15.

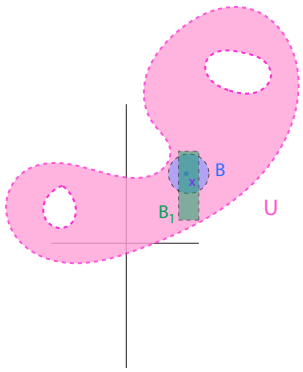


Figure 2.14: Given an open subset U of $\mathbb{R}^2$ and $x \in U$, there exists an open ball B containing x with $B \subset U$. There also exists an open rectangle B_1 containing x with $B_1 \subset U$.

We now consider the fundamental property of continuity.

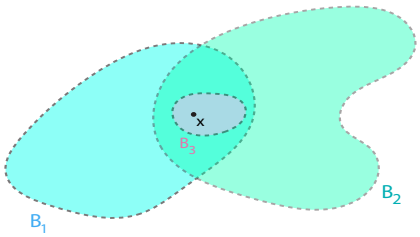


Figure 2.15: A schematic illustration of Condition (b) in Proposition 2.9.

2.4 Continuous Functions

Definition 2.14. Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, and let $f\colon E \to F$ be a function. For every $a \in E$, we say that f *is continuous at* a, if for every open set $V \in \mathcal{O}_F$ containing $f(a)$, there is some open set $U \in \mathcal{O}_E$ containing a, such that, $f(U) \subseteq V$. See Figure 2.16. We say that f *is continuous* if it is continuous at every $a \in E$.

If $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ are topological spaces, and $f\colon E \to F$ is a function, for every nonempty subset $A \subseteq E$ of E, we say that f *is continuous on* A if the restriction of f to A is continuous with respect to $(A, \mathcal{U})$ and $(F, \mathcal{O}_F)$, where $\mathcal{U}$ is the subspace topology induced by $\mathcal{O}_E$ on A.

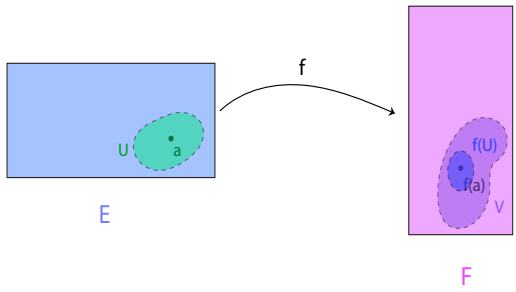


Figure 2.16: A schematic illustration of Definition 2.14.

Definition 2.15. Let $(E, \mathcal{O}_E)$ be a topological space. Define a *neighborhood of* $a \in E$ as any subset N of E containing some open set $O \in \mathcal{O}$ such that $a \in O$.

Now if f is continuous at a and N is any neighborhood of $f(a)$, there is some open set $V \subseteq N$ containing $f(a)$, and since f is continuous at a, there is some open set U containing a, such that $f(U) \subseteq V$. Since $V \subseteq N$, the open set U is a subset of $f^{-1}(N)$ containing a, and $f^{-1}(N)$ is a neighborhood of a. Conversely, if $f^{-1}(N)$ is a neighborhood of a whenever N is any neighborhood of $f(a)$, it is immediate that f is continuous at a. See Figure 2.17.

It is easy to see that Definition 2.14 is equivalent to the following statements.

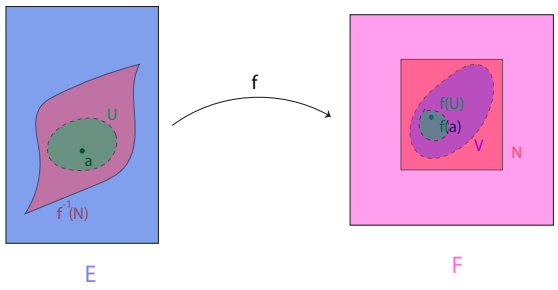


Figure 2.17: A schematic illustration of the neighborhood condition.

Proposition 2.10. *Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, and let $f\colon E \to F$ be a function. For every $a \in E$, the function f is continuous at $a \in E$ iff for every neighborhood N of $f(a) \in F$, then $f^{-1}(N)$ is a neighborhood of a. The function f is continuous on E iff $f^{-1}(V)$ is an open set in $\mathcal{O}_E$ for every open set $V \in \mathcal{O}_F$.*

If E and F are metric spaces, Proposition 2.10 can be restated as follows.

Proposition 2.11. *Let E and F be metric spaces defined by metrics d_1 and d_2. The function $f\colon E \to F$ is continuous at $a \in E$ iff for every $\epsilon > 0$, there is some $\eta > 0$ such that for every $x \in E$,*

$$\text{if } d_1(a, x) \leq \eta, \text{ then } d_2(f(a), f(x)) \leq \epsilon.$$

If E and F are normed vector spaces, Proposition 2.10 can be restated as follows.

Proposition 2.12. *Let E and F be normed vector spaces defined by norms $\| \|_1$ and $\| \|_2$. The function $f\colon E \to F$ is continuous at $a \in E$ iff for every $\epsilon > 0$, there is some $\eta > 0$ such that for every $x \in E$,*

$$\text{if } \|x - a\|_1 \leq \eta, \text{ then } \|f(x) - f(a)\|_2 \leq \epsilon.$$

It is worth noting that continuity is a topological notion, in the sense that equivalent metrics (or equivalent norms) define exactly the same notion of continuity.

An important example of a continuous function is the distance function in a metric space. One can show that in a metric space (E, d), the distance $d\colon E \times E \to \mathbb{R}$ is continuous, where $E \times E$ has the product topology. By the triangle inequality, we have

$$d(x, y) \leq d(x, x_0) + d(x_0, y_0) + d(y_0, y) = d(x_0, y_0) + d(x_0, x) + d(y_0, y)$$

and

$$d(x_0, y_0) \leq d(x_0, x) + d(x, y) + d(y, y_0) = d(x, y) + d(x_0, x) + d(y_0, y).$$

Consequently,

$$|d(x, y) - d(x_0, y_0)| \leq d(x_0, x) + d(y_0, y),$$

which proves that d is continuous at (x_0, y_0). In fact this shows that d is uniformly continuous; see Definition 2.21.

Similarly, for a normed vector space $(E, \| \|)$, the norm $\| \| \colon E \to \mathbb{R}$ is (uniformly) continuous.

Another important example of a continuous function is the projection of a product space. Given a product $E_1 \times \cdots \times E_n$ of topological spaces, as usual, we let $\pi_i \colon E_1 \times \cdots \times E_n \to E_i$ be the projection function such that, $\pi_i(x_1, \ldots, x_n) = x_i$. It is immediately verified that each π_i is continuous.

Definition 2.16. Given a topological space $(E, \mathcal{O})$, we say that a point $a \in E$ is *isolated* if $\{a\}$ is an open set in $\mathcal{O}$.

If $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ are topological spaces, any function $f \colon E \to F$ is continuous at every isolated point $a \in E$. In the discrete topology, every point is isolated.

As the following proposition shows, isolated points *do not* occur in nontrivial metric spaces.

Proposition 2.13. *In a nontrivial normed vector space $(E, \| \|)$ (with $E \neq \{0\}$), no point is isolated.*

Proof. To show this, we show that every open ball $B_0(u, \rho,)$ contains some vectors different from u. Indeed, since E is nontrivial, there is some $v \in E$ such that $v \neq 0$, and thus $\lambda = \|v\| > 0$ (by (N1)). Let

$$w = u + \frac{\rho}{\lambda + 1} v.$$

Since $v \neq 0$ and $\rho > 0$, we have $w \neq u$. Then,

$$\|w - u\| = \left\| \frac{\rho}{\lambda + 1} v \right\| = \frac{\rho\lambda}{\lambda + 1} < \rho,$$

which shows that $\|w - u\| < \rho$, for $w \neq u$. □

The following proposition shows that composition behaves well with respect to continuity.

Proposition 2.14. *Given topological spaces $(E, \mathcal{O}_E)$, $(F, \mathcal{O}_F)$, and $(G, \mathcal{O}_G)$, and two functions $f \colon E \to F$ and $g \colon F \to G$, if f is continuous at $a \in E$ and g is continuous at $f(a) \in F$, then $g \circ f \colon E \to G$ is continuous at $a \in E$. Given n topological spaces $(F_i, \mathcal{O}_i)$, for every function $f \colon E \to F_1 \times \cdots \times F_n$, then f is continuous at $a \in E$ iff every $f_i \colon E \to F_i$ is continuous at a, where $f_i = \pi_i \circ f$.*

Given a function $f \colon E_1 \times \cdots \times E_n \to F$, we can fix $n - 1$ of the arguments, say $a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n$, and view f as a function of the remaining argument,

$$x_i \mapsto f(a_1, \ldots, a_{i-1}, x_i, a_{i+1}, \ldots, a_n),$$

where $x_i \in E_i$. If f is continuous, it is clear that each f_i is continuous.

One should be careful that the converse is false! For example, consider the function $f\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$, defined such that,

$$f(x,y)=\frac{xy}{x^2+y^2}\quad \text{if } (x,y)\neq(0,0),\quad \text{and}\quad f(0,0)=0.$$

The function f is continuous on $\mathbb{R}\times\mathbb{R}-\{(0,0)\}$, but on the line $y=mx$, with $m\neq 0$, we have $f(x,y)=\frac{m}{1+m^2}\neq 0$, and thus, on this line, $f(x,y)$ does not approach 0 when (x,y) approaches $(0,0)$. See Figure 2.18.

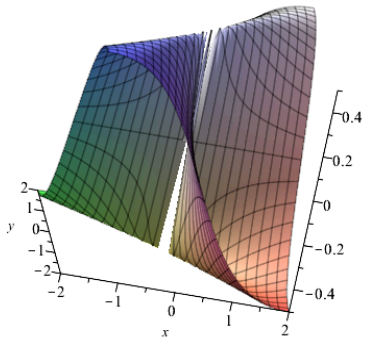


Figure 2.18: The graph of $f(x,y)=\frac{xy}{x^2+y^2}$ for $(x,y)\neq(0,0)$. The bottom of this graph, which shows the approach along the line $y=-x$, does not have a z value of 0.

The following proposition is useful for showing that real-valued functions are continuous.

Proposition 2.15. *If E is a topological space, and $(\mathbb{R},|x-y|)$ the reals under the standard topology, for any two functions $f\colon E\to\mathbb{R}$ and $g\colon E\to\mathbb{R}$, for any $a\in E$, for any $\lambda\in\mathbb{R}$, if f and g are continuous at a, then $f+g$, λf, $f\cdot g$ are continuous at a, and f/g is continuous at a if $g(a)\neq 0$.*

Proof. Left as an exercise. □

Using Proposition 2.15, we can show easily that every real polynomial function is continuous.

The notion of isomorphism of topological spaces is defined as follows.

Definition 2.17. Let $(E,\mathcal{O}_E)$ and $(F,\mathcal{O}_F)$ be topological spaces, and let $f\colon E\to F$ be a function. We say that f *is a homeomorphism between E and F* if f is bijective, and both $f\colon E\to F$ and $f^{-1}\colon F\to E$ are continuous.

One should be careful that a bijective continuous function $f: E \to F$ is not necessarily a homeomorphism. For example, if $E = \mathbb{R}$ with the discrete topology, and $F = \mathbb{R}$ with the standard topology, the identity is not a homeomorphism. Another interesting example involving a parametric curve is given below. Let $L: \mathbb{R} \to \mathbb{R}^2$ be the function, defined such that

$$L_1(t) = \frac{t(1+t^2)}{1+t^4},$$
$$L_2(t) = \frac{t(1-t^2)}{1+t^4}.$$

If we think of $(x(t), y(t)) = (L_1(t), L_2(t))$ as a geometric point in $\mathbb{R}^2$, the set of points $(x(t), y(t))$ obtained by letting t vary in $\mathbb{R}$ from $-\infty$ to $+\infty$, defines a curve having the shape of a "figure eight," with self-intersection at the origin, called the "lemniscate of Bernoulli." See Figure 2.19. The map L is continuous, and in fact bijective, but its inverse L^{-1} is not continuous. Indeed, when we approach the origin on the branch of the curve in the upper left quadrant (i.e., points such that, $x \leq 0$, $y \geq 0$), then t goes to $-\infty$, and when we approach the origin on the branch of the curve in the lower right quadrant (i.e., points such that, $x \geq 0$, $y \leq 0$), then t goes to $+\infty$.

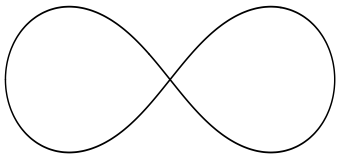


Figure 2.19: The lemniscate of Bernoulli.

2.5 Limits and Continuity; Uniform Continuity

The definition of continuity utilizes open sets (or neighborhoods) to capture the notion of "closeness." Another way to quantify this notion of "closeness" is through the limit of a sequence.

Definition 2.18. Given any set E, a *sequence* is any function $x: \mathbb{N} \to E$, usually denoted by $(x_n)_{n\in\mathbb{N}}$, or $(x_n)_{n\geq 0}$, or even by (x_n).

Definition 2.19. Given a topological space $(E, \mathcal{O})$, we say that *a sequence* $(x_n)_{n\in\mathbb{N}}$ *converges to some* $a \in E$ if for every open set U containing a, there is some $n_0 \geq 0$, such that, $x_n \in U$, for all $n \geq n_0$. We also say that *a is a limit of* $(x_n)_{n\in\mathbb{N}}$. See Figure 2.20.

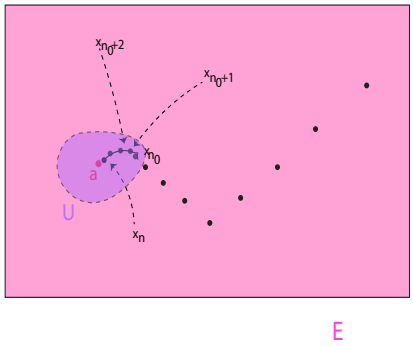


Figure 2.20: A schematic illustration of Definition 2.19.

When E is a metric space, Definition 2.19 is equivalent to the following proposition.

Proposition 2.16. *Let E be a metric space with metric d. A sequence $(x_n)_{n\in\mathbb{N}} \subset E$ converges to some $a \in E$ iff*

for every $\epsilon > 0$, there is some $n_0 \geq 0$, such that, $d(x_n, a) \leq \epsilon$, for all $n \geq n_0$.

When E is a normed vector space, Definition 2.19 is equivalent to the following proposition.

Proposition 2.17. *Let E be a normed vector space with norm $\|\ \|$. A sequence $(x_n)_{n\in\mathbb{N}} \subset E$ converges to some $a \in E$ iff*

for every $\epsilon > 0$, there is some $n_0 \geq 0$, such that, $\|x_n - a\| \leq \epsilon$, for all $n \geq n_0$.

The following proposition shows the importance of the Hausdorff separation axiom.

Proposition 2.18. *Given a topological space $(E, \mathcal{O})$, if the Hausdorff separation axiom holds, then every sequence has at most one limit.*

Proof. Left as an exercise. □

It is worth noting that the notion of limit is topological, in the sense that a sequence converge to a limit b iff it converges to the same limit b in any equivalent metric (and similarly for equivalent norms).

If E is a metric space and if A is a subset of E, there is a convenient way of showing that a point $x \in E$ belongs to the closure $\overline{A}$ of A in terms of sequences.

Proposition 2.19. *Given any metric space (E, d), for any subset A of E and any point $x \in E$, we have $x \in \overline{A}$ iff there is a sequence (a_n) of points $a_n \in A$ converging to x.*

Proof. If the sequence (a_n) of points $a_n \in A$ converges to x, then for every open subset U of E containing x, there is some n_0 such that $a_n \in U$ for all $n \geq n_0$, so $U \cap A \neq \emptyset$, and Proposition 2.3 implies that $x \in \overline{A}$.

Conversely, assume that $x \in \overline{A}$. Then for every $n \geq 1$, consider the open ball $B_0(x, 1/n)$. By Proposition 2.3, we have $B_0(x, 1/n) \cap A \neq \emptyset$, so we can pick some $a_n \in B_0(x, 1/n) \cap A$. This way, we define a sequence (a_n) of points in A, and by construction $d(x, a_n) < 1/n$ for all $n \geq 1$, so the sequence (a_n) converges to x. □

Before stating continuity in terms of limits, we still need one more concept, that of limit for functions.

Definition 2.20. Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, let A be some nonempty subset of E, and let $f\colon A \to F$ be a function. For any $a \in \overline{A}$ and any $b \in F$, we say that $f(x)$ *approaches* b *as* x *approaches* a *with values in* A if for every open set $V \in \mathcal{O}_F$ containing b, there is some open set $U \in \mathcal{O}_E$ containing a, such that, $f(U \cap A) \subseteq V$. See Figure 2.21. This is denoted by

$$\lim_{x \to a, x \in A} f(x) = b.$$

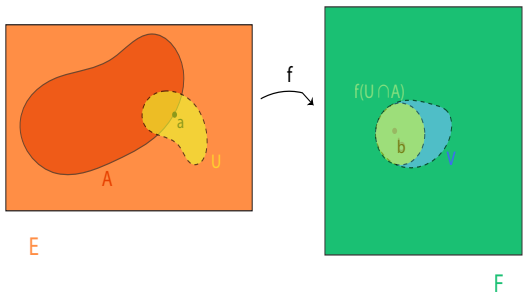


Figure 2.21: A schematic illustration of Definition 2.20.

Note that by Proposition 2.3, since $a \in \overline{A}$, for every open set U containing a, we have $U \cap A \neq \emptyset$, and the definition is nontrivial. Also, even if $a \in A$, the value $f(a)$ of f at a plays no role in this definition.

When E and F are metric spaces, Definition 2.20 can be restated as follows.

Proposition 2.20. *Let E and F be metric spaces with metrics d_1 and d_2. Let A be some nonempty subset of E, and let $f\colon A \to F$ be a function. For any $a \in \overline{A}$ and any $b \in F$, $f(x)$ approaches b as x approaches a with values in A iff*

for every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in A$,

$$\text{if } d_1(x, a) \leq \eta, \text{ then } d_2(f(x), b) \leq \epsilon.$$

When E and F are normed vector spaces, Definition 2.20 can be restated as follows.

Proposition 2.21. *Let E and F be normed vector spaces with norms $\| \|_1$ and $\| \|_2$. Let A be some nonempty subset of E, and let $f\colon A \to F$ be a function. For any $a \in \overline{A}$ and any $b \in F$, $f(x)$ approaches b as x approaches a with values in A iff*

for every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in A$,

$$\text{if } \|x - a\|_1 \leq \eta, \text{ then } \|f(x) - b\|_2 \leq \epsilon.$$

We have the following result relating continuity at a point and the previous notion.

Proposition 2.22. *Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be two topological spaces, and let $f\colon E \to F$ be a function. For any $a \in E$, the function f is continuous at a iff $f(x)$ approaches $f(a)$ when x approaches a (with values in E).*

Proof. Left as a trivial exercise. □

Another important proposition relating the notion of convergence of a sequence to continuity is stated without proof.

Proposition 2.23. *Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be two topological spaces, and let $f\colon E \to F$ be a function.*

(1) If f is continuous, then for every sequence $(x_n)_{n\in\mathbb{N}}$ in E, if (x_n) converges to a, then $(f(x_n))$ converges to $f(a)$.

(2) If E is a metric space, and $(f(x_n))$ converges to $f(a)$ whenever (x_n) converges to a, for every sequence $(x_n)_{n\in\mathbb{N}}$ in E, then f is continuous.

A special case of Definition 2.20 will be used when E and F are (nontrivial) normed vector spaces with norms $\| \|_1$ and $\| \|_2$. Let U be any nonempty open subset of E. We showed earlier that E has no isolated points and that every set $\{v\}$ is closed, for every $v \in E$. Since E is nontrivial, for every $v \in U$, there is a nontrivial open ball contained in U (an open ball not reduced to its center). Then for every $v \in U$, $A = U - \{v\}$ is open and nonempty, and clearly, $v \in \overline{A}$. For any $v \in U$, if $f(x)$ approaches b when x approaches v with values in $A = U - \{v\}$, we say that $f(x)$ *approaches b when x approaches v with values $\neq v$ in U*. This is denoted by

$$\lim_{x \to v, x \in U, x \neq v} f(x) = b.$$

Remark: Variations of the above case show up in the following case: $E = \mathbb{R}$, and F is some arbitrary topological space. Let A be some nonempty subset of $\mathbb{R}$, and let $f\colon A \to F$ be some function. For any $a \in A$, we say that f *is continuous on the right at a* if

$$\lim_{x \to a, x \in A \cap [a, +\infty)} f(x) = f(a).$$

We can define *continuity on the left* at a in a similar fashion, namely

$$\lim_{x \to a, x \in A \cap (-\infty, a]} f(x) = f(a).$$

For example, the function $f : \mathbb{R} \to \mathbb{R}$

$$\begin{cases} f(x) = x & \text{if } x < 1 \\ f(x) = 2 & \text{if } x \geq 1, \end{cases}$$

is continuous on the right at 1, but not continuous on the left at 1. See Figure 2.22.

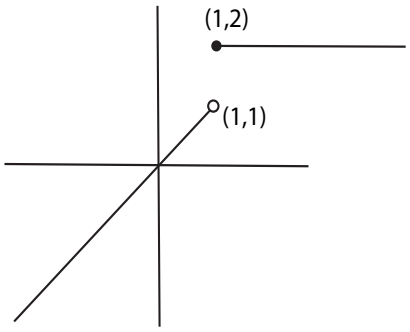


Figure 2.22: The graph of the piecewise function $f(x) = x$ when $x < 1$ and $f(x) = 2$ when $x \geq 1$.

Let us consider another variation. Let A be some nonempty subset of $\mathbb{R}$, and let $f : A \to F$ be some function. For any $a \in A$, we say that f *has a discontinuity of the first kind at* a if

$$\lim_{x \to a, x \in A \cap (-\infty, a)} f(x) = f(a_-)$$

and

$$\lim_{x \to a, x \in A \cap (a, +\infty)} f(x) = f(a_+)$$

both exist, and either $f(a_-) \neq f(a)$, or $f(a_+) \neq f(a)$. For example, the function $f : \mathbb{R} \to \mathbb{R}$

$$\begin{cases} f(x) = x & \text{if } x < 1 \\ f(x) = 2 & \text{if } x \geq 1, \end{cases}$$

has a discontinuity of the first kind at 1; both directional limits exits, namely $\lim_{x \to a, x \in A \cap (-\infty, a)} f(x) = 1$ and $\lim_{x \to a, x \in A \cap (a, +\infty)} f(x) = 2$, but $f(1_-) \neq f(1) = 2$. See Figure 2.22.

Note that it is possible that $f(a_-) = f(a_+)$, but f is still discontinuous at a if this common value differs from $f(a)$. Functions defined on a nonempty subset of $\mathbb{R}$, and that are continuous, except for some points of discontinuity of the first kind, play an important role in analysis.

In a metric space there is another important notion of continuity, namely uniform continuity.

Definition 2.21. Given two metric spaces, (E, d_E) and (F, d_F), a function, $f\colon E \to F$, is *uniformly continuous* if for every $\epsilon > 0$, there is some $\eta > 0$, such that for all $a, b \in E$,

$$\text{if} \quad d_E(a, b) \leq \eta \quad \text{then} \quad d_F(f(a), f(b)) \leq \epsilon.$$

See Figures 2.23 and 2.24.

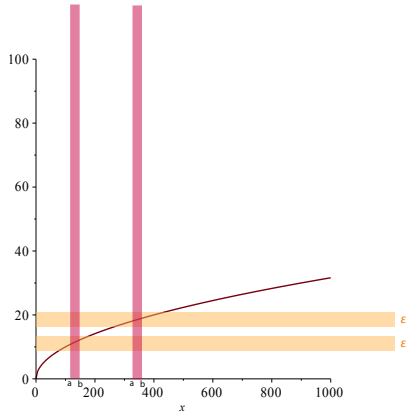


Figure 2.23: The real valued function $f(x) = \sqrt{x}$ is uniformly continuous over $(0, \infty)$. Fix ϵ. If the x values lie within the rose colored η strip, the y values always lie within the peach ϵ strip.

As we saw earlier, the metric on a metric space is uniformly continuous, and the norm on a normed metric space is uniformly continuous.

Before considering differentials, we need to look at the continuity of linear maps.

2.6 Continuous Linear and Multilinear Maps

If E and F are normed vector spaces, we first characterize when a linear map $f\colon E \to F$ is continuous.

Proposition 2.24. *Given two normed vector spaces E and F, for any linear map $f\colon E \to F$, the following conditions are equivalent:*

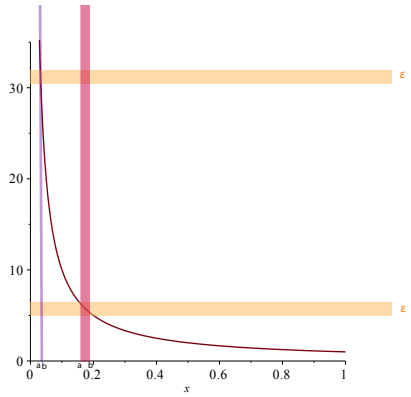


Figure 2.24: The real valued function $f(x) = 1/x$ is not uniformly continuous over $(0, \infty)$. Fix ϵ. In order for the y values to lie within the peach epsilon strip, the widths of the eta strips decrease as $x \to 0$.

(1) The function f is continuous at 0.

(2) There is a constant $k \geq 0$ such that,

$$\|f(u)\| \leq k, \text{ for every } u \in E \text{ such that } \|u\| \leq 1.$$

(3) There is a constant $k \geq 0$ such that,

$$\|f(u)\| \leq k\|u\|, \text{ for every } u \in E.$$

(4) The function f is continuous at every point of E.

Proof. Assume (1). Then for every $\epsilon > 0$, there is some $\eta > 0$ such that, for every $u \in E$, if $\|u\| \leq \eta$, then $\|f(u)\| \leq \epsilon$. Pick $\epsilon = 1$, so that there is some $\eta > 0$ such that, if $\|u\| \leq \eta$, then $\|f(u)\| \leq 1$. If $\|u\| \leq 1$, then $\|\eta u\| \leq \eta\|u\| \leq \eta$, and so, $\|f(\eta u)\| \leq 1$, that is, $\eta\|f(u)\| \leq 1$, which implies $\|f(u)\| \leq \eta^{-1}$. Thus Condition (2) holds with $k = \eta^{-1}$.

Assume that (2) holds. If $u = 0$, then by linearity, $f(0) = 0$, and thus $\|f(0)\| \leq k\|0\|$ holds trivially for all $k \geq 0$. If $u \neq 0$, then $\|u\| > 0$, and since

$$\left\| \frac{u}{\|u\|} \right\| = 1,$$

we have

$$\left\| f\left(\frac{u}{\|u\|} \right) \right\| \leq k,$$

which implies that

$$\|f(u)\| \leq k\|u\|.$$

Thus Condition (3) holds.

If (3) holds, then for all $u, v \in E$, we have

$$\|f(v) - f(u)\| = \|f(v-u)\| \leq k\|v-u\|.$$

If $k = 0$, then f is the zero function, and continuity is obvious. Otherwise, if $k > 0$, for every $\epsilon > 0$, if $\|v-u\| \leq \frac{\epsilon}{k}$, then $\|f(v-u)\| \leq \epsilon$, which shows continuity at every $u \in E$. Finally it is obvious that (4) implies (1). □

Among other things, Proposition 2.24 shows that a linear map is continuous iff the image of the unit (closed) ball is bounded. Since a continuous linear map satisfies the condition $\|f(u)\| \leq k\|u\|$ (for some $k \geq 0$), it is also uniformly continuous.

Definition 2.22. If E and F are normed vector spaces, the set of all continuous linear maps $f\colon E \to F$ is denoted by $\mathcal{L}(E; F)$.

Using Proposition 2.24, we can define a norm on $\mathcal{L}(E; F)$ which makes it into a normed vector space. This definition has already been given in Chapter 8 (Vol. I) (Definition 8.7 (Vol. I)) but for the reader's convenience, we repeat it here.

Definition 2.23. Given two normed vector spaces E and F, for every continuous linear map $f\colon E \to F$, we define the *norm* $\|f\|$ *of* f as

$$\begin{aligned}\|f\| &= \inf\{k \geq 0 \mid \|f(x)\| \leq k\|x\|, \text{ for all } x \in E\} \\ &= \sup\{\|f(x)\| \mid \|x\| \leq 1\} \\ &= \sup\{\|f(x)\| \mid \|x\| = 1\}.\end{aligned}$$

From Definition 2.23, for every continuous linear map $f \in \mathcal{L}(E; F)$, we have

$$\|f(x)\| \leq \|f\|\|x\|,$$

for every $x \in E$. It is easy to verify that $\mathcal{L}(E; F)$ is a normed vector space under the norm of Definition 2.23. Furthermore, if E, F, G are normed vector spaces, and $f\colon E \to F$ and $g\colon F \to G$ are continuous linear maps, we have

$$\|g \circ f\| \leq \|g\|\|f\|.$$

We can now show that when $E = \mathbb{R}^n$ or $E = \mathbb{C}^n$, with any of the norms $\| \|_1$, $\| \|_2$, or $\| \|_\infty$, then every linear map $f\colon E \to F$ is continuous.

Proposition 2.25. *If $E = \mathbb{R}^n$ or $E = \mathbb{C}^n$, with any of the norms $\| \|_1$, $\| \|_2$, or $\| \|_\infty$, and F is any normed vector space, then every linear map $f\colon E \to F$ is continuous.*

Proof. Let $(e_1, \ldots, e_n)$ be the standard basis of $\mathbb{R}^n$ (a similar proof applies to $\mathbb{C}^n$). In view of Proposition 8.3 (Vol. I), it is enough to prove the proposition for the norm

$$\|x\|_\infty = \max\{|x_i| \mid 1 \leq i \leq n\}.$$

We have

$$\|f(v) - f(u)\| = \|f(v-u)\| = \left\| f\Big(\sum_{1\leq i\leq n} (v_i - u_i)e_i\Big) \right\| = \left\| \sum_{1\leq i\leq n} (v_i - u_i) f(e_i) \right\|,$$

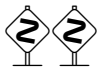
and so,

$$\|f(v) - f(u)\| \leq \Big(\sum_{1\leq i\leq n} \|f(e_i)\|\Big) \max_{1\leq i\leq n} |v_i - u_i| = \Big(\sum_{1\leq i\leq n} \|f(e_i)\|\Big) \|v-u\|_\infty.$$

By the argument used in Proposition 2.24 to prove that (3) implies (4), f is continuous. □

Actually, we proved in Theorem 8.5 (Vol. I) that if E is a vector space of finite dimension, then any two norms are equivalent, so that they define the same topology. This fact together with Proposition 2.25 prove the following.

Theorem 2.26. *If E is a vector space of finite dimension (over $\mathbb{R}$ or $\mathbb{C}$), then all norms are equivalent (define the same topology). Furthermore, for any normed vector space F, every linear map $f\colon E \to F$ is continuous.*



If E is a normed vector space of infinite dimension, a linear map $f\colon E \to F$ may not be continuous.

As an example, let E be the infinite vector space of all polynomials over $\mathbb{R}$. Let

$$\|P(X)\| = \sup_{0\leq x\leq 1} |P(x)|.$$

We leave as an exercise to show that this is indeed a norm. Let $F = \mathbb{R}$, and let $f\colon E \to F$ be the map defined such that, $f(P(X)) = P(3)$. It is clear that f is linear. Consider the sequence of polynomials

$$P_n(X) = \left(\frac{X}{2}\right)^n.$$

It is clear that $\|P_n\| = \left(\frac{1}{2}\right)^n$, and thus, the sequence P_n has the null polynomial as a limit. However, we have

$$f(P_n(X)) = P_n(3) = \left(\frac{3}{2}\right)^n,$$

and the sequence $f(P_n(X))$ diverges to $+\infty$. Consequently, in view of Proposition 2.23 (1), f is not continuous.

We now consider the continuity of multilinear maps. We treat explicitly bilinear maps, the general case being a straightforward extension.

Proposition 2.27. *Given normed vector spaces E, F and G, for any bilinear map $f\colon E \times F \to G$, the following conditions are equivalent:*

(1) The function f is continuous at $\langle 0, 0\rangle$.

(2) There is a constant $k \geq 0$ such that,

$$\|f(u, v)\| \leq k, \text{ for all } u \in E, v \in F \text{ such that } \|u\|, \|v\| \leq 1.$$

(3) There is a constant $k \geq 0$ such that,

$$\|f(u, v)\| \leq k\|u\|\|v\|, \text{ for all } u \in E, v \in F.$$

(4) The function f is continuous at every point of $E \times F$.

Proof. It is similar to that of Proposition 2.24, with a small subtlety in proving that (3) implies (4), namely that two different η's that are not independent are needed. □

In contrast to continuous linear maps, which must be uniformly continuous, nonzero continuous bilinear maps are **not** uniformly continuous. Let $f\colon E \times F \to G$ be a continuous bilinear map such that $f(a, b) \neq 0$ for some $a \in E$ and some $b \in F$. Consider the sequences (u_n) and (v_n) (with $n \geq 1$) given by

$$\begin{aligned} u_n &= (x_n, y_n) = (na, nb) \\ v_n &= (x'_n, y'_n) = \left(\left(n + \frac{1}{n}\right) a, \left(n + \frac{1}{n}\right) b\right). \end{aligned}$$

Obviously

$$\|v_n - u_n\| \leq \frac{1}{n}(\|a\| + \|b\|),$$

so $\lim_{n\mapsto\infty} \|v_n - u_n\| = 0$. On the other hand

$$f(x'_n, y'_n) - f(x_n, y_n) = \left(2 + \frac{1}{n^2}\right) f(a, b),$$

and thus $\lim_{n\mapsto\infty} \|f(x'_n, y'_n) - f(x_n, y_n)\| = 2\,\|f(a, b)\| \neq 0$, which shows that f is not uniformly continuous, because if this was the case, this limit would be zero.

Definition 2.24. If E, F, and G are normed vector spaces, we denote the set of all continuous bilinear maps $f\colon E \times F \to G$ by $\mathcal{L}_2(E, F; G)$.

Using Proposition 2.27, we can define a norm on $\mathcal{L}_2(E, F; G)$ which makes it into a normed vector space.

Definition 2.25. Given normed vector spaces E, F, and G, for every continuous bilinear map $f\colon E \times F \to G$, we define the *norm* $\|f\|$ *of* f as

$$\begin{aligned}\|f\| &= \inf\{k \geq 0 \mid \|f(x,y)\| \leq k\|x\|\|y\|, \text{ for all } x \in E, y \in F\}\\ &= \sup\{\|f(x,y)\| \mid \|x\|, \|y\| \leq 1\}\\ &= \sup\{\|f(x,y)\| \mid \|x\| = \|y\| = 1\}.\end{aligned}$$

From Definition 2.25, for every continuous bilinear map $f \in \mathcal{L}_2(E,F;G)$, we have

$$\|f(x,y)\| \leq \|f\|\|x\|\|y\|,$$

for all $x \in E, y \in F$. It is easy to verify that $\mathcal{L}_2(E,F;G)$ is a normed vector space under the norm of Definition 2.25.

Given a bilinear map $f\colon E \times F \to G$, for every $u \in E$, we obtain a linear map denoted $fu\colon F \to G$, defined such that, $fu(v) = f(u,v)$. Furthermore, since

$$\|f(x,y)\| \leq \|f\|\|x\|\|y\|,$$

it is clear that fu is continuous. We can then consider the map $\varphi\colon E \to \mathcal{L}(F;G)$, defined such that, $\varphi(u) = fu$, for any $u \in E$, or equivalently, such that,

$$\varphi(u)(v) = f(u,v).$$

Actually, it is easy to show that φ is linear and continuous, and that $\|\varphi\| = \|f\|$. Thus, $f \mapsto \varphi$ defines a map from $\mathcal{L}_2(E,F;G)$ to $\mathcal{L}(E;\mathcal{L}(F;G))$. We can also go back from $\mathcal{L}(E;\mathcal{L}(F;G))$ to $\mathcal{L}_2(E,F;G)$. We summarize all this in the following proposition.

Proposition 2.28. *Let E, F, G be three normed vector spaces. The map $f \mapsto \varphi$, from $\mathcal{L}_2(E,F;G)$ to $\mathcal{L}(E;\mathcal{L}(F;G))$, defined such that, for every $f \in \mathcal{L}_2(E,F;G)$,*

$$\varphi(u)(v) = f(u,v),$$

is an isomorphism of vector spaces, and furthermore, $\|\varphi\| = \|f\|$.

As a corollary of Proposition 2.28, we get the following proposition which will be useful when we define second-order derivatives.

Proposition 2.29. *Let E and F be normed vector spaces. The map app from $\mathcal{L}(E;F) \times E$ to F, defined such that, for every $f \in \mathcal{L}(E;F)$, for every $u \in E$,*

$$app(f,u) = f(u),$$

is a continuous bilinear map.

Remark: If E and F are nontrivial, it can be shown that $\|\text{app}\| = 1$. It can also be shown that composition

$$\circ\colon \mathcal{L}(E;F) \times \mathcal{L}(F;G) \to \mathcal{L}(E;G),$$

is bilinear and continuous.

The above propositions and definition generalize to arbitrary n-multilinear maps, with $n \geq 2$. Proposition 2.27 extends in the obvious way to any n-multilinear map $f\colon E_1 \times \cdots \times E_n \to F$, but condition (3) becomes:

There is a constant $k \geq 0$ such that,

$$\|f(u_1, \ldots, u_n)\| \leq k\|u_1\| \cdots \|u_n\|, \text{ for all } u_1 \in E_1, \ldots, u_n \in E_n.$$

Definition 2.25 also extends easily to

$$\begin{aligned}\|f\| &= \inf\{k \geq 0 \mid \|f(x_1, \ldots, x_n)\| \leq k\|x_1\| \cdots \|x_n\|, \text{ for all } x_i \in E_i, 1 \leq i \leq n\} \\ &= \sup\{\|f(x_1, \ldots, x_n)\| \mid \|x_1\|, \ldots, \|x_n\| \leq 1\} \\ &= \sup\{\|f(x_1, \ldots, x_n)\| \mid \|x_1\| = \cdots = \|x_n\| = 1\}.\end{aligned}$$

Proposition 2.28 is also easily extended, and we get an isomorphism between continuous n-multilinear maps in $\mathcal{L}_n(E_1, \ldots, E_n; F)$, and continuous linear maps in

$$\mathcal{L}(E_1; \mathcal{L}(E_2; \ldots; \mathcal{L}(E_n; F))).$$

An obvious extension of Proposition 2.29 also holds.

Complete metric spaces and complete normed vector spaces are important tools in analysis and optimization theory, so we include some sections covering the basics.

2.7 Complete Metric Spaces and Banach Spaces

Definition 2.26. Given a metric space, (E, d), a sequence, $(x_n)_{n\in\mathbb{N}}$, in E is a *Cauchy sequence* if the following condition holds: for every $\epsilon > 0$, there is some $p \geq 0$, such that for all $m, n \geq p$, then $d(x_m, x_n) \leq \epsilon$.

If every Cauchy sequence in (E, d) converges we say that (E, d) is a *complete metric space*. A normed vector space $(E, \| \|)$ over $\mathbb{R}$ (or $\mathbb{C}$) which is a complete metric space for the distance $d(u, v) = \|v - u\|$, is called a *Banach space*.

The standard example of a complete metric space is the set $\mathbb{R}$ of real numbers. As a matter of fact, the set $\mathbb{R}$ can be defined as the "completion" of the set $\mathbb{Q}$ of rationals. The spaces $\mathbb{R}^n$ and $\mathbb{C}^n$ under their standard topology are complete metric spaces.

It can be shown that every normed vector space of finite dimension is a Banach space (is complete). It can also be shown that if E and F are normed vector spaces, and F is a Banach space, then $\mathcal{L}(E;F)$ is a Banach space. If E, F and G are normed vector spaces, and G is a Banach space, then $\mathcal{L}_2(E,F;G)$ is a Banach space.

An arbitrary metric space (E,d) is not necessarily complete, but there is a construction of a metric space $(\widehat{E},\widehat{d})$ such that $\widehat{E}$ is complete, and there is a continuous (injective) distance-preserving map $\varphi\colon E \to \widehat{E}$ such that $\varphi(E)$ is dense in $\widehat{E}$. This is a generalization of the construction of the set $\mathbb{R}$ of real numbers from the set $\mathbb{Q}$ of rational numbers in terms of Cauchy sequences. This construction can be immediately adapted to a normed vector space $(E, \| \|)$ to embed $(E, \| \|)$ into a complete normed vector space $(\widehat{E}, \| \|_{\widehat{E}})$ (a Banach space). This construction is used heavily in integration theory where E is a set of functions.

2.8 Completion of a Metric Space

In order to prove a kind of uniqueness result for the completion $(\widehat{E},\widehat{d})$ of a metric space (E,d), we need the following result about extending a uniformly continuous function.

Recall that E_0 is dense in E iff $\overline{E_0} = E$. Since E is a metric space, by Proposition 2.19, this means that for every $x \in E$, there is some sequence (x_n) converging to x, with $x_n \in E_0$.

Theorem 2.30. *Let E and F be two metric spaces, let E_0 be a dense subspace of E, and let $f_0\colon E_0 \to F$ be a continuous function. If f_0 is uniformly continuous and if F is complete, then there is a unique uniformly continuous function $f\colon E \to F$ extending f_0.*

Proof. We follow Schwartz's proof; see Schwartz [68] (Chapter XI, Section 3, Theorem 1).

Step 1. We begin by constructing a function $f\colon E \to F$ extending f_0. Since E_0 is dense in E, for every $x \in E$, there is some sequence (x_n) converging to x, with $x_n \in E_0$. Then the sequence (x_n) is a Cauchy sequence in E. We claim that $(f_0(x_n))$ is a Cauchy sequence in F.

Proof of the claim. For every $\epsilon > 0$, since f_0 is uniformly continuous, there is some $\eta > 0$ such that for all $(y,z) \in E_0$, if $d(y,z) \leq \eta$, then $d(f_0(y), f_0(z)) \leq \epsilon$. Since (x_n) is a Cauchy sequence with $x_n \in E_0$, there is some integer $p > 0$ such that if $m, n \geq p$, then $d(x_m, x_n) \leq \eta$, thus $d(f_0(x_m), f_0(x_n)) \leq \epsilon$, which proves that $(f_0(x_n))$ is a Cauchy sequence in F. □

Since F is complete and $(f_0(x_n))$ is a Cauchy sequence in F, the sequence $(f_0(x_n))$ converges to some element of F; denote this element by $f(x)$.

Step 2. Let us now show that $f(x)$ does not depend on the sequence (x_n) converging to x. Suppose that (x'_n) and (x''_n) are two sequences of elements in E_0 converging to x. Then the mixed sequence

$$x'_0, x''_0, x'_1, x''_1, \ldots, x'_n, x''_n, \ldots,$$

also converges to x. It follows that the sequence

$$f_0(x_0'), f_0(x_0''), f_0(x_1'), f_0(x_1''), \ldots, f_0(x_n'), f_0(x_n''), \ldots,$$

is a Cauchy sequence in F, and since F is complete, it converges to some element of F, which implies that the sequences $(f_0(x_n'))$ and $(f_0(x_n''))$ converge to the same limit.

As a summary, we have defined a function $f\colon E \to F$ by

$$f(x) = \lim_{n\mapsto\infty} f_0(x_n),$$

for any sequence (x_n) converging to x, with $x_n \in E_0$. See Figure 2.25.

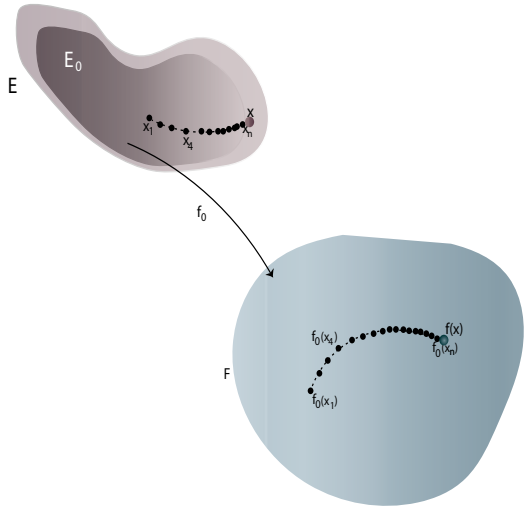


Figure 2.25: A schematic illustration of the construction of $f\colon E \to F$ where $f(x) = \lim_{n\mapsto\infty} f_0(x_n)$ for any sequence (x_n) converging to x, with $x_n \in E_0$.

Step 3. The function f extends f_0. Since every element $x \in E_0$ is the limit of the constant sequence (x_n) with $x_n = x$ for all $n \geq 0$, by definition $f(x)$ is the limit of the sequence $(f_0(x_n))$, which is the constant sequence with value $f_0(x)$, so $f(x) = f_0(x)$; that is, f extends f_0.

Step 4. We now prove that f is uniformly continuous. Since f_0 is uniformly continuous, for every $\epsilon > 0$, there is some $\eta > 0$ such that if $a, b \in E_0$ and $d(a, b) \leq \eta$, then $d(f_0(a), f_0(b)) \leq \epsilon$. Consider any two points $x, y \in E$ such that $d(x, y) \leq \eta/2$. We claim that $d(f(x), f(y)) \leq \epsilon$, which shows that f is uniformly continuous.

Let (x_n) be a sequence of points in E_0 converging to x, and let (y_n) be a sequence of points in E_0 converging to y. By the triangle inequality,

$$d(x_n, y_n) \leq d(x_n, x) + d(x, y) + d(y, y_n) = d(x, y) + d(x_n, x) + d(y_n, y),$$

and since (x_n) converges to x and (y_n) converges to y, there is some integer $p > 0$ such that for all $n \geq p$, we have $d(x_n, x) \leq \eta/4$ and $d(y_n, y) \leq \eta/4$, and thus

$$d(x_n, y_n) \leq d(x, y) + \frac{\eta}{2}.$$

Since we assumed that $d(x, y) \leq \eta/2$, we get $d(x_n, y_n) \leq \eta$ for all $n \geq p$, and by uniform continuity of f_0, we get

$$d(f_0(x_n), f_0(y_n)) \leq \epsilon$$

for all $n \geq p$. Since the distance function on F is also continuous, and since $(f_0(x_n))$ converges to $f(x)$ and $(f_0(y_n))$ converges to $f(y)$, we deduce that the sequence $(d(f_0(x_n), f_0(y_n)))$ converges to $d(f(x), f(y))$. This implies that $d(f(x), f(y)) \leq \epsilon$, as desired.

Step 5. It remains to prove that f is unique. Since E_0 is dense in E, for every $x \in E$, there is some sequence (x_n) converging to x, with $x_n \in E_0$. Since f extends f_0 and since f is continuous, we get

$$f(x) = \lim_{n \mapsto \infty} f_0(x_n),$$

which only depends on f_0 and x and shows that f is unique. □

Remark: It can be shown that the theorem no longer holds if we either omit the hypothesis that F is complete or omit that f_0 is uniformly continuous.

For example, if $E_0 \neq E$ and if we let $F = E_0$ and f_0 be the identity function, it is easy to see that f_0 cannot be extended to a continuous function from E to E_0 (for any $x \in E - E_0$, any continuous extension f of f_0 would satisfy $f(x) = x$, which is absurd since $x \notin E_0$).

If f_0 is continuous but not uniformly continuous, a counter-example can be given by using $E = \overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ made into a metric space, $E_0 = \mathbb{R}$, $F = \mathbb{R}$, and f_0 the identity function; for details, see Schwartz [68] (Chapter XI, Section 3, page 134).

Definition 2.27. If (E, d_E) and (F, d_F) are two metric spaces, then a function $f\colon E \to F$ is *distance-preserving*, or an *isometry*, if

$$d_F(f(x), f(y)) = d_E(x, y), \quad \text{for all for all } x, y \in E.$$

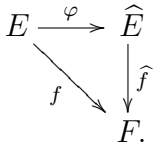
Observe that an isometry must be injective, because if $f(x) = f(y)$, then $d_F(f(x), f(y)) = 0$, and since $d_F(f(x), f(y)) = d_E(x, y)$, we get $d_E(x, y) = 0$, but $d_E(x, y) = 0$ implies that $x = y$. Also, an isometry is uniformly continuous (since we can pick $\eta = \epsilon$ to satisfy the condition of uniform continuity). However, an isometry is not necessarily surjective.

We now give a construction of the completion of a metric space. This construction is just a generalization of the classical construction of $\mathbb{R}$ from $\mathbb{Q}$ using Cauchy sequences.

Theorem 2.31. *Let (E, d) be any metric space. There is a complete metric space $(\widehat{E}, \widehat{d})$ called a completion of (E, d), and a distance-preserving (uniformly continuous) map $\varphi\colon E \to \widehat{E}$ such that $\varphi(E)$ is dense in $\widehat{E}$, and the following extension property holds: for every complete metric space F and for every uniformly continuous function $f\colon E \to F$, there is a unique uniformly continuous function $\widehat{f}\colon \widehat{E} \to F$ such that*

$$f = \widehat{f} \circ \varphi,$$

as illustrated in the following diagram.



As a consequence, for any two completions $(\widehat{E}_1, \widehat{d}_1)$ and $(\widehat{E}_2, \widehat{d}_2)$ of (E, d), there is a unique bijective isometry between $(\widehat{E}_1, \widehat{d}_1)$ and $(\widehat{E}_2, \widehat{d}_2)$.

Proof. Consider the set $\mathcal{E}$ of all Cauchy sequences (x_n) in E, and define the relation $\sim$ on $\mathcal{E}$ as follows:

$$(x_n) \sim (y_n) \quad \text{iff} \quad \lim_{n \mapsto \infty} d(x_n, y_n) = 0.$$

It is easy to check that $\sim$ is an equivalence relation on $\mathcal{E}$, and let $\widehat{E} = \mathcal{E}/\sim$ be the quotient set, that is, the set of equivalence classes modulo $\sim$. Our goal is to show that we can endow $\widehat{E}$ with a distance that makes it into a complete metric space satisfying the conditions of the theorem. We proceed in several steps.

Step 1. First let us construct the function $\varphi\colon E \to \widehat{E}$. For every $a \in E$, we have the constant sequence (a_n) such that $a_n = a$ for all $n \geq 0$, which is obviously a Cauchy sequence. Let $\varphi(a) \in \widehat{E}$ be the equivalence class $[(a_n)]$ of the constant sequence (a_n) with $a_n = a$ for all n. By definition of $\sim$, the equivalence class $\varphi(a)$ is also the equivalence class of all sequences converging to a. The map $a \mapsto \varphi(a)$ is injective because a metric space is Hausdorff, so if $a \neq b$, then a sequence converging to a does not converge to b. After having defined a distance on $\widehat{E}$, we will check that φ is an isometry.

Step 2. Let us now define a distance on $\widehat{E}$. Let $\alpha = [(a_n)]$ and $\beta = [(b_n)]$ be two equivalence classes of Cauchy sequences in E. The triangle inequality implies that

$$d(a_m, b_m) \leq d(a_m, a_n) + d(a_n, b_n) + d(b_n, b_m) = d(a_n, b_n) + d(a_m, a_n) + d(b_m, b_n)$$

and

$$d(a_n, b_n) \leq d(a_n, a_m) + d(a_m, b_m) + d(b_m, b_n) = d(a_m, b_m) + d(a_m, a_n) + d(b_m, b_n),$$

which implies that

$$|d(a_m, b_m) - d(a_n, b_n)| \leq d(a_m, a_n) + d(b_m, b_n).$$

Since (a_n) and (b_n) are Cauchy sequences, the above inequality shows that $(d(a_n, b_n))$ is a Cauchy sequence of nonnegative reals. Since $\mathbb{R}$ is complete, the sequence $(d(a_n, b_n))$ has a limit, which we denote by $\widehat{d}(\alpha, \beta)$; that is, we set

$$\widehat{d}(\alpha, \beta) = \lim_{n \mapsto \infty} d(a_n, b_n), \quad \alpha = [(a_n)],\ \beta = [(b_n)].$$

See Figure 2.26.

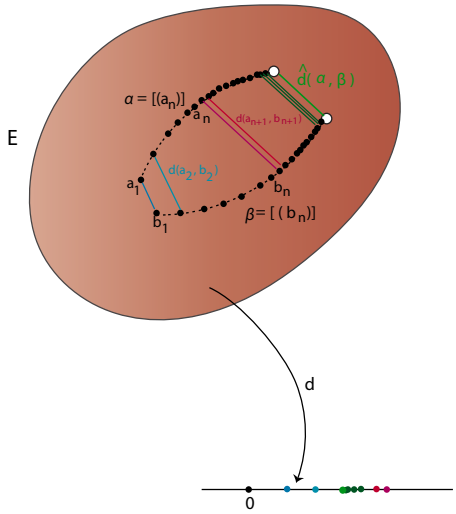


Figure 2.26: A schematic illustration of $\widehat{d}(\alpha, \beta)$ from the Cauchy sequence $(d(a_n, b_n))$.

Step 3. Let us check that $\widehat{d}(\alpha, \beta)$ does not depend on the Cauchy sequences (a_n) and (b_n) chosen in the equivalence classes α and β.

If $(a_n) \sim (a'_n)$ and $(b_n) \sim (b'_n)$, then $\lim_{n \mapsto \infty} d(a_n, a'_n) = 0$ and $\lim_{n \mapsto \infty} d(b_n, b'_n) = 0$, and since

$$d(a'_n, b'_n) \leq d(a'_n, a_n) + d(a_n, b_n) + d(b_n, b'_n) = d(a_n, b_n) + d(a_n, a'_n) + d(b_n, b'_n),$$

and

$$d(a_n, b_n) \leq d(a_n, a'_n) + d(a'_n, b'_n) + d(b'_n, b_n) = d(a'_n, b'_n) + d(a_n, a'_n) + d(b_n, b'_n),$$

we have

$$|d(a_n, b_n) - d(a'_n, b'_n)| \leq d(a_n, a'_n) + d(b_n, b'_n),$$

so we have $\lim_{n \mapsto \infty} d(a'_n, b'_n) = \lim_{n \mapsto \infty} d(a_n, b_n) = \widehat{d}(\alpha, \beta)$. Therefore, $\widehat{d}(\alpha, \beta)$ is indeed well defined.

Step 4. Let us check that φ is indeed an isometry.

Given any two elements $\varphi(a)$ and $\varphi(b)$ in $\widehat{E}$, since they are the equivalence classes of the constant sequences (a_n) and (b_n) such that $a_n = a$ and $b_n = b$ for all n, the constant sequence $(d(a_n, b_n))$ with $d(a_n, b_n) = d(a, b)$ for all n converges to $d(a, b)$, so by definition $\widehat{d}(\varphi(a), \varphi(b)) = \lim_{n\mapsto\infty} d(a_n, b_n) = d(a, b)$, which shows that φ is an isometry.

Step 5. Let us verify that $\widehat{d}$ is a metric on $\widehat{E}$. By definition it is obvious that $\widehat{d}(\alpha, \beta) = \widehat{d}(\beta, \alpha)$. If α and β are two distinct equivalence classes, then for any Cauchy sequence (a_n) in the equivalence class α and for any Cauchy sequence (b_n) in the equivalence class β, the sequences (a_n) and (b_n) are inequivalent, which means that $\lim_{n\mapsto\infty} d(a_n, b_n) \neq 0$, that is, $\widehat{d}(\alpha, \beta) \neq 0$. Obviously, $\widehat{d}(\alpha, \alpha) = 0$.

For any equivalence classes $\alpha = [(a_n)]$, $\beta = [(b_n)]$, and $\gamma = [(c_n)]$, we have the triangle inequality

$$d(a_n, c_n) \leq d(a_n, b_n) + d(b_n, c_n),$$

so by continuity of the distance function, by passing to the limit, we obtain

$$\widehat{d}(\alpha, \gamma) \leq \widehat{d}(\alpha, \beta) + \widehat{d}(\beta, \gamma),$$

which is the triangle inequality for $\widehat{d}$. Therefore, $\widehat{d}$ is a distance on $\widehat{E}$.

Step 6. Let us prove that $\varphi(E)$ is dense in $\widehat{E}$. For any $\alpha = [(a_n)]$, let (x_n) be the constant sequence such that $x_k = a_n$ for all $k \geq 0$, so that $\varphi(a_n) = [(x_n)]$. Then we have

$$\widehat{d}(\alpha, \varphi(a_n)) = \lim_{m\mapsto\infty} d(a_m, a_n) \leq \sup_{p,q\geq n} d(a_p, a_q).$$

Since (a_n) is a Cauchy sequence, $\sup_{p,q\geq n} d(a_p, a_q)$ tends to 0 as n goes to infinity, so

$$\lim_{n\mapsto\infty} d(\alpha, \varphi(a_n)) = 0,$$

which means that the sequence $(\varphi(a_n))$ converge to α, and $\varphi(E)$ is indeed dense in $\widehat{E}$.

Step 7. Finally let us prove that the metric space $\widehat{E}$ is complete.

Let (α_n) be a Cauchy sequence in $\widehat{E}$. Since $\varphi(E)$ is dense in $\widehat{E}$, for every $n > 0$, there some $a_n \in E$ such that

$$\widehat{d}(\alpha_n, \varphi(a_n)) \leq \frac{1}{n}.$$

Since

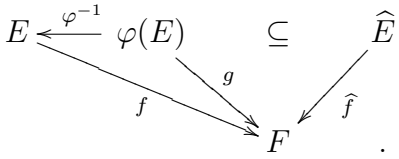
$$\widehat{d}(\varphi(a_m), \varphi(a_n)) \leq \widehat{d}(\varphi(a_m), \alpha_m) + \widehat{d}(\alpha_m, \alpha_n) + \widehat{d}(\alpha_n, \varphi(a_n)) \leq \widehat{d}(\alpha_m, \alpha_n) + \frac{1}{m} + \frac{1}{n},$$

and since (α_m) is a Cauchy sequence, so is $(\varphi(a_n))$, and as φ is an isometry, the sequence (a_n) is a Cauchy sequence in E. Let $\alpha \in \widehat{E}$ be the equivalence class of (a_n). Since

$$\widehat{d}(\alpha, \varphi(a_n)) = \lim_{m\mapsto\infty} d(a_m, a_n)$$

and (a_n) is a Cauchy sequence, we deduce that the sequence $(\varphi(a_n))$ converges to α, and since $d(\alpha_n, \varphi(a_n)) \leq 1/n$ for all $n > 0$, the sequence (α_n) also converges to α.

Step 8. Let us prove the extension property. Let F be any complete metric space and let $f\colon E \to F$ be any uniformly continuous function. The function $\varphi\colon E \to \widehat{E}$ is an isometry and a bijection between E and its image $\varphi(E)$, so its inverse $\varphi^{-1}\colon \varphi(E) \to E$ is also an isometry, and thus is uniformly continuous. If we let $g = f \circ \varphi^{-1}$, then $g\colon \varphi(E) \to F$ is a uniformly continuous function, and $\varphi(E)$ is dense in $\widehat{E}$, so by Theorem 2.30 there is a unique uniformly continuous function $\widehat{f}\colon \widehat{E} \to F$ extending $g = f \circ \varphi^{-1}$; see the diagram below:



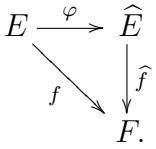
This means that

$$\widehat{f}|\varphi(E) = f \circ \varphi^{-1},$$

which implies that

$$(\widehat{f}|\varphi(E)) \circ \varphi = f,$$

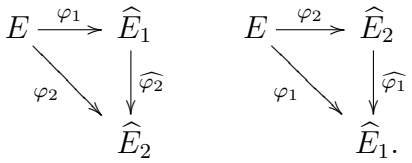
that is, $f = \widehat{f} \circ \varphi$, as illustrated in the diagram below:



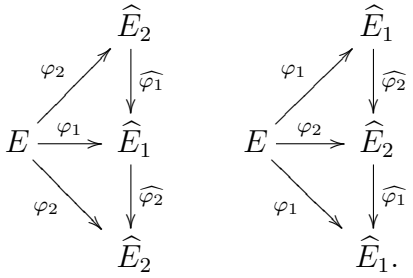
If $h\colon \widehat{E} \to F$ is any other uniformly continuous function such that $f = h \circ \varphi$, then $g = f \circ \varphi^{-1} = h|\varphi(E)$, so h is a uniformly continuous function extending g, and by Theorem 2.30, we have have $h = \widehat{f}$, so $\widehat{f}$ is indeed unique.

Step 9. Uniqueness of the completion $(\widehat{E}, \widehat{d})$ up to a bijective isometry.

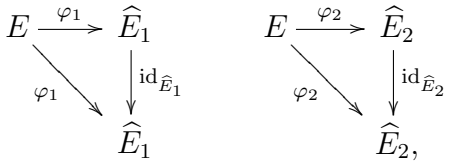
Let $(\widehat{E}_1, \widehat{d}_1)$ and $(\widehat{E}_2, \widehat{d}_2)$ be any two completions of (E, d). Then we have two uniformly continuous isometries $\varphi_1\colon E \to \widehat{E}_1$ and $\varphi_2\colon E \to \widehat{E}_2$, so by the unique extension property, there exist unique uniformly continuous maps $\widehat{\varphi_2}\colon \widehat{E}_1 \to \widehat{E}_2$ and $\widehat{\varphi_1}\colon \widehat{E}_2 \to \widehat{E}_1$ such that the following diagrams commute:



Consequently we have the following commutative diagrams:



However, $\mathrm{id}_{\widehat{E}_1}$ and $\mathrm{id}_{\widehat{E}_2}$ are uniformly continuous functions making the following diagrams commute



so by the uniqueness of extensions we must have

$$\widehat{\varphi_1} \circ \widehat{\varphi_2} = \mathrm{id}_{\widehat{E}_1} \quad \text{and} \quad \widehat{\varphi_2} \circ \widehat{\varphi_1} = \mathrm{id}_{\widehat{E}_2}.$$

This proves that $\widehat{\varphi_1}$ and $\widehat{\varphi_2}$ are mutual inverses. Now since $\varphi_2 = \widehat{\varphi_2} \circ \varphi_1$, we have

$$\widehat{\varphi_2}|\varphi_1(E) = \varphi_2 \circ \varphi_1^{-1},$$

and since φ_1^{-1} and φ_2 are isometries, so is $\widehat{\varphi_2}|\varphi_1(E)$. But we showed in Step 8 that $\widehat{\varphi_2}$ is the uniform continuous extension of $\widehat{\varphi_2}|\varphi_1(E)$ and $\varphi_1(E)$ is dense in $\widehat{E}_1$, so for any two elements $\alpha, \beta \in \widehat{E}_1$, if (a_n) and (b_n) are sequences in $\varphi_1(E)$ converging to α and β, we have

$$\widehat{d}_2((\widehat{\varphi_2}|\varphi_1(E))(a_n), ((\widehat{\varphi_2}|\varphi_1(E))(b_n)) = \widehat{d}_1(a_n, b_n),$$

and by passing to the limit we get

$$\widehat{d}_2(\widehat{\varphi_2}(\alpha), \widehat{\varphi_2}(\beta)) = \widehat{d}_1(\alpha, \beta),$$

which shows that $\widehat{\varphi_2}$ is an isometry (similarly, $\widehat{\varphi_1}$ is an isometry). □

Remarks:

1. Except for Step 8 and Step 9, the proof of Theorem 2.31 is the proof given in Schwartz [68] (Chapter XI, Section 4, Theorem 1), and Kolmogorov and Fomin [45] (Chapter 2, Section 7, Theorem 4).

2. The construction of $\widehat{E}$ relies on the completeness of $\mathbb{R}$, and so it cannot be used to construct $\mathbb{R}$ from $\mathbb{Q}$. However, this construction can be modified to yield a construction of $\mathbb{R}$ from $\mathbb{Q}$.

We show in Section 2.9 that Theorem 2.31 yields a construction of the completion of a normed vector space.

2.9 Completion of a Normed Vector Space

An easy corollary of Theorem 2.31 and Theorem 2.30 is that every normed vector space can be embedded in a complete normed vector space, that is, a Banach space.

Theorem 2.32. *If $(E, \| \|)$ is a normed vector space, then its completion $(\widehat{E}, \widehat{d})$ as a metric space (where E is given the metric $d(x, y) = \|x - y\|$) can be given a unique vector space structure extending the vector space structure on E, and a norm $\| \|_{\widehat{E}}$, so that $(\widehat{E}, \| \|_{\widehat{E}})$ is a Banach space, and the metric $\widehat{d}$ is associated with the norm $\| \|_{\widehat{E}}$. Furthermore, the isometry $\varphi \colon E \to \widehat{E}$ is a linear isometry.*

Proof. The addition operation $+ \colon E \times E \to E$ is uniformly continuous because

$$\|(u' + v') - (u'' + v'')\| \leq \|u' - u''\| + \|v' - v''\|.$$

It is not hard to show that $\widehat{E} \times \widehat{E}$ is a complete metric space and that $E \times E$ is dense in $\widehat{E} \times \widehat{E}$. Then by Theorem 2.30, the uniformly continuous function $+$ has a unique continuous extension $+ \colon \widehat{E} \times \widehat{E} \to \widehat{E}$.

The map $\cdot \colon \mathbb{R} \times E \to E$ is not uniformly continuous, but for any fixed $\lambda \in \mathbb{R}$, the map $L_\lambda \colon E \to E$ given by $L_\lambda(u) = \lambda \cdot u$ is uniformly continuous, so by Theorem 2.30 the function L_λ has a unique continuous extension $L_\lambda \colon \widehat{E} \to \widehat{E}$, which we use to define the scalar multiplication $\cdot \colon \mathbb{R} \times \widehat{E} \to \widehat{E}$. It is easily checked that with the above addition and scalar multiplication, $\widehat{E}$ is a vector space.

Since the norm $\| \|$ on E is uniformly continuous, it has a unique continuous extension $\| \|_{\widehat{E}} \colon \widehat{E} \to \mathbb{R}_+$. The identities $\|u + v\| \leq \|u\| + \|v\|$ and $\|\lambda u\| \leq |\lambda| \, \|u\|$ extend to $\widehat{E}$ by continuity. The equation

$$d(u, v) = \|u - v\|$$

also extends to $\widehat{E}$ by continuity and yields

$$\widehat{d}(\alpha, \beta) = \|\alpha - \beta\|_{\widehat{E}},$$

which shows that $\| \|_{\widehat{E}}$ is indeed a norm and that the metric $\widehat{d}$ is associated to it. Finally, it is easy to verify that the map φ is linear. The uniqueness of the structure of normed vector space follows from the uniqueness of continuous extensions in Theorem 2.30. □

Theorem 2.32 and Theorem 2.30 will be used to show that every Hermitian space can be embedded in a Hilbert space.

We refer the readers to the references cited at the end of this chapter for a discussion of the concepts of compactness and connectedness. They are important, but of less immediate concern.

2.10 The Contraction Mapping Theorem

If (E, d) is a nonempty complete metric space, every map $f\colon E \to E$, for which there is some k such that $0 \leq k < 1$ and

$$d(f(x), f(y)) \leq kd(x, y) \quad \text{for all } x, y \in E$$

has the very important property that it has a unique fixed point, that is, there is a unique, $a \in E$, such that $f(a) = a$.

Definition 2.28. Let (E, d) be a metric space. A map $f\colon E \to E$ is a *contraction* (or a *contraction mapping*) if there is some real number k such that $0 \leq k < 1$ and

$$d(f(u), f(v)) \leq kd(u, v) \quad \text{for all } u, v \in E.$$

The number k is often called a *Lipschitz constant*.

Furthermore, the fixed point of a contraction mapping can be computed as the limit of a fast converging sequence.

The fixed point property of contraction mappings is used to show some important theorems of analysis, such as the implicit function theorem and the existence of solutions to certain differential equations. It can also be used to show the existence of fractal sets defined in terms of iterated function systems. Since the proof is quite simple, we prove the fixed point property of contraction mappings. First observe that a contraction mapping is (uniformly) continuous.

Theorem 2.33. *(Contraction Mapping Theorem) If (E, d) is a nonempty complete metric space, every contraction mapping, $f\colon E \to E$, has a unique fixed point. Furthermore, for every $x_0 \in E$, if we define the sequence $(x_n)_{\geq 0}$ such that $x_{n+1} = f(x_n)$ for all $n \geq 0$, then $(x_n)_{n\geq 0}$ converges to the unique fixed point of f.*

Proof. First we prove that f has at most one fixed point. Indeed, if $f(a) = a$ and $f(b) = b$, since

$$d(a, b) = d(f(a), f(b)) \leq kd(a, b)$$

and $0 \leq k < 1$, we must have $d(a, b) = 0$, that is, $a = b$.

Next we prove that (x_n) is a Cauchy sequence. Observe that

$$\begin{aligned} d(x_2, x_1) &\leq kd(x_1, x_0), \\ d(x_3, x_2) &\leq kd(x_2, x_1) \leq k^2 d(x_1, x_0), \\ \vdots &\qquad\quad \vdots \\ d(x_{n+1}, x_n) &\leq kd(x_n, x_{n-1}) \leq \cdots \leq k^n d(x_1, x_0). \end{aligned}$$

Thus, we have

$$\begin{aligned} d(x_{n+p}, x_n) &\leq d(x_{n+p}, x_{n+p-1}) + d(x_{n+p-1}, x_{n+p-2}) + \cdots + d(x_{n+1}, x_n) \\ &\leq (k^{p-1} + k^{p-2} + \cdots + k + 1)k^n d(x_1, x_0) \\ &\leq \frac{k^n}{1-k}\, d(x_1, x_0). \end{aligned}$$

We conclude that $d(x_{n+p}, x_n)$ converges to 0 when n goes to infinity, which shows that (x_n) is a Cauchy sequence. Since E is complete, the sequence (x_n) has a limit, a. Since f is continuous, the sequence $(f(x_n))$ converges to $f(a)$. But $x_{n+1} = f(x_n)$ converges to a and so $f(a) = a$, the unique fixed point of f. □

The above theorem is also called the *Banach fixed point theorem*. Note that no matter how the starting point x_0 of the sequence (x_n) is chosen, (x_n) converges to the unique fixed point of f. Also, the convergence is fast, since

$$d(x_n, a) \leq \frac{k^n}{1-k}\, d(x_1, x_0).$$

2.11 Further Readings

A thorough treatment of general topology can be found in Munkres [59, 58], Dixmier [28], Lang [50], Schwartz [69, 68], Bredon [19], and the classic, Seifert and Threlfall [73].

2.12 Summary

The main concepts and results of this chapter are listed below:

- *Metric space*, *distance*, *metric*.
- *Euclidean metric*, *discrete metric*.
- *Closed ball*, *open ball*, *sphere*, *bounded subset*.
- *Normed vector space*, *norm*.
- *Open* and *closed* sets.
- *Topology*, *topological space*.
- *Hausdorff separation axiom*, *Hausdorff space*.
- *Discrete topology*.
- *Closure*, *dense subset*, *interior*, *frontier* or *boundary*.

- *Subspace topology*.
- *Product topology*.
- *Basis of a topology*, *subbasis of a topology*.
- *Continuous* functions.
- *Neighborhood* of a point.
- *Homeomorphisms*.
- *Limits* of sequences.
- *Continuous linear maps*.
- The *norm* of a continuous linear map.
- *Continuous bilinear maps*.
- The *norm* of a continuous bilinear map.
- The isomorphism between $\mathcal{L}(E, F; G)$ and $\mathcal{L}(E, \mathcal{L}(F; G))$.
- *Cauchy sequences*
- *Complete metric spaces* and *Banach spaces*.
- *Completion* of a metric space or of a normed vector space.
- *Contractions*.
- *The contraction mapping theorem*.

2.13 Problems

Problem 2.1. Prove Proposition 2.1.

Problem 2.2. Give an example of a countably infinite family of closed sets whose union is not closed.

Problem 2.3. Prove Proposition 2.4.

Problem 2.4. Prove Proposition 2.5.

Problem 2.5. Prove Proposition 2.6.

Problem 2.6. Prove Proposition 2.7.

Problem 2.7. Prove Proposition 2.8.

Problem 2.8. Prove Proposition 2.9.

Problem 2.9. Prove Proposition 2.10.

Problem 2.10. Prove Proposition 2.11 and Proposition 2.12.

Problem 2.11. Prove Proposition 2.14.

Problem 2.12. Prove Proposition 2.15.

Problem 2.13. Prove Proposition 2.16 and Proposition 2.17.

Problem 2.14. Prove Proposition 2.18.

Problem 2.15. Prove Proposition 2.20 and Proposition 2.21.

Problem 2.16. Prove Proposition 2.22.

Problem 2.17. Prove Proposition 2.23.

Chapter 3

Differential Calculus

This chapter contains a review of basic notions of differential calculus. First we review the definition of the derivative of a function $f: \mathbb{R} \to \mathbb{R}$. Next we define directional derivatives and the total derivative of a function $f: E \to F$ between normed vector spaces. Basic properties of derivatives are shown, including the chain rule. We show how derivatives are represented by Jacobian matrices. The mean value theorem is stated, as well as the implicit function theorem and the inverse function theorem. Diffeomorphisms and local diffeomorphisms are defined. Higher-order derivatives are defined, as well as the Hessian. Schwarz's lemma (about the commutativity of partials) is stated. Several versions of Taylor's formula are stated, and a famous formula due to Faà di Bruno's is given.

3.1 Directional Derivatives, Total Derivatives

We first review the notion of the derivative of a real-valued function whose domain is an open subset of $\mathbb{R}$.

Let $f: A \to \mathbb{R}$, where A is a nonempty open subset of $\mathbb{R}$, and consider any $a \in A$. The main idea behind the concept of the derivative of f at a, denoted by $f'(a)$, is that locally around a (that is, in some small open set $U \subseteq A$ containing a), the function f is approximated linearly[1] by the map

$$x \mapsto f(a) + f'(a)(x - a).$$

As pointed out by Dieudonné in the early 1960s, it is an "unfortunate accident" that if V is vector space of dimension one, then there is a bijection between the space V^* of linear forms defined on V and the field of scalars. As a consequence, the derivative of a real-valued function f defined on an open subset A of the reals can be defined as the scalar $f'(a)$ (for any $a \in A$). But as soon as f is a function of several arguments, the scalar interpretation of the derivative breaks down.

[1] Actually, the approximation is affine, but everybody commits this abuse of language.

Part of the difficulty in extending the idea of derivative to more complex spaces is to give an adequate notion of linear approximation. The key idea is to use linear maps. This could be carried out in terms of matrices but it turns out that this neither shortens nor simplifies proofs. In fact, this is often the opposite.

We admit that the more intrinsic definition of the notion of derivative f'_a at a point a of a function $f\colon E \to F$ between two normed vector spaces E and F as a linear map requires a greater effort to be grasped, but we feel that the advantages of this definition outweigh its degree of abstraction. In particular, it yields a clear notion of the derivative of a function $f\colon \mathrm{M}_m(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ defined from $m \times m$ matrices to $n \times n$ matrices (many definitions make use of partial derivatives with respect to matrices that do not make any sense). But more importantly, the definition of the derivative as a linear map makes it clear that whether the space E or the space F is infinite dimensional does not matter. This is important in optimization theory where the natural space of solutions of the problem is often an infinite dimensional function space. Of course, to carry out computations one need to pick finite bases and to use Jacobian matrices, but this is a different matter.

Let us now review the formal definition of the derivative of a real-valued function.

Definition 3.1. Let A be any nonempty open subset of $\mathbb{R}$, and let $a \in A$. For any function $f\colon A \to \mathbb{R}$, the *derivative of f at $a \in A$* is the limit (if it exists)

$$f'(a) = \lim_{h \to 0,\, h \in U} \frac{f(a+h) - f(a)}{h},$$

where $U = \{h \in \mathbb{R} \mid a + h \in A,\, h \neq 0\}$. This limit is denoted by $f'(a)$, or $\mathrm{D}f(a)$, or $\frac{df}{dx}(a)$. If $f'(a)$ exists for every $a \in A$, we say that f *is differentiable on A*. In this case, the map $a \mapsto f'(a)$ is denoted by f', or $\mathrm{D}f$, or $\frac{df}{dx}$.

Note that since A is assumed to be open, $A - \{a\}$ is also open, and since the function $h \mapsto a + h$ is continuous and U is the inverse image of $A - \{a\}$ under this function, U is indeed open and the definition makes sense.

We can also define $f'(a)$ as follows: there is some function ϵ, such that,

$$f(a+h) = f(a) + f'(a) \cdot h + \epsilon(h)h,$$

whenever $a + h \in A$, where $\epsilon(h)$ is defined for all h such that $a + h \in A$, and

$$\lim_{h \to 0,\, h \in U} \epsilon(h) = 0.$$

Remark: We can also define the notion of *derivative of f at a on the left*, and *derivative of f at a on the right*. For example, we say that the *derivative of f at a on the left* is the limit $f'(a_-)$ (if it exists)

$$f'(a_-) = \lim_{h \to 0,\, h \in U} \frac{f(a+h) - f(a)}{h},$$

where $U = \{h \in \mathbb{R} \mid a + h \in A,\ h < 0\}$.

If a function f as in Definition 3.1 has a derivative $f'(a)$ at a, then it is continuous at a. If f is differentiable on A, then f is continuous on A. The composition of differentiable functions is differentiable.

Remark: A function f has a derivative $f'(a)$ at a iff the derivative of f on the left at a and the derivative of f on the right at a exist and if they are equal. Also, if the derivative of f on the left at a exists, then f is continuous on the left at a (and similarly on the right).

We would like to extend the notion of derivative to functions $f\colon A \to F$, where E and F are normed vector spaces, and A is some nonempty open subset of E. The first difficulty is to make sense of the quotient

$$\frac{f(a+h) - f(a)}{h}$$

if E has dimension greater than 1.

Since F is a normed vector space, $f(a+h) - f(a)$ makes sense. But how do we define the quotient by a vector? Well, we don't!

A first possibility is to consider the *directional derivative* with respect to a vector $u \neq 0$ in E. We can consider the vector $f(a+tu) - f(a)$, where $t \in \mathbb{R}$. Now,

$$\frac{f(a+tu) - f(a)}{t}$$

makes sense.

The idea is that in E, the points of the form $a+tu$ for t in some small interval $[-\epsilon, +\epsilon]$ in $\mathbb{R}$ form a line segment $[r, s]$ in A containing a, and that the image of this line segment defines a small curve segment on $f(A)$. This curve segment is defined by the map $t \mapsto f(a+tu)$, from $[r, s]$ to F, and the directional derivative $\mathrm{D}_u f(a)$ defines the direction of the tangent line at a to this curve; see Figure 3.1. This leads us to the following definition.

Definition 3.2. Let E and F be two normed vector spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, for any $u \neq 0$ in E, the *directional derivative of f at a w.r.t. the vector u*, denoted by $\mathrm{D}_u f(a)$, is the limit (if it exists)

$$\mathrm{D}_u f(a) = \lim_{t \to 0,\, t \in U} \frac{f(a+tu) - f(a)}{t},$$

where $U = \{t \in \mathbb{R} \mid a + tu \in A,\ t \neq 0\}$ (or $U = \{t \in \mathbb{C} \mid a + tu \in A,\ t \neq 0\}$).

Since the map $t \mapsto a + tu$ is continuous, and since $A - \{a\}$ is open, the inverse image U of $A - \{a\}$ under the above map is open, and the definition of the limit in Definition 3.2 makes sense. The directional derivative is sometimes called the *Gâteaux derivative*.

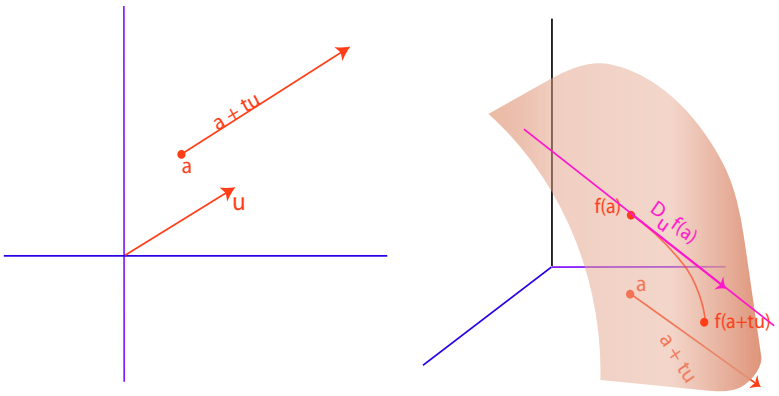


Figure 3.1: Let $f\colon \mathbb{R}^2 \to \mathbb{R}$. The graph of f is the peach surface in $\mathbb{R}^3$, and $t \mapsto f(a+tu)$ is the embedded orange curve connecting $f(a)$ to $f(a+tu)$. Then $\mathrm{D}_u f(a)$ is the slope of the pink tangent line in the direction of u.

Remark: Since the notion of limit is purely topological, the existence and value of a directional derivative is independent of the choice of norms in E and F, as long as they are equivalent norms.

In the special case where $E = \mathbb{R}$ and $F = \mathbb{R}$, and we let $u = 1$ (i.e., the real number 1, viewed as a vector), it is immediately verified that $\mathrm{D}_1 f(a) = f'(a)$, in the sense of Definition 3.1. When $E = \mathbb{R}$ (or $E = \mathbb{C}$) and F is any normed vector space, the derivative $\mathrm{D}_1 f(a)$, also denoted by $f'(a)$, provides a suitable generalization of the notion of derivative.

However, when E has dimension ≥ 2, directional derivatives present a serious problem, which is that their definition is not sufficiently uniform. Indeed, there is no reason to believe that the directional derivatives w.r.t. all nonnull vectors u share something in common. As a consequence, a function can have all directional derivatives at a, and yet not be continuous at a. Two functions may have all directional derivatives in some open sets, and yet their composition may not.

Example 3.1. Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be the function given by

$$f(x,y) = \begin{cases} \frac{x^2 y}{x^4+y^2} & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

The graph of $f(x,y)$ is illustrated in Figure 3.2.

For any $u \neq 0$, letting $u = \begin{pmatrix} h \\ k \end{pmatrix}$, we have

$$\frac{f(0+tu) - f(0)}{t} = \frac{h^2 k}{t^2 h^4 + k^2},$$

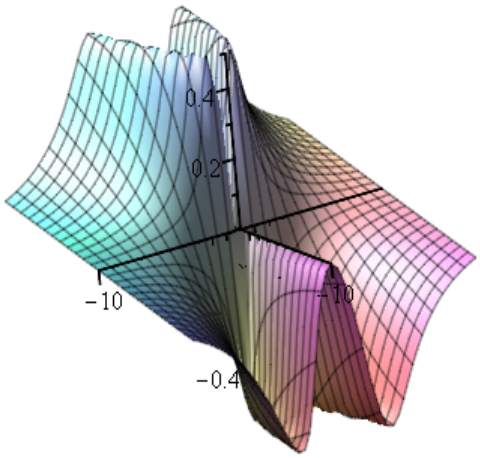


Figure 3.2: The graph of the function from Example 3.1. Note that f is not continuous at $(0,0)$, despite the existence of $\mathrm{D}_u f(0,0)$ for all $u \neq 0$.

so that

$$\mathrm{D}_u f(0,0) = \begin{cases} \frac{h^2}{k} & \text{if } k \neq 0 \\ 0 & \text{if } k = 0. \end{cases}$$

Thus, $\mathrm{D}_u f(0,0)$ exists for all $u \neq 0$.

On the other hand, if $\mathrm{D}f(0,0)$ existed, it would be a linear map $\mathrm{D}f(0,0)\colon \mathbb{R}^2 \to \mathbb{R}$ represented by a row matrix $(\alpha \ \beta)$, and we would have $\mathrm{D}_u f(0,0) = \mathrm{D}f(0,0)(u) = \alpha h + \beta k$, but the explicit formula for $\mathrm{D}_u f(0,0)$ is not linear. As a matter of fact, the function f is not continuous at $(0,0)$. For example, on the parabola $y = x^2$, $f(x,y) = \frac{1}{2}$, and when we approach the origin on this parabola, the limit is $\frac{1}{2}$, but $f(0,0) = 0$.

To avoid the problems arising with directional derivatives we introduce a more uniform notion.

Given two normed spaces E and F, recall that a linear map $f\colon E \to F$ is *continuous* iff there is some constant $C \geq 0$ such that

$$\|f(u)\| \leq C\,\|u\| \quad \text{for all } u \in E.$$

Definition 3.3. Let E and F be two normed vector spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, we say that f is *differentiable at* $a \in A$ if there is a *continuous linear* map $L\colon E \to F$ and a function $h \mapsto \epsilon(h)$, such that

$$f(a+h) = f(a) + L(h) + \epsilon(h)\|h\|$$

for every $a + h \in A$, where $\epsilon(h)$ is defined for every h such that $a + h \in A$, and

$$\lim_{h \to 0,\, h \in U} \epsilon(h) = 0,$$

where $U = \{h \in E \mid a + h \in A,\, h \neq 0\}$. The linear map L is denoted by $\mathrm{D}f(a)$, or $\mathrm{D}f_a$, or $df(a)$, or df_a, or $f'(a)$, and it is called the *Fréchet derivative*, or *derivative*, or *total derivative*, or *total differential*, or *differential* of f at a; see Figure 3.3.

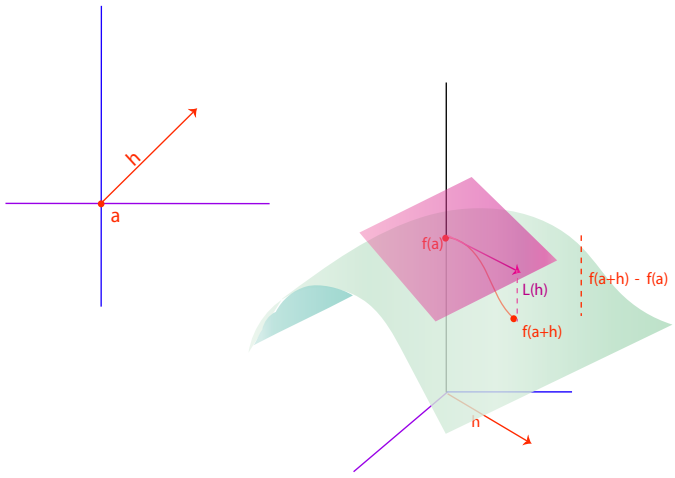


Figure 3.3: Let $f\colon \mathbb{R}^2 \to \mathbb{R}$. The graph of f is the green surface in $\mathbb{R}^3$. The linear map $L = \mathrm{D}f(a)$ is the pink tangent plane. For any vector $h \in \mathbb{R}^2$, $L(h)$ is approximately equal to $f(a + h) - f(a)$. Note that $L(h)$ is also the direction tangent to the curve $t \mapsto f(a + th)$.

Since the map $h \mapsto a + h$ from E to E is continuous, and since A is open in E, the inverse image U of $A - \{a\}$ under the above map is open in E, and it makes sense to say that

$$\lim_{h \to 0,\, h \in U} \epsilon(h) = 0.$$

Note that for every $h \in U$, since $h \neq 0$, $\epsilon(h)$ is uniquely determined since

$$\epsilon(h) = \frac{f(a + h) - f(a) - L(h)}{\|h\|},$$

and that the value $\epsilon(0)$ plays absolutely no role in this definition. The condition for f to be differentiable at a amounts to the fact that

$$\lim_{h \mapsto 0} \frac{\|f(a + h) - f(a) - L(h)\|}{\|h\|} = 0 \qquad (\dagger)$$

as $h \neq 0$ approaches 0, when $a + h \in A$. However, it does no harm to assume that $\epsilon(0) = 0$, and we will assume this from now on.

Again, we note that the derivative $\mathrm{D}f(a)$ of f at a provides an affine approximation of f, locally around a.

If $E = F = \mathbb{R}$, a linear map $L\colon \mathbb{R} \to \mathbb{R}$ is uniquely determined by some fixed real number $c \in \mathbb{R}$ and we have

$$L(u) = cu, \quad u \in \mathbb{R}.$$

Since

$$|\epsilon(h)| = \frac{|f(a+h) - f(a) - ch|}{|h|} = \left| \frac{f(a+h) - f(a)}{h} - c \right|,$$

if $\mathrm{D}f_a$ exists, $|\epsilon(h)|$ tends to zero if $|h|$ tends to zero, and we deduce that

$$c = f'(a),$$

so our new notion of derivative agrees with the old (standard) notion of derivative, as it should.

Remarks:

(1) Since the notion of limit is purely topological, the existence and value of a derivative is independent of the choice of norms in E and F, as long as they are equivalent norms.

(2) If $h\colon (-a, a) \to \mathbb{R}$ is a real-valued function defined on some open interval containing 0, we say that h *is* $o(t)$ *for* $t \to 0$, and we write $h(t) = o(t)$, if

$$\lim_{t \mapsto 0,\, t \neq 0} \frac{h(t)}{t} = 0.$$

With this notation (the *little o notation*), the function f is differentiable at a iff

$$f(a+h) - f(a) - L(h) = o(\|h\|),$$

which is also written as

$$f(a+h) = f(a) + L(h) + o(\|h\|).$$

The following proposition shows that our new definition is consistent with the definition of the directional derivative and that *the continuous linear map L is unique*, if it exists.

Proposition 3.1. *Let E and F be two normed spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, if $\mathrm{D}f(a)$ is defined, then f is continuous at a and f has a directional derivative $\mathrm{D}_u f(a)$ for every $u \neq 0$ in E. Furthermore,*

$$\mathrm{D}_u f(a) = \mathrm{D}f(a)(u),$$

which can also be written as

$$\mathrm{D}_u f(a) = \mathrm{D}f_a(u),$$

and thus, $\mathrm{D}f(a)$ is uniquely defined.

Proof. If $L = \mathrm{D}f(a)$ exists, then for any nonzero vector $u \in E$, because A is open, for any $t \in \mathbb{R} - \{0\}$ (or $t \in \mathbb{C} - \{0\}$) small enough, $a + tu \in A$, so

$$\begin{aligned} f(a+tu) &= f(a) + L(tu) + \epsilon(tu)\|tu\| \\ &= f(a) + tL(u) + |t|\epsilon(tu)\|u\| \end{aligned}$$

which implies that

$$L(u) = \frac{f(a+tu) - f(a)}{t} - \frac{|t|}{t}\epsilon(tu)\|u\|,$$

and since $\lim_{t \mapsto 0} \epsilon(tu) = 0$, we deduce that

$$L(u) = \mathrm{D}f(a)(u) = \mathrm{D}_u f(a).$$

Because

$$f(a+h) = f(a) + L(h) + \epsilon(h)\|h\|$$

for all h such that $\|h\|$ is small enough, L is continuous, and $\lim_{h \mapsto 0} \epsilon(h)\|h\| = 0$, we have $\lim_{h \mapsto 0} f(a+h) = f(a)$, that is, f is continuous at a. □

When E is of finite dimension, every linear map is continuous (see Proposition 8.8 (Vol. I) or Theorem 2.26), and this assumption is then redundant.

Observe that in the equation

$$\mathrm{D}_u f(a) = \mathrm{D}f_a(u),$$

the directional derivative $\mathrm{D}_u f(a)$ is a *vector*, and $\mathrm{D}f_a(u)$ is the *result of evaluating the linear map* $\mathrm{D}f_a$ *on the vector* u. The linear map $\mathrm{D}f_a$ "knows" about all the directional derivatives, it is a global object. So after all, the derivative $\mathrm{D}f_a$ of a function f at a is *not a number or a vector, it is a linear map*.

Although this may not be immediately obvious, the reason for requiring the linear map $\mathrm{D}f_a$ to be continuous is to ensure that if a function f is differentiable at a, then it is continuous at a. This is certainly a desirable property of a differentiable function. In finite dimension this holds, but in infinite dimension this is not the case. The following proposition shows that if $\mathrm{D}f_a$ exists at a and if f is continuous at a, then $\mathrm{D}f_a$ must be a continuous map. So if a function is differentiable at a, then it is continuous iff the linear map $\mathrm{D}f_a$ is continuous. We chose to include the second condition rather that the first in the definition of a differentiable function.

Proposition 3.2. *Let E and F be two normed spaces, let A be a nonempty open subset of E, and let $f \colon A \to F$ be any function. For any $a \in A$, if $\mathrm{D}f_a$ is defined, then f is continuous at a iff $\mathrm{D}f_a$ is a continuous linear map.*

Proof. Proposition 3.1 shows that if $\mathrm{D}f_a$ is defined and continuous then f is continuous at a. Conversely, assume that $\mathrm{D}f_a$ exists and that f is continuous at a. Since f is continuous at a and since $\mathrm{D}f_a$ exists, for any $\eta > 0$ there is some ρ with $0 < \rho < 1$ such that if $\|h\| \leq \rho$ then

$$\|f(a+h) - f(a)\| \leq \frac{\eta}{2},$$

and

$$\|f(a+h) - f(a) - \mathrm{D}_a(h)\| \leq \frac{\eta}{2}\|h\| \leq \frac{\eta}{2},$$

so we have

$$\begin{aligned}\|\mathrm{D}_a(h)\| &= \|\mathrm{D}_a(h) - (f(a+h) - f(a)) + f(a+h) - f(a)\| \\ &\leq \|f(a+h) - f(a) - \mathrm{D}_a(h)\| + \|f(a+h) - f(a)\| \\ &\leq \frac{\eta}{2} + \frac{\eta}{2} = \eta,\end{aligned}$$

which proves that $\mathrm{D}f_a$ is continuous at 0. By Proposition 2.24, $\mathrm{D}f_a$ is a continuous linear map. □

In practice, to find the linear map $L = \mathrm{D}f_a = df_a$, we try to expand $f(a+h) - f(a)$ as a function of h and to isolate the part of $f(a+h) - f(a)$ which is linear in h. For functions on matrices, this is typically not too hard. Then we need to show that the error term $\epsilon(h)$ tends to zero as $\|h\|$ tends to zero, which relies on suitable properties of the norms involved and may be quite challenging.

Example 3.2. Consider the map $f\colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ given by

$$f(A) = A^\top A - I,$$

where $\mathrm{M}_n(\mathbb{R})$ denotes the vector space of all $n \times n$ matrices with real entries equipped with any matrix norm, since they are all equivalent; for example, pick the Frobenius norm $\|A\|_F = \sqrt{\mathrm{tr}(A^\top A)}$. We claim that

$$Df(A)(H) = A^\top H + H^\top A, \quad \text{for all } A \text{ and } H \text{ in } \mathrm{M}_n(\mathbb{R}).$$

We have

$$\begin{aligned}f(A+H) - f(A) - (A^\top H + H^\top A) &= (A+H)^\top(A+H) - I - (A^\top A - I) - A^\top H - H^\top A \\ &= A^\top A + A^\top H + H^\top A + H^\top H - A^\top A - A^\top H - H^\top A \\ &= H^\top H.\end{aligned}$$

It follows that

$$\epsilon(H) = \frac{f(A+H) - f(A) - (A^\top H + H^\top A)}{\|H\|} = \frac{H^\top H}{\|H\|},$$

and since our norm is the Frobenius norm,

$$\|\epsilon(H)\| = \left\| \frac{H^\top H}{\|H\|} \right\| \leq \frac{\left\|H^\top\right\| \|H\|}{\|H\|} = \left\|H^\top\right\| = \|H\|,$$

so

$$\lim_{H \mapsto 0} \epsilon(H) = 0,$$

and we conclude that

$$Df(A)(H) = A^\top H + H^\top A.$$

Definition 3.4. If the derivative $\mathrm{D}f_a$ of a function $f\colon A \to F$ exists for every $a \in A$, we get a map $\mathrm{D}f\colon A \to \mathcal{L}(E;F)$, called the *derivative of f on A*, also denoted by *df*. Here $\mathcal{L}(E;F)$ denotes the vector space of continuous linear maps from E to F.

Note that according to the old notion of the derivative of a function $f\colon A \to \mathbb{R}$, the derivative f' of f is the map $f'\colon A \to \mathbb{R}$ given by $a \mapsto f'(a)$. But the "right" notion is that the derivative $\mathrm{D}f$ of f is a map from A to the space $\mathcal{L}(E;F)$ of continuous linear maps from E to F. When $E = F = \mathbb{R}$, there is an accidental isomorphism between $\mathcal{L}(\mathbb{R};\mathbb{R})$ and $\mathbb{R}$.

We now consider a number of standard results about derivatives.

3.2 Properties of Derivatives

A function $f\colon E \to F$ is said to be *affine* if there is some linear map $\overrightarrow{f}\colon E \to F$ and some fixed vector $c \in F$, such that

$$f(u) = \overrightarrow{f}(u) + c$$

for all $u \in E$. We call $\overrightarrow{f}$ the *linear map associated with f*.

Proposition 3.3. *Given two normed spaces E and F, if $f\colon E \to F$ is a constant function, then $\mathrm{D}f(a) = 0$, for every $a \in E$ (here, 0 denotes the linear map from E to F whose value is $0 \in F$ for all $u \in E$). If $f\colon E \to F$ is a continuous affine map, then $\mathrm{D}f(a) = \overrightarrow{f}$, for every $a \in E$, where $\overrightarrow{f}$ denotes the linear map associated with f. In particular, if $f\colon E \to F$ is a continuous linear map, then $\mathrm{D}f(a) = f$ for all $a \in E$.*

Proof. If $f\colon E \to F$ is the constant function such that $f(u) = c$ for all $u \in E$ for some fixed $c \in F$, then

$$f(a+h) - f(a) = c - c = 0 = 0\,\|h\|,$$

so $\epsilon(h) = 0$ for all $h \in E$, and by definition, $\mathrm{D}f(a)(h) = 0$ for all $a, h \in E$.

If $f\colon E \to F$ is a continuous affine map, since $\overrightarrow{f}$ is linear, we have

$$\begin{aligned} f(a+h) - f(a) &= \overrightarrow{f}(a+h) + c - (\overrightarrow{f}(a) + c) = \overrightarrow{f}(a) + \overrightarrow{f}(h) + c - (\overrightarrow{f}(a) + c) \\ &= \overrightarrow{f}(h) = \overrightarrow{f}(h) + 0\,\|h\|, \end{aligned}$$

so $\epsilon(h) = 0$ for all $h \in E$, and by definition, $\mathrm{D}f(a) = \overrightarrow{f}$ for all $a \in E$. □

Example 3.3. Consider the function $f\colon \mathbb{R}^n \to \mathrm{M}_{1,m}(\mathbb{R})$ given by

$$\varphi(v) = (Cv - d)^\top, \quad v \in \mathbb{R}^n,$$

where C is an $m \times n$ matrix and $d \in \mathbb{R}^m$. Since transposition is linear,

$$\varphi(v) = (Cv)^\top - d^\top$$

is an affine map with linear part given by $v \mapsto (Cv)^\top$, so by Proposition 3.3,

$$\mathrm{D}\varphi_v(w) = (Cw)^\top, \quad v, w \in \mathbb{R}^n.$$

Proposition 3.4. *Given a normed space E and a normed vector space F, for any two functions $f, g\colon E \to F$, for every $a \in E$, if $\mathrm{D}f(a)$ and $\mathrm{D}g(a)$ exist, then $\mathrm{D}(f+g)(a)$ and $\mathrm{D}(\lambda f)(a)$ exist, and*

$$\mathrm{D}(f+g)(a) = \mathrm{D}f(a) + \mathrm{D}g(a),$$
$$\mathrm{D}(\lambda f)(a) = \lambda \mathrm{D}f(a).$$

Given two normed vector spaces $(E_1, \| \|_1)$ and $(E_2, \| \|_2)$, there are three natural and equivalent norms that can be used to make $E_1 \times E_2$ into a normed vector space:

1. $\|(u_1, u_2)\|_1 = \|u_1\|_1 + \|u_2\|_2$.
2. $\|(u_1, u_2)\|_2 = (\|u_1\|_1^2 + \|u_2\|_2^2)^{1/2}$.
3. $\|(u_1, u_2)\|_\infty = \max(\|u_1\|_1, \|u_2\|_2)$.

We usually pick the first norm. If E_1, E_2, and F are three normed vector spaces, recall that a bilinear map $f\colon E_1 \times E_2 \to F$ is *continuous* iff there is some constant $C \geq 0$ such that

$$\|f(u_1, u_2)\| \leq C \|u_1\|_1 \|u_2\|_2 \quad \text{for all } u_1 \in E_1 \text{ and all } u_2 \in E_2.$$

Proposition 3.5. *Given three normed vector spaces E_1, E_2, and F, for any continuous bilinear map $f\colon E_1 \times E_2 \to F$, for every $(a, b) \in E_1 \times E_2$, $\mathrm{D}f(a, b)$ exists, and for every $u \in E_1$ and $v \in E_2$,*

$$\mathrm{D}f(a, b)(u, v) = f(u, b) + f(a, v).$$

Proof. Since f is bilinear, a simple computation implies that

$$\begin{aligned} f((a, b) + (u, v)) - f(a, b) - (f(u, b) + f(a, v)) &= f(a + u, b + v) - f(a, b) - f(u, b) - f(a, v) \\ &= f(a + u, b) + f(a + u, v) - f(a, b) - f(u, b) - f(a, v) \\ &= f(a, b) + f(u, b) + f(a, v) + f(u, v) - f(a, b) - f(u, b) - f(a, v) \\ &= f(u, v). \end{aligned}$$

We define

$$\epsilon(u,v) = \frac{f((a,b)+(u,v)) - f(a,b) - (f(u,b)+f(a,v))}{\|(u,v)\|_1},$$

and observe that the continuity of f implies

$$\begin{aligned}\|f((a,b)+(u,v)) - f(a,b) - (f(u,b)+f(a,v))\| &= \|f(u,v)\| \\ &\leq C\,\|u\|_1\,\|v\|_2 \leq C\,(\|u\|_1 + \|v\|_2)^2.\end{aligned}$$

Hence

$$\|\epsilon(u,v)\| = \left\| \frac{f(u,v)}{\|(u,v)\|_1} \right\| = \frac{\|f(u,v)\|}{\|(u,v)\|_1} \leq \frac{C\,(\|u\|_1 + \|v\|_2)^2}{\|u\|_1 + \|v\|_2} = C\,(\|u\|_1 + \|v\|_2) = C\,\|(u,v)\|_1,$$

which in turn implies

$$\lim_{(u,v)\mapsto(0,0)} \epsilon(u,v) = 0.$$

□

Example 3.4. Consider the function $f\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ given by

$$f(v,\lambda) = \lambda^\top C v = v^\top C^\top \lambda, \quad v \in \mathbb{R}^n, \lambda \in \mathbb{R}^m,$$

where C is an $m \times n$ matrix. Since the function f is bilinear, by Proposition 3.5, its derivative is given by

$$df_{(v,\lambda)}(w,\mu) = f(w,\lambda) + f(v,\mu) = \lambda^\top C w + \mu^\top C v = (C^\top\lambda)^\top w + (Cv)^\top \mu,$$

with $w \in \mathbb{R}^n, \mu \in \mathbb{R}^m$.

We now state the very useful *chain rule*.

Theorem 3.6. *Given three normed spaces E, F, and G, let A be an open set in E, and let B an open set in F. For any functions $f\colon A \to F$ and $g\colon B \to G$, such that $f(A) \subseteq B$, for any $a \in A$, if $\mathrm{D}f(a)$ exists and $\mathrm{D}g(f(a))$ exists, then $\mathrm{D}(g \circ f)(a)$ exists, and*

$$\mathrm{D}(g \circ f)(a) = \mathrm{D}g(f(a)) \circ \mathrm{D}f(a),$$

which can also be written as

$$\mathrm{D}(g \circ f)_a = \mathrm{D}g_{f(a)} \circ \mathrm{D}f_a \quad \textit{or} \quad d(g \circ f)_a = dg_{f(a)} \circ df_a.$$

Proof. Since f is differentiable at a and g is differentiable at $b = f(a)$, for every η such that $0 < \eta < 1$ there is some $\rho > 0$ such that for all s, t, if $\|s\| \leq \rho$ and $\|t\| \leq \rho$ then

$$\begin{aligned} f(a+s) &= f(a) + \mathrm{D}f_a(s) + \epsilon_1(s) \\ g(b+t) &= g(b) + \mathrm{D}g_b(t) + \epsilon_2(t), \end{aligned}$$

with $\|\epsilon_1(s)\| \leq \eta \|s\|$ and $\|\epsilon_2(t)\| \leq \eta \|t\|$. Since $\mathrm{D}f_a$ and $\mathrm{D}g_b$ are continuous, we have

$$\|\mathrm{D}f_a(s)\| \leq \|\mathrm{D}f_a\| \|s\| \quad \text{and} \quad \|\mathrm{D}g_b(t)\| \leq \|\mathrm{D}g_b\| \|t\|,$$

which, since $\|\epsilon_1(s)\| \leq \eta \|s\|$ and $\eta < 1$, implies that

$$\|\mathrm{D}f_a(s) + \epsilon_1(s)\| \leq \|\mathrm{D}f_a\| \|s\| + \|\epsilon_1(s)\| \leq \|\mathrm{D}f_a\| \|s\| + \eta \|s\| \leq (\|\mathrm{D}f_a\| + 1) \|s\|.$$

Consequently, if $\|s\| < \rho/(\|\mathrm{D}f_a\| + 1)$, we have

$$\|\epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s))\| \leq \eta(\|\mathrm{D}f_a\| + 1) \|s\|, \tag{$*_1$}$$

and

$$\|\mathrm{D}g_b(\epsilon_1(s))\| \leq \|\mathrm{D}g_b\| \|\epsilon_1(s)\| \leq \eta \|\mathrm{D}g_b\| \|s\|. \tag{$*_2$}$$

Then since $b = f(a)$, using the above we have

$$\begin{aligned}(g \circ f)(a + s) &= g(f(a + s)) = g(b + \mathrm{D}f_a(s) + \epsilon_1(s)) \\ &= g(b) + \mathrm{D}g_b(\mathrm{D}f_a(s) + \epsilon_1(s)) + \epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s)) \\ &= g(b) + (\mathrm{D}g_b \circ \mathrm{D}f_a)(s) + \mathrm{D}g_b(\epsilon_1(s)) + \epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s)).\end{aligned}$$

Now by $(*_1)$ and $(*_2)$ we have

$$\begin{aligned}\|\mathrm{D}g_b(\epsilon_1(s)) + \epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s))\| &\leq \|\mathrm{D}g_b(\epsilon_1(s))\| + \|\epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s))\| \\ &\leq \eta \|\mathrm{D}g_b\| \|s\| + \eta(\|\mathrm{D}f_a\| + 1) \|s\| \\ &= \eta(\|\mathrm{D}f_a\| + \|\mathrm{D}g_b\| + 1) \|s\|,\end{aligned}$$

so if we write $\epsilon_3(s) = \mathrm{D}g_b(\epsilon_1(s)) + \epsilon_2(\mathrm{D}f_a(s) + \epsilon_1(s))$ we proved that

$$(g \circ f)(a + s) = g(b) + (\mathrm{D}g_b \circ \mathrm{D}f_a)(s) + \epsilon_3(s)$$

with $\epsilon_3(s) \leq \eta(\|\mathrm{D}f_a\| + \|\mathrm{D}g_b\| + 1) \|s\|$, which proves that $\mathrm{D}g_b \circ \mathrm{D}f_a$ is the derivative of $g \circ f$ at a. Since $\mathrm{D}f_a$ and $\mathrm{D}g_b$ are continuous, so is $\mathrm{D}g_b \circ \mathrm{D}f_a$, which proves our proposition. □

In the special case where $E = F = G = \mathbb{R}$ and A and B are open subsets of $\mathbb{R}$ such that $f(A) \subseteq B$, the chain rule gives us back the standard version of the chain rule for functions $f\colon A \to \mathbb{R}$ and $g\colon B \to \mathbb{R}$, namely for any $a \in A$,

$$(g \circ f)'(a) = g'(f(a)) \cdot f'(a),$$

the product of the real numbers $g'(f(a))$ and $f'(a)$.

This is because for any $a \in A$ and $b \in B$, $\mathrm{D}f_a$ is the linear map given by $\mathrm{D}f_a(u) = f'(a)u$ and $\mathrm{D}g_b$ is the linear map given by $\mathrm{D}g_b(v) = g'(b)v$, with $u, v \in \mathbb{R}$, so the chain rule with $b = f(a)$ and $v = f'(a)u$ says that

$$\mathrm{D}(g \circ f)_a(u) = \mathrm{D}g_{f(a)}(\mathrm{D}f_a(u)) = g'(f(a))(f'(a)u) = (g'(f(a))f'(a))u,$$

so indeed

$$(g \circ f)'(a) = g'(f(a))f'(a).$$

Theorem 3.6 has many interesting consequences. We mention two corollaries.

Proposition 3.7. *Given three normed vector spaces E, F, and G, for any open subset A in E, for any $a \in A$, let $f\colon A \to F$ such that $\mathrm{D}f(a)$ exists, and let $g\colon F \to G$ be a continuous affine map. Then $\mathrm{D}(g \circ f)(a)$ exists, and*

$$\mathrm{D}(g \circ f)(a) = \overrightarrow{g} \circ \mathrm{D}f(a),$$

where $\overrightarrow{g}$ is the linear map associated with the affine map g.

Proposition 3.8. *Given two normed vector spaces E and F, let A be some open subset in E, let B be some open subset in F, let $f\colon A \to B$ be a bijection from A to B, and assume that $\mathrm{D}f$ exists on A and that $\mathrm{D}f^{-1}$ exists on B. Then for every $a \in A$,*

$$\mathrm{D}f^{-1}(f(a)) = (\mathrm{D}f(a))^{-1}.$$

Proposition 3.8 has the remarkable consequence that the two vector spaces E and F have the same dimension. In other words, a local property, the existence of a bijection f between an open set A of E and an open set B of F, such that f is differentiable on A and f^{-1} is differentiable on B, implies a global property, that the two vector spaces E and F have the same dimension.

Let us mention two more rules about derivatives that are used all the time.

Let $\iota\colon \mathbf{GL}(n,\mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ be the function (inversion) defined on invertible $n \times n$ matrices by

$$\iota(A) = A^{-1}.$$

Observe that $\mathbf{GL}(n,\mathbb{C})$ is indeed an open subset of the normed vector space $\mathrm{M}_n(\mathbb{C})$ of complex $n \times n$ matrices, since its complement is the closed set of matrices $A \in \mathrm{M}_n(\mathbb{C})$ satisfying $\det(A) = 0$. Then we have

$$d\iota_A(H) = -A^{-1}HA^{-1},$$

for all $A \in \mathbf{GL}(n,\mathbb{C})$ and for all $H \in \mathrm{M}_n(\mathbb{C})$.

To prove the preceding line observe that for H with sufficiently small norm, we have

$$\begin{aligned}
\iota(A+H) - \iota(A) + A^{-1}HA^{-1} &= (A+H)^{-1} - A^{-1} + A^{-1}HA^{-1} \\
&= (A+H)^{-1}[I - (A+H)A^{-1} + (A+H)A^{-1}HA^{-1}] \\
&= (A+H)^{-1}[I - I - HA^{-1} + HA^{-1} + HA^{-1}HA^{-1}] \\
&= (A+H)^{-1}HA^{-1}HA^{-1}.
\end{aligned}$$

Consequently, we get

$$\epsilon(H) = \frac{\iota(A+H) - \iota(A) + A^{-1}HA^{-1}}{\|H\|} = \frac{(A+H)^{-1}HA^{-1}HA^{-1}}{\|H\|},$$

and since

$$\left\|(A+H)^{-1}HA^{-1}HA^{-1}\right\| \leq \|H\|^2 \left\|A^{-1}\right\|^2 \left\|(A+H)^{-1}\right\|,$$

it is clear that $\lim_{H\mapsto 0} \epsilon(H) = 0$, which proves that

$$d\iota_A(H) = -A^{-1}HA^{-1}.$$

In particular, if $A = I$, then $d\iota_I(H) = -H$.

Next, if $f\colon \mathrm{M}_n(\mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ and $g\colon \mathrm{M}_n(\mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ are differentiable matrix functions, then

$$d(fg)_A(B) = df_A(B)g(A) + f(A)dg_A(B),$$

for all $A, B \in \mathbf{M}_n(\mathbb{C})$. This is known as the *product rule*.

In preparation for the next section on Jacobian matrices and the section on the implicit function theorem we need the following definitions.

When E is of finite dimension n, for any basis $(u_1, \ldots, u_n)$ of E, we can define the directional derivatives with respect to the vectors in the basis $(u_1, \ldots, u_n)$ (actually, we can also do it for an infinite basis). This way we obtain the definition of partial derivatives as follows:

Definition 3.5. For any two normed spaces E and F, if E is of finite dimension n, for every basis $(u_1, \ldots, u_n)$ for E, for every $a \in E$, for every function $f\colon E \to F$, the directional derivatives $\mathrm{D}_{u_j} f(a)$ (if they exist) are called the *partial derivatives of f with respect to the basis* $(u_1, \ldots, u_n)$. The partial derivative $\mathrm{D}_{u_j} f(a)$ is also denoted by $\partial_j f(a)$, or $\dfrac{\partial f}{\partial x_j}(a)$.

The notation $\dfrac{\partial f}{\partial x_j}(a)$ for a partial derivative, although customary and going back to Leibniz, is a "logical obscenity." Indeed, the variable x_j really has nothing to do with the formal definition. This is just another of these situations where tradition is just too hard to overthrow!

More generally we now consider the situation where E is a finite direct sum. Given a normed vector space $E = E_1 \oplus \cdots \oplus E_n$ and a normed vector space F, given any open subset A of E, for any $c = (c_1, \ldots, c_n) \in A$, we define the continuous functions $i_j^c\colon E_j \to E$, such that

$$i_j^c(x) = (c_1, \ldots, c_{j-1}, x, c_{j+1}, \ldots, c_n).$$

For any function $f\colon A \to F$, we have functions $f \circ i_j^c\colon E_j \to F$ defined on $(i_j^c)^{-1}(A)$, which contains c_j.

Definition 3.6. If $\mathrm{D}(f \circ i_j^c)(c_j)$ exists, we call it the *partial derivative of f w.r.t. its jth argument, at c*. We also denote this derivative by $\mathrm{D}_j f(c)$ of $\dfrac{\partial f}{\partial x_j}(c)$. Note that $\mathrm{D}_j f(c) \in \mathcal{L}(E_j; F)$.

This notion is a generalization of the notion defined in Definition 3.5. In fact, when E is of dimension n, and a basis $(u_1, \ldots, u_n)$ has been chosen, we can write $E = Ku_1 \oplus \cdots \oplus Ku_n$, (with $K = \mathbb{R}$ or $K = \mathbb{C}$), and then

$$\mathrm{D}_j f(c)(\lambda u_j) = \lambda \partial_j f(c),$$

and the two notions are consistent. We will use freely the notation $\dfrac{\partial f}{\partial x_j}(c)$ instead of $\mathrm{D}_j f(c)$.

The notion $\partial_j f(c)$ introduced in Definition 3.5 is really that of the vector derivative, whereas $\mathrm{D}_j f(c)$ $\left(= \dfrac{\partial f}{\partial x_j}(c)\right)$ is the corresponding linear map. The following proposition holds.

Proposition 3.9. *Given a normed vector space $E = E_1 \oplus \cdots \oplus E_n$, and a normed vector space F, given any open subset A of E, for any function $f\colon A \to F$, for every $c \in A$, if $\mathrm{D}f(c)$ exists, then each $\dfrac{\partial f}{\partial x_j}(c)$ exists, and*

$$\mathrm{D}f(c)(u_1, \ldots, u_n) = \frac{\partial f}{\partial x_1}(c)(u_1) + \cdots + \frac{\partial f}{\partial x_n}(c)(u_n),$$

for every $u_i \in E_i$, $1 \leq i \leq n$. The same result holds for the finite product $E_1 \times \cdots \times E_n$.

Proof. If $i_j\colon E_j \to E$ is the linear map given by

$$i_j(x) = (0, \ldots, 0, x, 0, \ldots, 0),$$

then

$$i_j^c(x) = (c_1, \ldots, c_{j-1}, 0, c_{j+1}, \ldots, c_n) + i_j(x),$$

which shows that i_j^c is affine, so $\mathrm{D}i_j^c(x) = i_j$. The proposition is then a simple application of Theorem 3.6. □

In the special case where F is a normed vector space of finite dimension m, for any basis $(v_1, \ldots, v_m)$ of F, every vector $x \in F$ can be expressed uniquely as

$$x = x_1 v_1 + \cdots + x_m v_m,$$

where $(x_1, \ldots, x_m) \in K^m$, the coordinates of x in the basis $(v_1, \ldots, v_m)$ (where $K = \mathbb{R}$ or $K = \mathbb{C}$). Thus, letting F_i be the standard normed vector space K with its natural structure, we note that F is isomorphic to the direct sum $F = K \oplus \cdots \oplus K$. Then every function $f\colon E \to F$ is represented by m functions $(f_1, \ldots, f_m)$, where $f_i\colon E \to K$ (where $K = \mathbb{R}$ or $K = \mathbb{C}$), and

$$f(x) = f_1(x)v_1 + \cdots + f_m(x)v_m,$$

for every $x \in E$. The following proposition is easily shown.

Proposition 3.10. *For any two normed vector spaces E and F, if F is of finite dimension m, for any basis $(v_1, \ldots, v_m)$ of F, a function $f\colon E \to F$ is differentiable at a iff each f_i is differentiable at a, and*

$$\mathrm{D}f(a)(u) = \mathrm{D}f_1(a)(u)v_1 + \cdots + \mathrm{D}f_m(a)(u)v_m,$$

for every $u \in E$.

3.3 Jacobian Matrices

If both E and F are of finite dimension, for any basis $(u_1, \ldots, u_n)$ of E and any basis $(v_1, \ldots, v_m)$ of F, every function $f\colon E \to F$ is determined by m functions $f_i\colon E \to \mathbb{R}$ (or $f_i\colon E \to \mathbb{C}$), where

$$f(x) = f_1(x)v_1 + \cdots + f_m(x)v_m,$$

for every $x \in E$. From Proposition 3.1, we have

$$\mathrm{D}f(a)(u_j) = \mathrm{D}_{u_j} f(a) = \partial_j f(a),$$

and from Proposition 3.10, we have

$$\mathrm{D}f(a)(u_j) = \mathrm{D}f_1(a)(u_j)v_1 + \cdots + \mathrm{D}f_i(a)(u_j)v_i + \cdots + \mathrm{D}f_m(a)(u_j)v_m,$$

that is,

$$\mathrm{D}f(a)(u_j) = \partial_j f_1(a)v_1 + \cdots + \partial_j f_i(a)v_i + \cdots + \partial_j f_m(a)v_m.$$

Since the j-th column of the $m \times n$-matrix representing $\mathrm{D}f(a)$ w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ is equal to the components of the vector $\mathrm{D}f(a)(u_j)$ over the basis $(v_1, \ldots, v_m)$, the linear map $\mathrm{D}f(a)$ is determined by the $m \times n$-matrix $J(f)(a) = (\partial_j f_i(a))$, (or $J(f)(a) = (\partial f_i/\partial x_j)(a)$):

$$J(f)(a) = \begin{pmatrix} \partial_1 f_1(a) & \partial_2 f_1(a) & \ldots & \partial_n f_1(a) \\ \partial_1 f_2(a) & \partial_2 f_2(a) & \ldots & \partial_n f_2(a) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 f_m(a) & \partial_2 f_m(a) & \ldots & \partial_n f_m(a) \end{pmatrix}$$

or

$$J(f)(a) = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1}(a) & \dfrac{\partial f_1}{\partial x_2}(a) & \ldots & \dfrac{\partial f_1}{\partial x_n}(a) \\ \dfrac{\partial f_2}{\partial x_1}(a) & \dfrac{\partial f_2}{\partial x_2}(a) & \ldots & \dfrac{\partial f_2}{\partial x_n}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1}(a) & \dfrac{\partial f_m}{\partial x_2}(a) & \ldots & \dfrac{\partial f_m}{\partial x_n}(a) \end{pmatrix}.$$

Definition 3.7. The matrix $J(f)(a)$ is called the *Jacobian matrix* of $\mathrm{D}f$ at a. When $m = n$, the determinant, $\det(J(f)(a))$, of $J(f)(a)$ is called the *Jacobian* of $\mathrm{D}f(a)$.

From a standard fact of linear algebra, we know that this determinant in fact only depends on $\mathrm{D}f(a)$, and not on specific bases. However, partial derivatives give a means for computing it.

When $E = \mathbb{R}^n$ and $F = \mathbb{R}^m$, for any function $f\colon \mathbb{R}^n \to \mathbb{R}^m$, it is easy to compute the partial derivatives $(\partial f_i/\partial x_j)(a)$. We simply treat the function $f_i\colon \mathbb{R}^n \to \mathbb{R}$ as a function of its j-th argument, leaving the others fixed, and compute the derivative as in Definition 3.1, that is, the usual derivative.

Example 3.5. For example, consider the function $f\colon \mathbb{R}^2 \to \mathbb{R}^2$, defined such that

$$f(r, \theta) = (r\cos(\theta), r\sin(\theta)).$$

Then we have

$$J(f)(r, \theta) = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix},$$

and the Jacobian (determinant) has value $\det(J(f)(r, \theta)) = r$.

In the case where $E = \mathbb{R}$ (or $E = \mathbb{C}$), for any function $f\colon \mathbb{R} \to F$ (or $f\colon \mathbb{C} \to F$), the Jacobian matrix of $\mathrm{D}f(a)$ is a column vector. In fact, this column vector is just $\mathrm{D}_1 f(a)$. Then for every $\lambda \in \mathbb{R}$ (or $\lambda \in \mathbb{C}$),

$$\mathrm{D}f(a)(\lambda) = \lambda \mathrm{D}_1 f(a).$$

This case is sufficiently important to warrant a definition.

Definition 3.8. Given a function $f\colon \mathbb{R} \to F$ (or $f\colon \mathbb{C} \to F$), where F is a normed vector space, the vector

$$\mathrm{D}f(a)(1) = \mathrm{D}_1 f(a)$$

is called the *vector derivative or velocity vector (in the real case)* at a. We usually identify $\mathrm{D}f(a)$ with its Jacobian matrix $\mathrm{D}_1 f(a)$, which is the column vector corresponding to $\mathrm{D}_1 f(a)$. By abuse of notation, we also let $\mathrm{D}f(a)$ denote the vector $\mathrm{D}f(a)(1) = \mathrm{D}_1 f(a)$.

When $E = \mathbb{R}$, the physical interpretation is that f defines a (parametric) curve that is the trajectory of some particle moving in $\mathbb{R}^m$ as a function of time, and the vector $\mathrm{D}_1 f(a)$ is the *velocity* of the moving particle $f(t)$ at $t = a$; see Figure 3.4.

It is often useful to consider functions $f\colon [a, b] \to F$ from a closed interval $[a, b] \subseteq \mathbb{R}$ to a normed vector space F, and its derivative $\mathrm{D}f(a)$ on $[a, b]$, even though $[a, b]$ is not open. In this case, as in the case of a real-valued function, we define the right derivative $\mathrm{D}_1 f(a_+)$ at a, and the left derivative $\mathrm{D}_1 f(b_-)$ at b, and we assume their existence.

Example 3.6.

1. When $A = (0,1)$ and $F = \mathbb{R}^3$, a function $f\colon (0,1) \to \mathbb{R}^3$ defines a (parametric) curve in $\mathbb{R}^3$. If $f = (f_1, f_2, f_3)$, its Jacobian matrix at $a \in \mathbb{R}$ is

$$J(f)(a) = \begin{pmatrix} \dfrac{\partial f_1}{\partial t}(a) \\ \dfrac{\partial f_2}{\partial t}(a) \\ \dfrac{\partial f_3}{\partial t}(a) \end{pmatrix}.$$

See Figure 3.4.

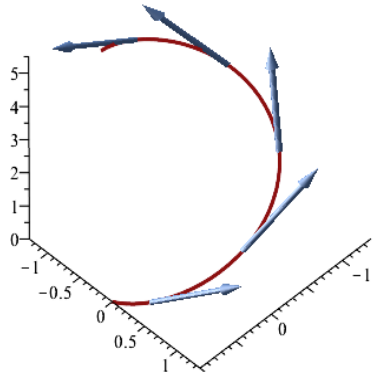


Figure 3.4: The red space curve $f(t) = (\cos(t), \sin(t), t)$.

The velocity vectors $J(f)(a) = \begin{pmatrix} -\sin(t) \\ \cos(t) \\ 1 \end{pmatrix}$ are represented by the blue arrows.

2. When $E = \mathbb{R}^2$ and $F = \mathbb{R}^3$, a function $\varphi\colon \mathbb{R}^2 \to \mathbb{R}^3$ defines a parametric surface. Letting $\varphi = (f, g, h)$, its Jacobian matrix at $a \in \mathbb{R}^2$ is

$$J(\varphi)(a) = \begin{pmatrix} \dfrac{\partial f}{\partial u}(a) & \dfrac{\partial f}{\partial v}(a) \\ \dfrac{\partial g}{\partial u}(a) & \dfrac{\partial g}{\partial v}(a) \\ \dfrac{\partial h}{\partial u}(a) & \dfrac{\partial h}{\partial v}(a) \end{pmatrix}.$$

See Figure 3.5. The Jacobian matrix is $J(f)(a) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2u & 2v \end{pmatrix}$. The first column is the

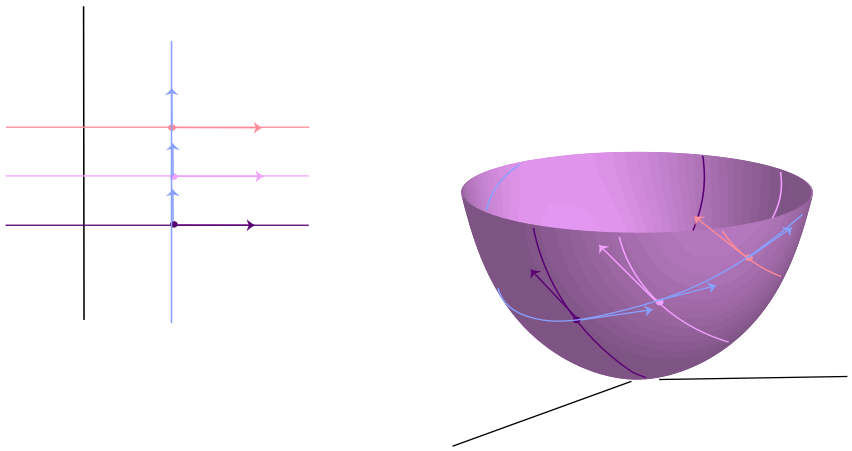


Figure 3.5: The parametric surface $x = u, y = v, z = u^2 + v^2$.

vector tangent to the pink u-direction curve, while the second column is the vector tangent to the blue v-direction curve.

3. When $E = \mathbb{R}^3$ and $F = \mathbb{R}$, for a function $f\colon \mathbb{R}^3 \to \mathbb{R}$, the Jacobian matrix at $a \in \mathbb{R}^3$ is

$$J(f)(a) = \left(\frac{\partial f}{\partial x}(a) \;\; \frac{\partial f}{\partial y}(a) \;\; \frac{\partial f}{\partial z}(a)\right).$$

Definition 3.9. More generally, when $f\colon \mathbb{R}^n \to \mathbb{R}$, the Jacobian matrix at $a \in \mathbb{R}^n$ is the row vector

$$J(f)(a) = \left(\frac{\partial f}{\partial x_1}(a) \;\cdots\; \frac{\partial f}{\partial x_n}(a)\right).$$

Its transpose is a column vector called the *gradient* of f at a, denoted by $\mathrm{grad} f(a)$ or $\nabla f(a)$. Then given any $v \in \mathbb{R}^n$, note that

$$\mathrm{D}f(a)(v) = \frac{\partial f}{\partial x_1}(a)\, v_1 + \cdots + \frac{\partial f}{\partial x_n}(a)\, v_n = \mathrm{grad} f(a) \cdot v,$$

the scalar product of $\mathrm{grad} f(a)$ and v.

Example 3.7. Consider the function $f\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ from Example 3.4 given by

$$f(v, \lambda) = \lambda^\top C v = v^\top C^\top \lambda, \quad v \in \mathbb{R}^n, \lambda \in \mathbb{R}^m,$$

where C is an $m \times n$ matrix. We showed that its derivative is given by

$$df_{(v,\lambda)}(w, \mu) = (C^\top \lambda)^\top w + (Cv)^\top \mu = \begin{pmatrix} C^\top \lambda & Cv \end{pmatrix}^\top \begin{pmatrix} w \\ \mu \end{pmatrix},$$

so the gradient of f at (v, λ) is

$$\nabla f_{(v,\lambda)} = \begin{pmatrix} C^\top \lambda \\ Cv \end{pmatrix}.$$

Example 3.8. Consider the quadratic function $f\colon \mathbb{R}^n \to \mathbb{R}$ given by

$$f(x) = x^\top Ax, \quad x \in \mathbb{R}^n,$$

where A is a real $n \times n$ symmetric matrix. We claim that

$$df_u(h) = 2u^\top Ah \quad \text{for all } u, h \in \mathbb{R}^n.$$

Since A is symmetric, we have

$$\begin{aligned} f(u+h) &= (u^\top + h^\top)A(u+h) \\ &= u^\top Au + u^\top Ah + h^\top Au + h^\top Ah \\ &= u^\top Au + 2u^\top Ah + h^\top Ah, \end{aligned}$$

so we have

$$f(u+h) - f(u) - 2u^\top Ah = h^\top Ah.$$

If we write

$$\epsilon(h) = \frac{h^\top Ah}{\|h\|}$$

for $h \notin 0$ where $\| \|$ is the 2-norm, by Cauchy–Schwarz we have

$$|\epsilon(h)| \leq \frac{\|h\| \, \|Ah\|}{\|h\|} \leq \frac{\|h\|^2 \, \|A\|}{\|h\|} = \|h\| \, \|A\|,$$

which shows that $\lim_{h \mapsto 0} \epsilon(h) = 0$. Therefore,

$$df_u(h) = 2u^\top Ah \quad \text{for all } u, h \in \mathbb{R}^n,$$

as claimed. This formula shows that the gradient ∇f_u of f at u is given by

$$\nabla f_u = 2Au.$$

As a first corollary we obtain the gradient of a function of the form

$$f(x) = \frac{1}{2} x^\top Ax - b^\top x,$$

where A is a symmetric $n \times n$ matrix and b is some vector $b \in \mathbb{R}^n$. Since the derivative of a linear function is itself, we obtain

$$df_u(h) = u^\top Ah - b^\top h,$$

and the gradient of $f(x) = \frac{1}{2}x^\top Ax - b^\top x$, is given by

$$\nabla f_u = Au - b.$$

As a second corollary we obtain the gradient of the function

$$f(x) = \|Ax - b\|_2^2 = (Ax - b)^\top (Ax - b) = (x^\top A^\top - b^\top)(Ax - b)$$

which is the function to minimize in a least squares problem, where A is an $m \times n$ matrix. We have

$$f(x) = x^\top A^\top Ax - x^\top A^\top b - b^\top Ax + b^\top b = x^\top A^\top Ax - 2b^\top Ax + b^\top b,$$

and since the derivative of a constant function is 0 and the derivative of a linear function is itself, we get

$$df_u(h) = 2u^\top A^\top Ah - 2b^\top Ah.$$

Consequently, the gradient of $f(x) = \|Ax - b\|_2^2$ is given by

$$\nabla f_u = 2A^\top Au - 2A^\top b.$$

These two results will be heavily used in quadratic optimization.

Example 3.9. In Example 4.3 from Section 4.1, we need to find the gradient of the function

$$L(v, \lambda) = \frac{1}{2}v^\top Av - v^\top b + v^\top C^\top \lambda - d^\top \lambda,$$

where A is an $n \times n$ symmetric matrix and C is an $m \times n$ matrix of rank m. From Example 3.7 and Example 3.8, we have

$$\nabla L_{((v,\lambda)} = \begin{pmatrix} Av - b \\ 0 \end{pmatrix} + \begin{pmatrix} C^\top \lambda \\ Cv \end{pmatrix} + \begin{pmatrix} 0 \\ -d \end{pmatrix} = \begin{pmatrix} Av - b + C^\top \lambda \\ Cv - d \end{pmatrix}.$$

When E, F, and G have finite dimensions, and $(u_1, \ldots, u_p)$ is a basis for E, $(v_1, \ldots, v_n)$ is a basis for F, and $(w_1, \ldots, w_m)$ is a basis for G, if A is an open subset of E, B is an open subset of F, for any functions $f\colon A \to F$ and $g\colon B \to G$, such that $f(A) \subseteq B$, for any $a \in A$, letting $b = f(a)$, and $h = g \circ f$, if $\mathrm{D}f(a)$ exists and $\mathrm{D}g(b)$ exists, by Theorem 3.6, the Jacobian matrix $J(h)(a) = J(g \circ f)(a)$ w.r.t. the bases $(u_1, \ldots, u_p)$ and $(w_1, \ldots, w_m)$ is the product of the Jacobian matrices $J(g)(b)$ w.r.t. the bases $(v_1, \ldots, v_n)$ and $(w_1, \ldots, w_m)$, and $J(f)(a)$ w.r.t. the bases $(u_1, \ldots, u_p)$ and $(v_1, \ldots, v_n)$:

$$J(h)(a) = \begin{pmatrix} \partial_1 g_1(b) & \partial_2 g_1(b) & \ldots & \partial_n g_1(b) \\ \partial_1 g_2(b) & \partial_2 g_2(b) & \ldots & \partial_n g_2(b) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 g_m(b) & \partial_2 g_m(b) & \ldots & \partial_n g_m(b) \end{pmatrix} \begin{pmatrix} \partial_1 f_1(a) & \partial_2 f_1(a) & \ldots & \partial_p f_1(a) \\ \partial_1 f_2(a) & \partial_2 f_2(a) & \ldots & \partial_p f_2(a) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 f_n(a) & \partial_2 f_n(a) & \ldots & \partial_p f_n(a) \end{pmatrix}$$

or

$$J(h)(a) = \begin{pmatrix} \dfrac{\partial g_1}{\partial y_1}(b) & \dfrac{\partial g_1}{\partial y_2}(b) & \dots & \dfrac{\partial g_1}{\partial y_n}(b) \\ \dfrac{\partial g_2}{\partial y_1}(b) & \dfrac{\partial g_2}{\partial y_2}(b) & \dots & \dfrac{\partial g_2}{\partial y_n}(b) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial g_m}{\partial y_1}(b) & \dfrac{\partial g_m}{\partial y_2}(b) & \dots & \dfrac{\partial g_m}{\partial y_n}(b) \end{pmatrix} \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1}(a) & \dfrac{\partial f_1}{\partial x_2}(a) & \dots & \dfrac{\partial f_1}{\partial x_p}(a) \\ \dfrac{\partial f_2}{\partial x_1}(a) & \dfrac{\partial f_2}{\partial x_2}(a) & \dots & \dfrac{\partial f_2}{\partial x_p}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_n}{\partial x_1}(a) & \dfrac{\partial f_n}{\partial x_2}(a) & \dots & \dfrac{\partial f_n}{\partial x_p}(a) \end{pmatrix}.$$

Thus, we have the familiar formula

$$\frac{\partial h_i}{\partial x_j}(a) = \sum_{k=1}^{k=n} \frac{\partial g_i}{\partial y_k}(b) \frac{\partial f_k}{\partial x_j}(a).$$

Given two normed vector spaces E and F of finite dimension, given an open subset A of E, if a function $f\colon A \to F$ is differentiable at $a \in A$, then its Jacobian matrix is well defined.

One should be warned that the converse is false. As evidenced by Example 3.1, there are functions such that all the partial derivatives exist at some $a \in A$, but yet, the function is not differentiable at a, and not even continuous at a. However, there are sufficient conditions on the partial derivatives for $\mathrm{D}f(a)$ to exist, namely, continuity of the partial derivatives.

If f is differentiable on A, then f defines a function $\mathrm{D}f\colon A \to \mathcal{L}(E; F)$. It turns out that the continuity of the partial derivatives on A is a necessary and sufficient condition for $\mathrm{D}f$ to exist and to be continuous on A.

If $f\colon [a, b] \to \mathbb{R}$ is a function which is continuous on $[a, b]$ and differentiable on (a, b), then there is some c with $a < c < b$ such that

$$f(b) - f(a) = (b - a)f'(c).$$

This result is known as the *mean value theorem* and is a generalization of *Rolle's theorem*, which corresponds to the case where $f(a) = f(b)$.

Unfortunately, the mean value theorem fails for vector-valued functions. For example, the function $f\colon [0, 2\pi] \to \mathbb{R}^2$ given by

$$f(t) = (\cos t, \sin t)$$

is such that $f(2\pi) - f(0) = (0, 0)$, yet its derivative $f'(t) = (-\sin t, \cos t)$ does not vanish in $(0, 2\pi)$.

A suitable generalization of the mean value theorem to vector-valued functions is possible if we consider an inequality (an upper bound) instead of an equality. This generalized version

of the mean value theorem plays an important role in the proof of several major results of differential calculus.

If E is a vector space (over $\mathbb{R}$ or $\mathbb{C}$), given any two points $a, b \in E$, the *closed segment* $[a, b]$ is the set of all points $a + \lambda(b - a)$, where $0 \leq \lambda \leq 1$, $\lambda \in \mathbb{R}$, and the *open segment* (a, b) is the set of all points $a + \lambda(b - a)$, where $0 < \lambda < 1$, $\lambda \in \mathbb{R}$.

Proposition 3.11. *Let E and F be two normed vector spaces, let A be an open subset of E, and let $f\colon A \to F$ be a continuous function on A. Given any $a \in A$ and any $h \neq 0$ in E, if the closed segment $[a, a + h]$ is contained in A, if $f\colon A \to F$ is differentiable at every point of the open segment $(a, a + h)$, and*

$$\sup_{x \in (a, a+h)} \|\mathrm{D}f(x)\| \leq M,$$

for some $M \geq 0$, then

$$\|f(a + h) - f(a)\| \leq M\|h\|.$$

As a corollary, if $L\colon E \to F$ is a continuous linear map, then

$$\|f(a + h) - f(a) - L(h)\| \leq M\|h\|,$$

where $M = \sup_{x \in (a, a+h)} \|\mathrm{D}f(x) - L\|$.

The above proposition is sometimes called the "mean value theorem." Propostion 3.11 can be used to show the following important result.

Theorem 3.12. *Given two normed vector spaces E and F, where E is of finite dimension n, and where $(u_1, \ldots, u_n)$ is a basis of E, given any open subset A of E, given any function $f\colon A \to F$, the derivative $\mathrm{D}f\colon A \to \mathcal{L}(E; F)$ is defined and continuous on A iff every partial derivative $\partial_j f$ (or $\dfrac{\partial f}{\partial x_j}$) is defined and continuous on A, for all j, $1 \leq j \leq n$. As a corollary, if F is of finite dimension m, and $(v_1, \ldots, v_m)$ is a basis of F, the derivative $\mathrm{D}f\colon A \to \mathcal{L}(E; F)$ is defined and continuous on A iff every partial derivative $\partial_j f_i$ (or $\dfrac{\partial f_i}{\partial x_j}$) is defined and continuous on A, for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$.*

Theorem 3.12 gives a necessary and sufficient condition for the existence and continuity of the derivative of a function on an open set. It should be noted that a more general version of Theorem 3.12 holds, assuming that $E = E_1 \oplus \cdots \oplus E_n$, or $E = E_1 \times \cdots \times E_n$, and using the more general partial derivatives $\mathrm{D}_j f$ introduced before Proposition 3.9.

Definition 3.10. Given two normed vector spaces E and F, and an open subset A of E, we say that a function $f\colon A \to F$ is *of class C^0 on A or a C^0-function on A* if f is continuous on A. We say that $f\colon A \to F$ is *of class C^1 on A or a C^1-function on A* if $\mathrm{D}f$ exists and is continuous on A.

Since the existence of the derivative on an open set implies continuity, a C^1-function is of course a C^0-function. Theorem 3.12 gives a necessary and sufficient condition for a function f to be a C^1-function (when E is of finite dimension). It is easy to show that the composition of C^1-functions (on appropriate open sets) is a C^1-function.

3.4 The Implicit and The Inverse Function Theorems

Given three normed vector spaces E, F, and G, given a function $f\colon E \times F \to G$, given any $c \in G$, it may happen that the equation

$$f(x, y) = c$$

has the property that for some open sets $A \subseteq E$ and $B \subseteq F$, there is a function $g\colon A \to B$, such that

$$f(x, g(x)) = c,$$

for all $x \in A$. Such a situation is usually very rare, but if some solution $(a, b) \in E \times F$ such that $f(a, b) = c$ is known, under certain conditions, for some small open sets $A \subseteq E$ containing a and $B \subseteq F$ containing b, the existence of a unique $g\colon A \to B$ such that

$$f(x, g(x)) = c,$$

for all $x \in A$, can be shown. Under certain conditions, it can also be shown that g is continuous and differentiable. Such a theorem, known as the *implicit function theorem*, can be proven.

Example 3.10. Let $E = \mathbb{R}^2$, $F = G = \mathbb{R}$, $\Omega = \mathbb{R}^2 \times \mathbb{R} \cong \mathbb{R}^3$, $f\colon \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ given by

$$f((x_1, x_2), x_3) = x_1^2 + x_2^2 + x_3^2 - 1,$$

$a = \big(\sqrt{3}/(2\sqrt{2}), \sqrt{3}/(2\sqrt{2})\big)$, $b = 1/2$, and $c = 0$. The set of vectors $(x_1, x_2, x_3) \in \mathbb{R}^2$ such that

$$f((x_1, x_2), x_3) = x_1^2 + x_2^2 + x_3^2 - 1 = 0$$

is the unit sphere in $\mathbb{R}^3$. The vector (a, b) belongs to the unit sphere since $\|a\|_2^2 + b^2 - 1 = 0$. The function $g\colon \mathbb{R}^2 \to \mathbb{R}$ given by

$$g(x_1, x_2) = \sqrt{1 - x_1^2 - x_2^2}$$

satisfies the equation

$$f(x_1, x_2, g(x_1, x_2)) = 0$$

all for (x_1, x_2) in the open disk $\{(x_1, x_2) \in \mathbb{R}^2 \mid x_1^2 + x_2^2 < 1\}$, and $g(a) = b$. Observe that if we had picked $b = -1/2$, then we would need to consider the function

$$g(x_1, x_2) = -\sqrt{1 - x_1^2 - x_2^2}.$$

We now state a very general version of the implicit function theorem. The proof of this theorem is fairly involved and uses a fixed-point theorem for contracting mappings in complete metric spaces; it is given in Schwartz [70].

Theorem 3.13. *Let E, F, and G be normed vector spaces, let Ω be an open subset of $E \times F$, let $f\colon \Omega \to G$ be a function defined on Ω, let $(a, b) \in \Omega$, let $c \in G$, and assume that $f(a, b) = c$. If the following assumptions hold:*

(1) The function $f\colon \Omega \to G$ is continuous on Ω;

(2) F is a complete normed vector space;

(3) $\dfrac{\partial f}{\partial y}(x, y)$ exists for every $(x, y) \in \Omega$ and $\dfrac{\partial f}{\partial y}\colon \Omega \to \mathcal{L}(F; G)$ is continuous, where $\dfrac{\partial f}{\partial y}(x, y)$ is defined as in Definition 3.6;

(4) $\dfrac{\partial f}{\partial y}(a, b)$ is a bijection of $\mathcal{L}(F; G)$, and $\Big(\dfrac{\partial f}{\partial y}(a, b)\Big)^{-1} \in \mathcal{L}(G; F)$; this hypothesis implies that G is also a complete normed vector space;

then the following properties hold:

(a) There exist some open subset $A \subseteq E$ containing a and some open subset $B \subseteq F$ containing b, such that $A \times B \subseteq \Omega$, and for every $x \in A$, the equation $f(x, y) = c$ has a single solution $y = g(x)$, and thus there is a unique function $g\colon A \to B$ such that $f(x, g(x)) = c$, for all $x \in A$;

(b) The function $g\colon A \to B$ is continuous.

If we also assume that

(5) The derivative $\mathrm{D}f(a, b)$ exists;

then

(c) The derivative $\mathrm{D}g(a)$ exists, and

$$\mathrm{D}g(a) = -\Big(\frac{\partial f}{\partial y}(a, b)\Big)^{-1} \circ \frac{\partial f}{\partial x}(a, b);$$

and if in addition

(6) $\dfrac{\partial f}{\partial x}\colon \Omega \to \mathcal{L}(E; G)$ is also continuous (and thus, in view of (3), f is C^1 on Ω);

then

(d) The derivative $\mathrm{D}g\colon A \to \mathcal{L}(E;F)$ *is continuous, and*

$$\mathrm{D}g(x) = -\Big(\frac{\partial f}{\partial y}(x, g(x))\Big)^{-1} \circ \frac{\partial f}{\partial x}(x, g(x)),$$

for all $x \in A$.

Example 3.11. Going back to Example 3.10, write $x = (x_1, x_2)$ and $y = x_3$, so that the partial derivatives $\partial f/\partial x$ and $\partial f/\partial y$ are given in terms of their Jacobian matrices by

$$\frac{\partial f}{\partial x}(x, y) = \begin{pmatrix} 2x_1 & 2x_2 \end{pmatrix}$$
$$\frac{\partial f}{\partial y}(x, y) = 2x_3.$$

If $0 < |b| \leq 1$ and $\|a\|_2^2 + b^2 - 1 = 0$, then Conditions (3) and (4) are satisfied. Conditions (1) and (2) obviously hold. Since $df_{(a,b)}$ is given by its Jacobian matrix as

$$df_{(a,b)} = \begin{pmatrix} 2a_1 & 2a_2 & 2b \end{pmatrix},$$

Condition (5) holds, and clearly, Condition (6) also holds.

Theorem 3.13 implies that there is some open subset A of $\mathbb{R}^2$ containing a, some open subset B of $\mathbb{R}$ containing b, and a unique function $g\colon A \to B$ such that

$$f(x, g(x)) = 0$$

for all $x \in A$. In fact, we can pick A to be the open unit disk in $\mathbb{R}$, $B = (0, 2)$, and if $0 < b \leq 1$, then

$$g(x_1, x_2) = \sqrt{1 - x_1^2 - x_2^2},$$

else if $-1 \leq b < 0$, then

$$g(x_1, x_2) = -\sqrt{1 - x_1^2 - x_2^2}.$$

Assuming $0 < b \leq 1$, We have

$$\frac{\partial f}{\partial x}(x, g(x)) = (2x_1 \;\; 2x_2),$$

and

$$\left(\frac{\partial f}{\partial y}(x, g(x))\right)^{-1} = \frac{1}{2\sqrt{1 - x_1^2 - x_2^2}},$$

so according to the theorem,

$$dg_x = -\frac{1}{\sqrt{1 - x_1^2 - x_2^2}}(x_1 \;\; x_2),$$

which matches the derivative of g computed directly.

Observe that the functions $(x_1, x_2) \mapsto \sqrt{1 - x_1^2 - x_2^2}$ and $(x_1, x_2) \mapsto -\sqrt{1 - x_1^2 - x_2^2}$ are two differentiable parametrizations of the sphere, but the union of their ranges does not cover the entire sphere. Since $b \neq 0$, none of the points on the unit circle in the (x_1, x_2)-plane are covered. Our function f views b as lying on the x_3-axis. In order to cover the entire sphere using this method, we need four more maps, which correspond to b lying on the x_1-axis or on the x_2 axis. Then we get the additional (implicit) maps $(x_2, x_3) \mapsto \pm\sqrt{1 - x_2^2 - x_3^2}$ and $(x_1, x_3) \mapsto \pm\sqrt{1 - x_1^2 - x_3^2}$.

The implicit function theorem plays an important role in the calculus of variations.

We now consider another very important notion, that of a (local) diffeomorphism.

Definition 3.11. Given two topological spaces E and F and an open subset A of E, we say that a function $f\colon A \to F$ is a *local homeomorphism from A to F* if for every $a \in A$, there is an open set $U \subseteq A$ containing a and an open set V containing $f(a)$ such that f is a homeomorphism from U to $V = f(U)$. If B is an open subset of F, we say that $f\colon A \to F$ is a *(global) homeomorphism from A to B* if f is a homeomorphism from A to $B = f(A)$. If E and F are normed vector spaces, we say that $f\colon A \to F$ is a *local diffeomorphism from A to F* if for every $a \in A$, there is an open set $U \subseteq A$ containing a and an open set V containing $f(a)$ such that f is a bijection from U to V, f is a C^1-function on U, and f^{-1} is a C^1-function on $V = f(U)$. We say that $f\colon A \to F$ is a *(global) diffeomorphism from A to B* if f is a homeomorphism from A to $B = f(A)$, f is a C^1-function on A, and f^{-1} is a C^1-function on B.

Note that a local diffeomorphism is a local homeomorphism. Also, as a consequence of Proposition 3.8, if f is a diffeomorphism on A, then $\mathrm{D}f(a)$ is a bijection for every $a \in A$. The following theorem can be shown. In fact, there is a fairly simple proof using Theorem 3.13.

Theorem 3.14. *(Inverse Function Theorem) Let E and F be complete normed spaces, let A be an open subset of E, and let $f\colon A \to F$ be a C^1-function on A. The following properties hold:*

(1) For every $a \in A$, if $\mathrm{D}f(a)$ is a linear isomorphism (which means that both $\mathrm{D}f(a)$ and $(\mathrm{D}f(a))^{-1}$ are linear and continuous),[2] then there exist some open subset $U \subseteq A$ containing a, and some open subset V of F containing $f(a)$, such that f is a diffeomorphism from U to $V = f(U)$. Furthermore,

$$\mathrm{D}f^{-1}(f(a)) = (\mathrm{D}f(a))^{-1}.$$

For every neighborhood N of a, the image $f(N)$ of N is a neighborhood of $f(a)$, and for every open ball $U \subseteq A$ of center a, the image $f(U)$ of U contains some open ball of center $f(a)$.

[2] Actually, since E and F are Banach spaces, by the open mapping theorem, it is sufficient to assume that $\mathrm{D}f(a)$ is continuous and bijective; see Lang [49].

(2) *If* $\mathrm{D}f(a)$ *is invertible for every* $a \in A$*, then* $B = f(A)$ *is an open subset of* F*, and* f *is a local diffeomorphism from* A *to* B*. Furthermore, if* f *is injective, then* f *is a diffeomorphism from* A *to* B*.*

Proofs of the inverse function theorem can be found in Schwartz [70], Lang [49], Abraham and Marsden [1], and Cartan [21].

The idea of Schwartz's proof is that if we define the function $f_1 \colon F \times \Omega \to F$ by

$$f_1(y, z) = f(z) - y,$$

then an inverse $g = f^{-1}$ of f is an implicit solution of the equation $f_1(y, z) = 0$, since $f_1(y, g(y)) = f(g(y)) - y = 0$. Observe that the roles of E and F are switched, but this is not a problem since F is complete. The proof consists in checking that the conditions of Theorem 3.13 apply.

Part (1) of Theorem 3.14 is often referred to as the "(local) inverse function theorem." It plays an important role in the study of manifolds and (ordinary) differential equations.

If E and F are both of finite dimension, and some bases have been chosen, the invertibility of $\mathrm{D}f(a)$ is equivalent to the fact that the Jacobian determinant $\det(J(f)(a))$ is nonnull. The case where $\mathrm{D}f(a)$ is just injective or just surjective is also important for defining manifolds, using implicit definitions.

Definition 3.12. Let E and F be normed vector spaces, where E and F are of finite dimension (or both E and F are complete), and let A be an open subset of E. For any $a \in A$, a C^1-function $f \colon A \to F$ is an *immersion at* a if $\mathrm{D}f(a)$ is injective. A C^1-function $f \colon A \to F$ is a *submersion at* a if $\mathrm{D}f(a)$ is surjective. A C^1-function $f \colon A \to F$ is an *immersion on* A (resp. *a submersion on* A) if $\mathrm{D}f(a)$ is injective (resp. surjective) for every $a \in A$.

When E and F are finite dimensional with $\dim(E) = n$ and $\dim(F) = m$, if $m \geq n$, then f is an immersion iff the Jacobian matrix, $J(f)(a)$, has full rank n for all $a \in E$, and if $n \geq m$, then f is a submersion iff the Jacobian matrix, $J(f)(a)$, has full rank m for all $a \in E$.

Example 3.12. For example, $f \colon \mathbb{R} \to \mathbb{R}^2$ defined by $f(t) = (\cos(t), \sin(t))$ is an immersion since $J(f)(t) = \begin{pmatrix} -\sin(t) \\ \cos(t) \end{pmatrix}$ has rank 1 for all t. On the other hand, $f \colon \mathbb{R} \to \mathbb{R}^2$ defined by $f(t) = (t^2, t^2)$ is not an immersion since $J(f)(t) = \begin{pmatrix} 2t \\ 2t \end{pmatrix}$ vanishes at $t = 0$. See Figure 3.6. An example of a submersion is given by the projection map $f \colon \mathbb{R}^2 \to \mathbb{R}$, where $f(x, y) = x$, since $J(f)(x, y) = \begin{pmatrix} 1 & 0 \end{pmatrix}$.

The following results can be shown.

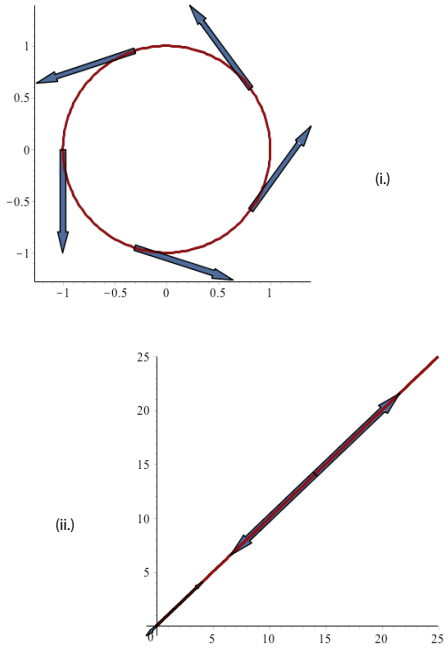


Figure 3.6: Figure $(i.)$ is the immersion of $\mathbb{R}$ into $\mathbb{R}^2$ given by $f(t) = (\cos(t), \sin(t))$. Figure $(ii.)$, the parametric curve $f(t) = (t^2, t^2)$, is not an immersion since the tangent vanishes at the origin.

Proposition 3.15. *Let A be an open subset of $\mathbb{R}^n$, and let $f\colon A \to \mathbb{R}^m$ be a function. For every $a \in A$, $f\colon A \to \mathbb{R}^m$ is a submersion at a iff there exists an open subset U of A containing a, an open subset $W \subseteq \mathbb{R}^{n-m}$, and a diffeomorphism $\varphi\colon U \to f(U) \times W$, such that,*

$$f = \pi_1 \circ \varphi,$$

where $\pi_1\colon f(U) \times W \to f(U)$ is the first projection. Equivalently,

$$(f \circ \varphi^{-1})(y_1, \ldots, y_m, \ldots, y_n) = (y_1, \ldots, y_m).$$

$$\begin{array}{ccc} U \subseteq A & \xrightarrow{\varphi} & f(U) \times W \\ & \searrow^{f} & \downarrow{\scriptstyle \pi_1} \\ & & f(U) \subseteq \mathbb{R}^m \end{array}$$

Furthermore, the image of every open subset of A under f is an open subset of F. (The same result holds for $\mathbb{C}^n$ and $\mathbb{C}^m$). See Figure 3.7.

Proposition 3.16. *Let A be an open subset of $\mathbb{R}^n$, and let $f\colon A \to \mathbb{R}^m$ be a function. For every $a \in A$, $f\colon A \to \mathbb{R}^m$ is an immersion at a iff there exists an open subset U of A*

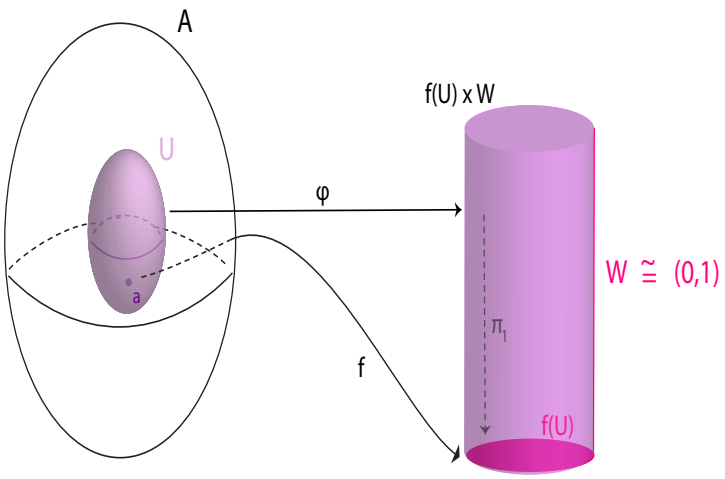


Figure 3.7: Let $n = 3$ and $m = 2$. The submersion maps the solid lavender egg in $\mathbb{R}^3$ onto the bottom pink circular face of the solid cylinder $f(U) \times W$.

containing a, an open subset V containing $f(a)$ such that $f(U) \subseteq V$, an open subset W containing 0 such that $W \subseteq \mathbb{R}^{m-n}$, and a diffeomorphism $\varphi\colon V \to U \times W$, such that,

$$\varphi \circ f = in_1,$$

where $in_1\colon U \to U \times W$ is the injection map such that $in_1(u) = (u, 0)$, or equivalently,

$$(\varphi \circ f)(x_1, \ldots, x_n) = (x_1, \ldots, x_n, 0, \ldots, 0).$$

$$\begin{array}{ccc} U \subseteq A & \xrightarrow{f} & f(U) \subseteq V \\ & \searrow^{in_1} & \downarrow \varphi \\ & & U \times W \end{array}$$

(The same result holds for $\mathbb{C}^n$ and $\mathbb{C}^m$). See Figure 3.8.

We now briefly consider second-order and higher-order derivatives.

3.5 Second-Order and Higher-Order Derivatives

Given two normed vector spaces E and F, and some open subset A of E, if $\mathrm{D}f(a)$ is defined for every $a \in A$, then we have a mapping $\mathrm{D}f\colon A \to \mathcal{L}(E; F)$. Since $\mathcal{L}(E; F)$ is a normed vector space, if $\mathrm{D}f$ exists on an open subset U of A containing a, we can consider taking the derivative of $\mathrm{D}f$ at some $a \in A$.

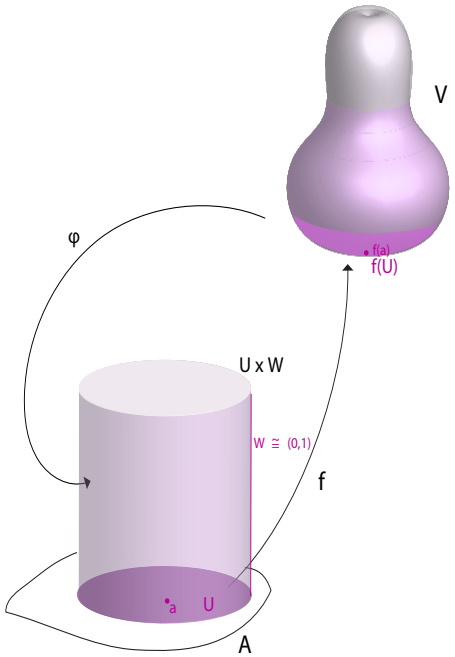


Figure 3.8: Let $n = 2$ and $m = 3$. The immersion maps the purple circular base of the cylinder $U \times W$ to circular cup on the surface of the solid purple gourd.

Definition 3.13. Given a function $f\colon A \to F$ defined on some open subset A of E such that $\mathrm{D}f(a)$ is defined for every $a \in A$, if $\mathrm{D}(\mathrm{D}f)(a)$ exists for every $a \in A$, we get a mapping $\mathrm{D}^2 f\colon A \to \mathcal{L}(E; \mathcal{L}(E; F))$ called the *second derivative of f on A*, where $\mathrm{D}^2 f(a) = \mathrm{D}(\mathrm{D}f)(a)$, for every $a \in A$.

As in the case of the first derivative $\mathrm{D}f_a$ where $\mathrm{D}f_a(u) = \mathrm{D}_u f(a)$, where $\mathrm{D}_u f(a)$ is the directional derivative of f at a in the direction u, it would be useful to express $\mathrm{D}^2 f(a)(u)(v)$ in terms of two directional derivatives. This can indeed be done. If $\mathrm{D}^2 f(a)$ exists, then for every $u \in E$,

$$\mathrm{D}^2 f(a)(u) = \mathrm{D}(\mathrm{D}f)(a)(u) = \mathrm{D}_u(\mathrm{D}f)(a) \in \mathcal{L}(E; F).$$

We have the following result.

Proposition 3.17. *If $\mathrm{D}^2 f(a)$ exists, then $\mathrm{D}_u(\mathrm{D}_v f)(a)$ exists and*

$$\mathrm{D}^2 f(a)(u)(v) = \mathrm{D}_u(\mathrm{D}_v f)(a), \quad \textit{for all } u, v \in E.$$

Proof. Recall from Proposition 2.29, that the map app from $\mathcal{L}(E; F) \times E$ to F, defined such that for every $L \in \mathcal{L}(E; F)$, for every $v \in E$,

$$\mathrm{app}(L, v) = L(v),$$

is a continuous bilinear map. Thus, in particular, given a fixed $v \in E$, the linear map $\mathrm{app}_v \colon \mathcal{L}(E;F) \to F$, defined such that $\mathrm{app}_v(L) = L(v)$, is a continuous map.

Also recall from Proposition 3.7, that if $h \colon A \to G$ is a function such that $\mathrm{D}h(a)$ exits, and $k \colon G \to H$ is a continuous linear map, then, $\mathrm{D}(k \circ h)(a)$ exists, and

$$k(\mathrm{D}h(a)(u)) = \mathrm{D}(k \circ h)(a)(u),$$

that is,

$$k(\mathrm{D}_u h(a)) = \mathrm{D}_u(k \circ h)(a),$$

Applying these two facts to $h = \mathrm{D}f$, and to $k = \mathrm{app}_v$, we have

$$\mathrm{app}_v(D_u(Df)(a)) = \mathrm{D}_u(\mathrm{D}f)(a)(v) = \mathrm{D}_u(\mathrm{app}_v \circ \mathrm{D}f)(a).$$

But $(\mathrm{app}_v \circ \mathrm{D}f)(x) = \mathrm{D}f(x)(v) = \mathrm{D}_v f(x)$, for every $x \in A$, that is, $\mathrm{app}_v \circ \mathrm{D}f = \mathrm{D}_v f$ on A. So we have

$$\mathrm{D}_u(\mathrm{D}f)(a)(v) = \mathrm{D}_u(\mathrm{D}_v f)(a),$$

and since $\mathrm{D}^2 f(a)(u) = \mathrm{D}_u(\mathrm{D}f)(a)$, we get

$$\mathrm{D}^2 f(a)(u)(v) = \mathrm{D}_u(\mathrm{D}_v f)(a). \qquad \square$$

Definition 3.14. We denote $\mathrm{D}_u(\mathrm{D}_v f)(a)$ by $\mathrm{D}^2_{u,v} f(a)$ (or $\mathrm{D}_u \mathrm{D}_v f(a)$).

Recall from Proposition 2.28, that the map from $\mathcal{L}_2(E, E; F)$ to $\mathcal{L}(E; \mathcal{L}(E; F))$ defined such that $g \mapsto \varphi$ iff for every $g \in \mathcal{L}_2(E, E; F)$,

$$\varphi(u)(v) = g(u, v),$$

is an isomorphism of vector spaces. *Thus, we will consider* $\mathrm{D}^2 f(a) \in \mathcal{L}(E; \mathcal{L}(E; F))$ *as a continuous bilinear map in* $\mathcal{L}_2(E, E; F)$*, and we write* $\mathrm{D}^2 f(a)(u, v)$*, instead of* $\mathrm{D}^2 f(a)(u)(v)$*.*

Then the above discussion can be summarized by saying that when $\mathrm{D}^2 f(a)$ is defined, we have

$$\mathrm{D}^2 f(a)(u, v) = \mathrm{D}_u \mathrm{D}_v f(a).$$

Definition 3.15. When E has finite dimension and $(e_1, \ldots, e_n)$ is a basis for E, we denote $\mathrm{D}_{e_j} \mathrm{D}_{e_i} f(a)$ by $\dfrac{\partial^2 f}{\partial x_i \partial x_j}(a)$, when $i \neq j$, and we denote $\mathrm{D}_{e_i} \mathrm{D}_{e_i} f(a)$ by $\dfrac{\partial^2 f}{\partial x_i^2}(a)$.

The following important result attributed to Schwarz can be shown using Proposition 3.11. Given a bilinear map $h \colon E \times E \to F$, recall that h is *symmetric* if

$$h(u, v) = h(v, u),$$

for all $u, v \in E$.

Proposition 3.18. *(Schwarz's lemma) Given two normed vector spaces E and F, given any open subset A of E, given any $f\colon A \to F$, for every $a \in A$, if $\mathrm{D}^2 f(a)$ exists, then $\mathrm{D}^2 f(a) \in \mathcal{L}_2(E, E; F)$ is a continuous symmetric bilinear map. As a corollary, if E is of finite dimension n, and $(e_1, \ldots, e_n)$ is a basis for E, we have*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \frac{\partial^2 f}{\partial x_j \partial x_i}(a).$$

Remark: There is a variation of the above result which does not assume the existence of $\mathrm{D}^2 f(a)$, but instead assumes that $\mathrm{D}_u\mathrm{D}_v f$ and $\mathrm{D}_v\mathrm{D}_u f$ exist on an open subset containing a and are continuous at a, and concludes that $\mathrm{D}_u\mathrm{D}_v f(a) = \mathrm{D}_v\mathrm{D}_u f(a)$. This is a different result which does not imply Proposition 3.18 and is not a consequence of Proposition 3.18.

When $E = \mathbb{R}^2$, the existence of $\dfrac{\partial^2 f}{\partial x \partial y}(a)$ and $\dfrac{\partial^2 f}{\partial y \partial x}(a)$ is not sufficient to ensure the existence of $\mathrm{D}^2 f(a)$.

When E is of finite dimension n and $(e_1, \ldots, e_n)$ is a basis for E, if $\mathrm{D}^2 f(a)$ exists, for every $u = u_1 e_1 + \cdots + u_n e_n$ and $v = v_1 e_1 + \cdots + v_n e_n$ in E, since $\mathrm{D}^2 f(a)$ is a symmetric bilinear form, we have

$$\mathrm{D}^2 f(a)(u, v) = \sum_{i=1,j=1}^{n} u_i v_j \mathrm{D}^2 f(a)(e_i, e_j) = \sum_{i=1,j=1}^{n} u_i v_j \mathrm{D}_{e_j}\mathrm{D}_{e_i} f(a) = \sum_{i=1,j=1}^{n} u_i v_j \frac{\partial^2 f}{\partial x_i \partial x_j}(a),$$

which can be written in matrix form as:

$$\mathrm{D}^2 f(a)(u, v) = U^\top \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1^2}(a) & \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(a) & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(a) \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(a) & \dfrac{\partial^2 f}{\partial x_2^2}(a) & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(a) & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(a) & \cdots & \dfrac{\partial^2 f}{\partial x_n^2}(a) \end{pmatrix} V,$$

where U is the column matrix representing u, and V is the column matrix representing v, over the basis $(e_1, \ldots, e_n)$. Note that the entries in this matrix are *vectors* in F, so the above expression is an abuse of notation, but since the u_i and v_j are scalars, the above expression makes sense since it is a bilinear combination. In the special case where $m = 1$, that is, $F = \mathbb{R}$ or $F = \mathbb{C}$, the Hessian matrix is an $n \times n$ matrix with scalar entries.

Definition 3.16. The above symmetric matrix is called the *Hessian of f at a*.

Example 3.13. Consider the function f defined on real invertible 2×2 matrices such that $ad - bc > 0$ given by

$$f(a, b, c, d) = \log(ad - bc).$$

We immediately verify that the Jacobian matrix of f is given by

$$df_{a,b,c,d} = \frac{1}{ad - bc} \begin{pmatrix} d & -c & -b & a \end{pmatrix}.$$

It is easily checked that if

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad X = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix},$$

then

$$df_A(X) = \mathrm{tr}(A^{-1}X) = \frac{1}{ad - bc} \mathrm{tr}\left(\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix} \right).$$

Computing second-order derivatives, we find that the Hessian matrix of f is given by

$$Hf(A) = \frac{1}{(ad - bc)^2} \begin{pmatrix} -d^2 & cd & bd & -bc \\ cd & -c^2 & -ad & ac \\ bd & -ad & -b^2 & ab \\ -bc & ac & ab & -a^2 \end{pmatrix}.$$

Using the formula for the derivative of the inversion map and the chain rule we can show that

$$\mathrm{D}^2 f(A)(X_1, X_2) = -\mathrm{tr}(A^{-1}X_1A^{-1}X_2),$$

and so

$$Hf(A)(X_1, X_2) = -\mathrm{tr}(A^{-1}X_1A^{-1}X_2),$$

a formula which is far from obvious.

The function f can be generalized to matrices $A \in \mathbf{GL}^+(n, \mathbb{R})$, that is, $n \times n$ real invertible matrices of positive determinants, as

$$f(A) = \log \det(A).$$

It can be shown that the formulae

$$\begin{aligned} df_A(X) &= \mathrm{tr}(A^{-1}X) \\ \mathrm{D}^2 f(A)(X_1, X_2) &= -\mathrm{tr}(A^{-1}X_1A^{-1}X_2) \end{aligned}$$

also hold.

Example 3.14. If we restrict the function of Example 3.13 to symmetric positive definite matrices we obtain the function g defined by

$$g(a,b,c) = \log(ac - b^2).$$

We immediately verify that the Jacobian matrix of g is given by

$$dg_{a,b,c} = \frac{1}{ac-b^2}\begin{pmatrix} c & -2b & a \end{pmatrix}.$$

Computing second-order derivatives, we find that the Hessian matrix of g is given by

$$Hg(a,b,c) = \frac{1}{(ac-b^2)^2}\begin{pmatrix} -c^2 & 2bc & -b^2 \\ 2bc & -2(b^2+ac) & 2ab \\ -b^2 & 2ab & -a^2 \end{pmatrix}.$$

Although this is not obvious, it can be shown that if $ac - b^2 > 0$ and $a, c > 0$, then the matrix $-Hg(a,b,c)$ is symmetric positive definite.

We now indicate briefly how higher-order derivatives are defined. Let $m \geq 2$. Given a function $f\colon A \to F$ as before, for any $a \in A$, if the derivatives $\mathrm{D}^i f$ exist on A for all i, $1 \leq i \leq m-1$, by induction, $\mathrm{D}^{m-1}f$ can be considered to be a continuous function $\mathrm{D}^{m-1}f\colon A \to \mathcal{L}_{m-1}(E^{m-1}; F)$.

Definition 3.17. Define $\mathrm{D}^m f(a)$, the *m-th derivative of f at a*, as

$$\mathrm{D}^m f(a) = \mathrm{D}(\mathrm{D}^{m-1}f)(a).$$

Then $\mathrm{D}^m f(a)$ *can be identified with a continuous m-multilinear map in* $\mathcal{L}_m(E^m; F)$. We can then show (as we did before) that if $\mathrm{D}^m f(a)$ is defined, then

$$\mathrm{D}^m f(a)(u_1, \ldots, u_m) = \mathrm{D}_{u_1} \ldots \mathrm{D}_{u_m} f(a).$$

Definition 3.18. When E if of finite dimension n and $(e_1, \ldots, e_n)$ is a basis for E, if $\mathrm{D}^m f(a)$ exists, for every $j_1, \ldots, j_m \in \{1, \ldots, n\}$, we denote $\mathrm{D}_{e_{j_m}} \ldots \mathrm{D}_{e_{j_1}} f(a)$ by

$$\frac{\partial^m f}{\partial x_{j_1} \ldots \partial x_{j_m}}(a).$$

Example 3.15. Going back to the function f of Example 3.13 given by $f(A) = \log\det(A)$, using the formula for the derivative of the inversion map, the chain rule and the product rule, we can show that

$$\mathrm{D}^m f(A)(X_1, \ldots, X_m) = (-1)^{m-1} \sum_{\sigma \in \mathfrak{S}_{m-1}} \mathrm{tr}(A^{-1}X_1A^{-1}X_{\sigma(1)+1}A^{-1}X_{\sigma(2)+1} \cdots A^{-1}X_{\sigma(m-1)+1})$$

for any $m \geq 1$, where $A \in \mathbf{GL}^+(n, \mathbb{R})$ and $X_1, \ldots X_m$ are any $n \times n$ real matrices.

Given a m-multilinear map $h \in \mathcal{L}_m(E^m; F)$, recall that h is *symmetric* if

$$h(u_{\pi(1)}, \ldots, u_{\pi(m)}) = h(u_1, \ldots, u_m),$$

for all $u_1, \ldots, u_m \in E$, and all permutations π on $\{1, \ldots, m\}$. Then the following generalization of Schwarz's lemma holds.

Proposition 3.19. *Given two normed vector spaces E and F, given any open subset A of E, given any $f\colon A \to F$, for every $a \in A$, for every $m \geq 1$, if $\mathrm{D}^m f(a)$ exists, then $\mathrm{D}^m f(a) \in \mathcal{L}_m(E^m; F)$ is a continuous symmetric m-multilinear map. As a corollary, if E is of finite dimension n, and $(e_1, \ldots, e_n)$ is a basis for E, we have*

$$\frac{\partial^m f}{\partial x_{j_1} \ldots \partial x_{j_m}}(a) = \frac{\partial^m f}{\partial x_{\pi(j_1)} \ldots \partial x_{\pi(j_m)}}(a),$$

for every $j_1, \ldots, j_m \in \{1, \ldots, n\}$, and for every permutation π on $\{1, \ldots, m\}$.

Because the trace function is invariant under permutation of its arguments ($\mathrm{tr}(XY) = \mathrm{tr}(YX)$), we see that the m-th derivatives in Example 3.15 are indeed symmetric multilinear maps.

If E is of finite dimension n, and $(e_1, \ldots, e_n)$ is a basis for E, $\mathrm{D}^m f(a)$ is a symmetric m-multilinear map, and we have

$$\mathrm{D}^m f(a)(u_1, \ldots, u_m) = \sum_j u_{1,j_1} \cdots u_{m,j_m} \frac{\partial^m f}{\partial x_{j_1} \ldots \partial x_{j_m}}(a),$$

where j ranges over all functions $j\colon \{1, \ldots, m\} \to \{1, \ldots, n\}$, for any m vectors

$$u_j = u_{j,1} e_1 + \cdots + u_{j,n} e_n.$$

The concept of C^1-function is generalized to the concept of C^m-function, and Theorem 3.12 can also be generalized.

Definition 3.19. Given two normed vector spaces E and F, and an open subset A of E, for any $m \geq 1$, we say that a function $f\colon A \to F$ is *of class C^m on A or a C^m-function on A* if $\mathrm{D}^k f$ exists and is continuous on A for every k, $1 \leq k \leq m$. We say that $f\colon A \to F$ is *of class C^∞ on A or a C^∞-function on A* if $\mathrm{D}^k f$ exists and is continuous on A for every $k \geq 1$. A C^∞-function (on A) is also called a *smooth function* (on A). A *C^m-diffeomorphism $f\colon A \to B$ between A and B* (where A is an open subset of E and B is an open subset of B) is a bijection between A and $B = f(A)$, such that both $f\colon A \to B$ and its inverse $f^{-1}\colon B \to A$ are C^m-functions.

Equivalently, f is a C^m-function on A if f is a C^1-function on A and $\mathrm{D}f$ is a C^{m-1}-function on A.

We have the following theorem giving a necessary and sufficient condition for f to a C^m-function on A.

Theorem 3.20. *Given two normed vector spaces E and F, where E is of finite dimension n, and where $(u_1, \ldots, u_n)$ is a basis of E, given any open subset A of E, given any function $f\colon A \to F$, for any $m \geq 1$, the derivative $\mathrm{D}^m f$ is a C^m-function on A iff every partial derivative $\mathrm{D}_{u_{j_k}} \ldots \mathrm{D}_{u_{j_1}} f$ (or $\dfrac{\partial^k f}{\partial x_{j_1} \ldots \partial x_{j_k}}(a)$) is defined and continuous on A, for all k, $1 \leq k \leq m$, and all $j_1, \ldots, j_k \in \{1, \ldots, n\}$. As a corollary, if F is of finite dimension p, and $(v_1, \ldots, v_p)$ is a basis of F, the derivative $\mathrm{D}^m f$ is defined and continuous on A iff every partial derivative $\mathrm{D}_{u_{j_k}} \ldots \mathrm{D}_{u_{j_1}} f_i$ (or $\dfrac{\partial^k f_i}{\partial x_{j_1} \ldots \partial x_{j_k}}(a)$) is defined and continuous on A, for all k, $1 \leq k \leq m$, for all i, $1 \leq i \leq p$, and all $j_1, \ldots, j_k \in \{1, \ldots, n\}$.*

Definition 3.20. When $E = \mathbb{R}$ (or $E = \mathbb{C}$), for any $a \in E$, $\mathrm{D}^m f(a)(1, \ldots, 1)$ is a vector in F, called the *mth-order vector derivative*. As in the case $m = 1$, we will usually identify the multilinear map $\mathrm{D}^m f(a)$ with the vector $\mathrm{D}^m f(a)(1, \ldots, 1)$.

Some notational conventions can also be introduced to simplify the notation of higher-order derivatives, and we discuss such conventions very briefly.

Recall that when E is of finite dimension n, and $(e_1, \ldots, e_n)$ is a basis for E, $\mathrm{D}^m f(a)$ is a symmetric m-multilinear map, and we have

$$\mathrm{D}^m f(a)(u_1, \ldots, u_m) = \sum_j u_{1,j_1} \cdots u_{m,j_m} \frac{\partial^m f}{\partial x_{j_1} \ldots \partial x_{j_m}}(a),$$

where j ranges over all functions $j\colon \{1, \ldots, m\} \to \{1, \ldots, n\}$, for any m vectors

$$u_j = u_{j,1} e_1 + \cdots + u_{j,n} e_n.$$

We can then group the various occurrences of ∂x_{j_k} corresponding to the same variable x_{j_k}, and this leads to the notation

$$\Big(\frac{\partial}{\partial x_1}\Big)^{\alpha_1} \Big(\frac{\partial}{\partial x_2}\Big)^{\alpha_2} \cdots \Big(\frac{\partial}{\partial x_n}\Big)^{\alpha_n} f(a),$$

where $\alpha_1 + \alpha_2 + \cdots + \alpha_n = m$.

If we denote $(\alpha_1, \ldots, \alpha_n)$ simply by α, then we denote

$$\Big(\frac{\partial}{\partial x_1}\Big)^{\alpha_1} \Big(\frac{\partial}{\partial x_2}\Big)^{\alpha_2} \cdots \Big(\frac{\partial}{\partial x_n}\Big)^{\alpha_n} f$$

by

$$\partial^\alpha f, \quad \text{or} \quad \Big(\frac{\partial}{\partial x}\Big)^{\alpha} f.$$

If $\alpha = (\alpha_1, \ldots, \alpha_n)$, we let $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n$, $\alpha! = \alpha_1! \cdots \alpha_n!$, and if $h = (h_1, \ldots, h_n)$, we denote $h_1^{\alpha_1} \cdots h_n^{\alpha_n}$ by h^α.

In the next section we survey various versions of Taylor's formula.

3.6 Taylor's Formula, Faà di Bruno's Formula

We discuss, without proofs, several versions of Taylor's formula. The hypotheses required in each version become increasingly stronger. The first version can be viewed as a generalization of the notion of derivative. Given an m-linear map $f\colon E^m \to F$, for any vector $h \in E$, we abbreviate

$$f(\underbrace{h, \ldots, h}_{m})$$

by $f(h^m)$. The version of Taylor's formula given next is sometimes referred to as the *formula of Taylor–Young*.

Theorem 3.21. *(Taylor–Young) Given two normed vector spaces E and F, for any open subset $A \subseteq E$, for any function $f\colon A \to F$, for any $a \in A$, if $\mathrm{D}^k f$ exists in A for all k, $1 \leq k \leq m-1$, and if $\mathrm{D}^m f(a)$ exists, then we have:*

$$f(a+h) = f(a) + \frac{1}{1!}\mathrm{D}^1 f(a)(h) + \cdots + \frac{1}{m!}\mathrm{D}^m f(a)(h^m) + \|h\|^m \epsilon(h),$$

for any h such that $a + h \in A$, and where $\lim_{h \to 0,\, h \neq 0} \epsilon(h) = 0$.

The above version of Taylor's formula has applications to the study of relative maxima (or minima) of real-valued functions. It is also used to study the local properties of curves and surfaces.

The next version of Taylor's formula can be viewed as a generalization of Proposition 3.11. It is sometimes called the *Taylor formula with Lagrange remainder* or *generalized mean value theorem*.

Theorem 3.22. *(Generalized mean value theorem) Let E and F be two normed vector spaces, let A be an open subset of E, and let $f\colon A \to F$ be a function on A. Given any $a \in A$ and any $h \neq 0$ in E, if the closed segment $[a, a+h]$ is contained in A, $\mathrm{D}^k f$ exists in A for all k, $1 \leq k \leq m$, $\mathrm{D}^{m+1} f(x)$ exists at every point x of the open segment $(a, a+h)$, and*

$$\max_{x \in (a, a+h)} \left\|\mathrm{D}^{m+1} f(x)\right\| \leq M,$$

for some $M \geq 0$, then

$$\left\| f(a+h) - f(a) - \left(\frac{1}{1!}\mathrm{D}^1 f(a)(h) + \cdots + \frac{1}{m!}\mathrm{D}^m f(a)(h^m)\right)\right\| \leq M \frac{\|h\|^{m+1}}{(m+1)!}.$$

As a corollary, if $L\colon E^{m+1} \to F$ is a continuous $(m+1)$-linear map, then

$$\left\| f(a+h) - f(a) - \left(\frac{1}{1!}\mathrm{D}^1 f(a)(h) + \cdots + \frac{1}{m!}\mathrm{D}^m f(a)(h^m) + \frac{L(h^{m+1})}{(m+1)!}\right)\right\| \leq M \frac{\|h\|^{m+1}}{(m+1)!},$$

where $M = \max_{x \in (a, a+h)} \|\mathrm{D}^{m+1} f(x) - L\|$.

The above theorem is sometimes stated under the slightly stronger assumption that f is a C^m-function on A. If $f\colon A \to \mathbb{R}$ is a real-valued function, Theorem 3.22 can be refined a little bit. This version is often called the *formula of Taylor–Maclaurin*.

Theorem 3.23. *(Taylor–Maclaurin) Let E be a normed vector space, let A be an open subset of E, and let $f\colon A \to \mathbb{R}$ be a real-valued function on A. Given any $a \in A$ and any $h \neq 0$ in E, if the closed segment $[a, a+h]$ is contained in A, if $\mathrm{D}^k f$ exists in A for all k, $1 \leq k \leq m$, and $\mathrm{D}^{m+1} f(x)$ exists at every point x of the open segment $(a, a+h)$, then there is some $\theta \in \mathbb{R}$, with $0 < \theta < 1$, such that*

$$f(a+h) = f(a) + \frac{1}{1!}\mathrm{D}^1 f(a)(h) + \cdots + \frac{1}{m!}\mathrm{D}^m f(a)(h^m) + \frac{1}{(m+1)!}\mathrm{D}^{m+1} f(a+\theta h)(h^{m+1}).$$

Example 3.16. Going back to the function f of Example 3.13 given by $f(A) = \log\det(A)$, we know from Example 3.15 that

$$\mathrm{D}^m f(A)(X_1, \ldots, X_m) = (-1)^{m-1} \sum_{\sigma \in \mathfrak{S}_{m-1}} \mathrm{tr}(A^{-1}X_1 A^{-1} X_{\sigma(1)+1} \cdots A^{-1} X_{\sigma(m-1)+1}) \qquad (*)$$

for all $m \geq 1$, with $A \in \mathbf{GL}^+(n, \mathbb{R})$. If we make the stronger assumption that A is symmetric positive definite, then for any other symmetric positive definite matrix B, since the symmetric positive definite matrices form a convex set, the matrices $A + \theta(B - A) = (1-\theta)A + \theta B$ are also symmetric positive definite for $\theta \in [0, 1]$. Theorem 3.23 applies with $H = B - A$ (a symmetric matrix), and using $(*)$, we obtain

$$\begin{aligned}\log\det(A+H) = \log\det(A) + \mathrm{tr}\Big(&A^{-1}H - \frac{1}{2}(A^{-1}H)^2 + \cdots + \frac{(-1)^{m-1}}{m}(A^{-1}H)^m \\ &+ \frac{(-1)^m}{m+1}((A+\theta H)^{-1}H)^{m+1}\Big),\end{aligned}$$

for some θ such that $0 < \theta < 1$. In particular, if $A = I$, for any symmetric matrix H such that $I + H$ is symmetric positive definite, we obtain

$$\begin{aligned}\log\det(I+H) = \mathrm{tr}\Big(&H - \frac{1}{2}H^2 + \cdots + \frac{(-1)^{m-1}}{m}H^m \\ &+ \frac{(-1)^m}{m+1}((I+\theta H)^{-1}H)^{m+1}\Big),\end{aligned}$$

for some θ such that $0 < \theta < 1$. In the special case when $n = 1$, we have $I = 1$, H is a real such that $1 + H > 0$ and the trace function is the identity, so we recognize the partial sum of the series for $x \mapsto \log(1+x)$,

$$\begin{aligned}\log(1+H) = &H - \frac{1}{2}H^2 + \cdots + \frac{(-1)^{m-1}}{m}H^m \\ &+ \frac{(-1)^m}{m+1}(1+\theta H)^{-(m+1)}H^{m+1}.\end{aligned}$$

We also mention for "mathematical culture," a version with integral remainder, in the case of a real-valued function. This is usually called *Taylor's formula with integral remainder*.

Theorem 3.24. *(Taylor's formula with integral remainder) Let E be a normed vector space, let A be an open subset of E, and let $f\colon A \to \mathbb{R}$ be a real-valued function on A. Given any $a \in A$ and any $h \neq 0$ in E, if the closed segment $[a, a+h]$ is contained in A, and if f is a C^{m+1}-function on A, then we have*

$$f(a+h) = f(a) + \frac{1}{1!}\mathrm{D}^1 f(a)(h) + \cdots + \frac{1}{m!}\mathrm{D}^m f(a)(h^m) + \int_0^1 \frac{(1-t)^m}{m!}\Big[\mathrm{D}^{m+1} f(a+th)(h^{m+1})\Big]\, dt.$$

The advantage of the above formula is that it gives an explicit remainder.

We now examine briefly the situation where E is of finite dimension n, and $(e_1, \ldots, e_n)$ is a basis for E. In this case we get a more explicit expression for the expression

$$\sum_{i=0}^{k=m} \frac{1}{k!}\mathrm{D}^k f(a)(h^k)$$

involved in all versions of Taylor's formula, where by convention, $\mathrm{D}^0 f(a)(h^0) = f(a)$. If $h = h_1 e_1 + \cdots + h_n e_n$, then we have

$$\sum_{k=0}^{k=m} \frac{1}{k!}\mathrm{D}^k f(a)(h^k) = \sum_{k_1+\cdots+k_n \leq m} \frac{h_1^{k_1}\cdots h_n^{k_n}}{k_1!\cdots k_n!}\left(\frac{\partial}{\partial x_1}\right)^{k_1}\cdots\left(\frac{\partial}{\partial x_n}\right)^{k_n} f(a),$$

which, using the abbreviated notation introduced at the end of Section 3.5, can also be written as

$$\sum_{k=0}^{k=m} \frac{1}{k!}\mathrm{D}^k f(a)(h^k) = \sum_{|\alpha| \leq m} \frac{h^\alpha}{\alpha!}\,\partial^\alpha f(a).$$

The advantage of the above notation is that it is the same as the notation used when $n = 1$, i.e., when $E = \mathbb{R}$ (or $E = \mathbb{C}$). Indeed, in this case, the Taylor–Maclaurin formula reads as:

$$f(a+h) = f(a) + \frac{h}{1!}\mathrm{D}^1 f(a) + \cdots + \frac{h^m}{m!}\mathrm{D}^m f(a) + \frac{h^{m+1}}{(m+1)!}\mathrm{D}^{m+1} f(a+\theta h),$$

for some $\theta \in \mathbb{R}$, with $0 < \theta < 1$, where $\mathrm{D}^k f(a)$ is the value of the k-th derivative of f at a (and thus, as we have already said several times, this is the kth-order vector derivative, which is just a scalar, since $F = \mathbb{R}$).

In the above formula, the assumptions are that $f\colon [a, a+h] \to \mathbb{R}$ is a C^m-function on $[a, a+h]$, and that $\mathrm{D}^{m+1} f(x)$ exists for every $x \in (a, a+h)$.

Taylor's formula is useful to study the local properties of curves and surfaces. In the case of a curve, we consider a function $f\colon [r,s] \to F$ from a closed interval $[r,s]$ of $\mathbb{R}$ to some vector space F, the derivatives $\mathrm{D}^k f(a)(h^k)$ correspond to vectors $h^k \mathrm{D}^k f(a)$, where $\mathrm{D}^k f(a)$ is the kth vector derivative of f at a (which is really $\mathrm{D}^k f(a)(1,\ldots,1)$), and for any $a \in (r,s)$, Theorem 3.21 yields the following formula:

$$f(a+h) = f(a) + \frac{h}{1!}\mathrm{D}^1 f(a) + \cdots + \frac{h^m}{m!}\mathrm{D}^m f(a) + h^m \epsilon(h),$$

for any h such that $a + h \in (r,s)$, and where $\lim_{h\to 0,\, h\neq 0} \epsilon(h) = 0$.

In the case of functions $f\colon \mathbb{R}^n \to \mathbb{R}$, it is convenient to have formulae for the Taylor–Young formula and the Taylor–Maclaurin formula in terms of the gradient and the Hessian. Recall that the *gradient* $\nabla f(a)$ of f at $a \in \mathbb{R}^n$ is the column vector

$$\nabla f(a) = \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(a) \\ \dfrac{\partial f}{\partial x_2}(a) \\ \vdots \\ \dfrac{\partial f}{\partial x_n}(a) \end{pmatrix},$$

and that

$$f'(a)(u) = \mathrm{D}f(a)(u) = \nabla f(a) \cdot u,$$

for any $u \in \mathbb{R}^n$ (where $\cdot$ means inner product). The above equation shows that *the direction of the gradient $\nabla f(a)$ is the direction of maximal increase of the function f at a* and that *$\|\nabla f(a)\|$ is the rate of change of f in its direction of maximal increase*. This is the reason why methods of "gradient descent" pick the direction *opposite* to the gradient (we are trying to minimize f).

The *Hessian matrix* $\nabla^2 f(a)$ of f at $a \in \mathbb{R}^n$ is the $n \times n$ symmetric matrix

$$\nabla^2 f(a) = \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1^2}(a) & \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(a) & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(a) \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(a) & \dfrac{\partial^2 f}{\partial x_2^2}(a) & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(a) & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(a) & \cdots & \dfrac{\partial^2 f}{\partial x_n^2}(a) \end{pmatrix},$$

and we have

$$\mathrm{D}^2 f(a)(u,v) = u^\top \nabla^2 f(a)\, v = u \cdot \nabla^2 f(a) v = \nabla^2 f(a) u \cdot v,$$

for all $u, v \in \mathbb{R}^n$. This is the special case of Definition 3.16 where $E = \mathbb{R}^n$ and $F = \mathbb{R}$. Then we have the following three formulations of the formula of Taylor–Young of order 2:

$$f(a+h) = f(a) + \mathrm{D}f(a)(h) + \frac{1}{2}\mathrm{D}^2 f(a)(h,h) + \|h\|^2 \epsilon(h)$$
$$f(a+h) = f(a) + \nabla f(a) \cdot h + \frac{1}{2}(h \cdot \nabla^2 f(a) h) + (h \cdot h)\epsilon(h)$$
$$f(a+h) = f(a) + (\nabla f(a))^\top h + \frac{1}{2}(h^\top \nabla^2 f(a)\, h) + (h^\top h)\epsilon(h),$$

with $\lim_{h \mapsto 0} \epsilon(h) = 0$.

One should keep in mind that only the first formula is intrinsic (i.e., does not depend on the choice of a basis), whereas the other two depend on the basis and the inner product chosen on $\mathbb{R}^n$. As an exercise, the reader should write similar formulae for the Taylor–Maclaurin formula of order 2.

Another application of Taylor's formula is the derivation of a formula which gives the m-th derivative of the composition of two functions, usually known as "Faà di Bruno's formula." This formula is useful when dealing with geometric continuity of splines curves and surfaces.

Proposition 3.25. *Given any normed vector space E, for any function $f\colon \mathbb{R} \to \mathbb{R}$ and any function $g\colon \mathbb{R} \to E$, for any $a \in \mathbb{R}$, letting $b = f(a)$, $f^{(i)}(a) = \mathrm{D}^i f(a)$, and $g^{(i)}(b) = \mathrm{D}^i g(b)$, for any $m \geq 1$, if $f^{(i)}(a)$ and $g^{(i)}(b)$ exist for all i, $1 \leq i \leq m$, then $(g \circ f)^{(m)}(a) = \mathrm{D}^m(g \circ f)(a)$ exists and is given by the following formula:*

$$(g \circ f)^{(m)}(a) = \sum_{0 \leq j \leq m} \sum_{\substack{i_1+i_2+\cdots+i_m = j \\ i_1+2i_2+\cdots+mi_m = m \\ i_1, i_2, \cdots, i_m \geq 0}} \frac{m!}{i_1! \cdots i_m!}\, g^{(j)}(b) \left(\frac{f^{(1)}(a)}{1!}\right)^{i_1} \cdots \left(\frac{f^{(m)}(a)}{m!}\right)^{i_m}.$$

When $m = 1$, the above simplifies to the familiar formula

$$(g \circ f)'(a) = g'(b) f'(a),$$

and for $m = 2$, we have

$$(g \circ f)^{(2)}(a) = g^{(2)}(b)(f^{(1)}(a))^2 + g^{(1)}(b) f^{(2)}(a).$$

3.7 Further Readings

A thorough treatment of differential calculus can be found in Munkres [58], Lang [50], Schwartz [70], Cartan [21], and Avez [5]. The techniques of differential calculus have many applications, especially to the geometry of curves and surfaces and to differential geometry in general. For this, we recommend do Carmo [29, 30] (two beautiful classics on the subject), Kreyszig [46], Stoker [75], Gray [38], Berger and Gostiaux [8], Milnor [56], Lang [48], Warner [82] and Choquet-Bruhat [23].

3.8 Summary

The main concepts and results of this chapter are listed below:

- *Directional derivative* ($\mathrm{D}_u f(a)$).
- *Total derivative*, *Fréchet derivative*, *derivative*, *total differential*, *differential* ($df(a), df_a$).
- *Partial derivatives*.
- *Affine* functions.
- The *chain rule*.
- *Jacobian matrices* ($J(f)(a)$), *Jacobians*.
- *Gradient* of a function ($\operatorname{grad} f(a)$, $\nabla f(a)$).
- *Mean value theorem*.
- C^0*-functions*, C^1*-functions*.
- The *implicit function theorem*.
- *Local homeomorphisms, local diffeomorphisms*, *diffeomorphisms*.
- The *inverse function theorem*.
- *Immersions, submersions*.
- Second-order derivatives.
- *Schwarz's lemma*.
- *Hessian matrix*.
- C^∞*-functions*, *smooth functions*.
- *Taylor–Young's formula*.
- Generalized mean value theorem.
- *Taylor–MacLaurin's formula*.
- *Taylor's formula with integral remainder*.
- *Faà di Bruno's formula*.

3.9 Problems

Problem 3.1. Let $f\colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ be the function defined on $n \times n$ matrices by

$$f(A) = A^2.$$

Prove that

$$\mathrm{D}f_A(H) = AH + HA,$$

for all $A, H \in \mathrm{M}_n(\mathbb{R})$.

Problem 3.2. Let $f\colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ be the function defined on $n \times n$ matrices by

$$f(A) = A^3.$$

Prove that

$$\mathrm{D}f_A(H) = A^2H + AHA + HA^2,$$

for all $A, H \in \mathrm{M}_n(\mathbb{R})$.

Problem 3.3. If $f\colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ and $g\colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ are differentiable matrix functions, prove that

$$d(fg)_A(B) = df_A(B)g(A) + f(A)dg_A(B),$$

for all $A, B \in \mathrm{M}_n(\mathbb{R})$.

Problem 3.4. Recall that $\mathfrak{so}(3)$ denotes the vector space of real skew-symmetric $n \times n$ matrices ($B^\top = -B$). Let $C\colon \mathfrak{so}(n) \to \mathrm{M}_n(\mathbb{R})$ be the function given by

$$C(B) = (I - B)(I + B)^{-1}.$$

(1) Prove that if B is skew-symmetric, then $I - B$ and $I + B$ are invertible, and so C is well-defined. Prove that

(2) Prove that

$$dC(B)(A) = -[I + (I - B)(I + B)^{-1}]A(I + B)^{-1} = -2(I + B)^{-1}A(I + B)^{-1}.$$

(3) Prove that $dC(B)$ is injective for every skew-symmetric matrix B.

Problem 3.5. Prove that

$$\begin{aligned} &d^mC_B(H_1, \ldots, H_m) \\ &\quad = 2(-1)^m \sum_{\pi \in \mathfrak{S}_m} (I + B)^{-1}H_{\pi(1)}(I + B)^{-1}H_{\pi(2)}(I + B)^{-1} \cdots (I + B)^{-1}H_{\pi(m)}(I + B)^{-1}. \end{aligned}$$

Problem 3.6. Consider the function g defined for all $A \in \mathbf{GL}(n, \mathbb{R})$, that is, all $n \times n$ real invertible matrices, given by

$$g(A) = \det(A).$$

(1) Prove that

$$dg_A(X) = \det(A)\mathrm{tr}(A^{-1}X),$$

for all $n \times n$ real matrices X.

(2) Consider the function f defined for all $A \in \mathbf{GL}^+(n, \mathbb{R})$, that is, $n \times n$ real invertible matrices of positive determinants, given by

$$f(A) = \log g(A) = \log \det(A).$$

Prove that

$$\begin{aligned} df_A(X) &= \mathrm{tr}(A^{-1}X) \\ \mathrm{D}^2 f(A)(X_1, X_2) &= -\mathrm{tr}(A^{-1}X_1A^{-1}X_2), \end{aligned}$$

for all $n \times n$ real matrices X, X_1, X_2.

(3) Prove that

$$\mathrm{D}^m f(A)(X_1, \ldots, X_m) = (-1)^{m-1} \sum_{\sigma \in \mathfrak{S}_{m-1}} \mathrm{tr}(A^{-1}X_1A^{-1}X_{\sigma(1)+1}A^{-1}X_{\sigma(2)+1} \cdots A^{-1}X_{\sigma(m-1)+1})$$

for any $m \geq 1$, where $X_1, \ldots X_m$ are any $n \times n$ real matrices.

Chapter 4

Extrema of Real-Valued Functions

This chapter deals with extrema of real-valued functions. In most optimization problems, we need to find necessary conditions for a function $J: \Omega \to \mathbb{R}$ to have a local extremum with respect to a subset U of Ω (where Ω is open). This can be done in two cases:

(1) The set U is defined by a set of equations,

$$U = \{x \in \Omega \mid \varphi_i(x) = 0,\ \ 1 \leq i \leq m\},$$

where the functions $\varphi_i: \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

(2) The set U is defined by a set of inequalities,

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ \ 1 \leq i \leq m\},$$

where the functions $\varphi_i: \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

In (1), the equations $\varphi_i(x) = 0$ are called *equality constraints*, and in (2), the inequalities $\varphi_i(x) \leq 0$ are called *inequality constraints*. The case of equality constraints is much easier to deal with and is treated in this chapter.

If the functions φ_i are convex and Ω is convex, then U is convex. This is a very important case that we discuss later. In particular, if the functions φ_i are affine, then the equality constraints can be written as $Ax = b$, and the inequality constraints as $Ax \leq b$, for some $m \times n$ matrix A and some vector $b \in \mathbb{R}^m$. We will also discuss the case of affine constraints later.

In the case of equality constraints, a necessary condition for a local extremum with respect to U can be given in terms of *Lagrange multipliers*. In the case of inequality constraints, there is also a necessary condition for a local extremum with respect to U in terms of generalized Lagrange multipliers and the *Karush–Kuhn–Tucker* conditions. This will be discussed in Chapter 14.

4.1 Local Extrema, Constrained Local Extrema, and Lagrange Multipliers

Let $J: E \to \mathbb{R}$ be a real-valued function defined on a normed vector space E (or more generally, any topological space). Ideally we would like to find where the function J reaches a minimum or a maximum value, at least locally. In this chapter we will usually use the notations $dJ(u)$ or $J'(u)$ (or dJ_u or J'_u) for the derivative of J at u, instead of $\mathrm{D}J(u)$. Our presentation follows very closely that of Ciarlet [25] (Chapter 7), which we find to be one of the clearest.

Definition 4.1. If $J: E \to \mathbb{R}$ is a real-valued function defined on a normed vector space E, we say that J has a *local minimum* (or *relative minimum*) at the point $u \in E$ if there is some open subset $W \subseteq E$ containing u such that

$$J(u) \leq J(w) \quad \text{for all } w \in W.$$

Similarly, we say that J has a *local maximum* (or *relative maximum*) at the point $u \in E$ if there is some open subset $W \subseteq E$ containing u such that

$$J(u) \geq J(w) \quad \text{for all } w \in W.$$

In either case, we say that J has a *local extremum* (or *relative extremum*) at u. We say that J has a *strict local minimum* (resp. *strict local maximum*) at the point $u \in E$ if there is some open subset $W \subseteq E$ containing u such that

$$J(u) < J(w) \quad \text{for all } w \in W - \{u\}$$

(resp.

$$J(u) > J(w) \quad \text{for all } w \in W - \{u\}).$$

By abuse of language, we often say that the point u itself "is a local minimum" or a "local maximum," even though, strictly speaking, this does not make sense.

We begin with a well-known necessary condition for a local extremum.

Proposition 4.1. *Let E be a normed vector space and let $J: \Omega \to \mathbb{R}$ be a function, with Ω some open subset of E. If the function J has a local extremum at some point $u \in \Omega$ and if J is differentiable at u, then*

$$dJ_u = J'(u) = 0.$$

Proof. Pick any $v \in E$. Since Ω is open, for t small enough we have $u + tv \in \Omega$, so there is an open interval $I \subseteq \mathbb{R}$ such that the function φ given by

$$\varphi(t) = J(u + tv)$$

for all $t \in I$ is well-defined. By applying the chain rule, we see that φ is differentiable at $t = 0$, and we get

$$\varphi'(0) = dJ_u(v).$$

Without loss of generality, assume that u is a local minimum. Then we have

$$\varphi'(0) = \lim_{t \mapsto 0_-} \frac{\varphi(t) - \varphi(0)}{t} \leq 0$$

and

$$\varphi'(0) = \lim_{t \mapsto 0_+} \frac{\varphi(t) - \varphi(0)}{t} \geq 0,$$

which shows that $\varphi'(0) = dJ_u(v) = 0$. As $v \in E$ is arbitrary, we conclude that $dJ_u = 0$. □

Definition 4.2. A point $u \in \Omega$ such that $J'(u) = 0$ is called a *critical point* of J.

If $E = \mathbb{R}^n$, then the condition $dJ_u = 0$ is equivalent to the system

$$\frac{\partial J}{\partial x_1}(u_1, \ldots, u_n) = 0$$
$$\vdots$$
$$\frac{\partial J}{\partial x_n}(u_1, \ldots, u_n) = 0.$$

The condition of Proposition 4.1 is only a *necessary* condition for the existence of an extremum, but not a sufficient condition.

Here are some counter-examples. If $f : \mathbb{R} \to \mathbb{R}$ is the function given by $f(x) = x^3$, since $f'(x) = 3x^2$, we have $f'(0) = 0$, but 0 is neither a minimum nor a maximum of f as evidenced by the graph shown in Figure 4.1.

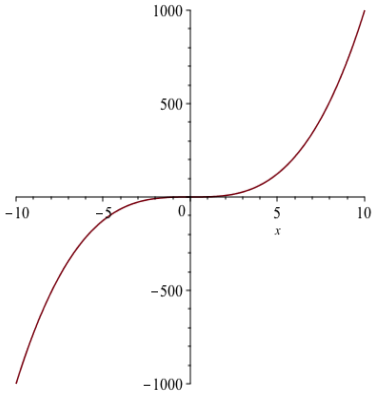


Figure 4.1: The graph of $f(x) = x^3$. Note that $x = 0$ is a saddle point and not a local extremum.

If $g\colon \mathbb{R}^2 \to \mathbb{R}$ is the function given by $g(x,y) = x^2 - y^2$, then $g'_{(x,y)} = (2x \quad -2y)$, so $g'_{(0,0)} = (0 \;\; 0)$, yet near $(0,0)$ the function g takes negative and positive values. See Figure 4.2.

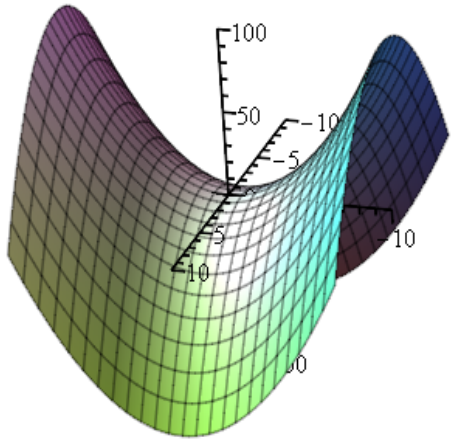


Figure 4.2: The graph of $g(x,y) = x^2 - y^2$. Note that $(0,0)$ is a saddle point and not a local extremum.

It is very important to note that the hypothesis that Ω *is open* is crucial for the validity of Proposition 4.1.

For example, if J is the identity function on $\mathbb{R}$ and $U = [0,1]$, a closed subset, then $J'(x) = 1$ for all $x \in [0,1]$, even though J has a minimum at $x = 0$ and a maximum at $x = 1$.

In many practical situations, we need to look for local extrema of a function J *under additional constraints*. This situation can be formalized conveniently as follows. We have a function $J\colon \Omega \to \mathbb{R}$ defined on some open subset Ω of a normed vector space, but we also have some subset U of Ω, and we are looking for the local extrema of J *with respect to the set* U.

The elements $u \in U$ are often called *feasible solutions* of the optimization problem consisting in finding the local extrema of some objective function J with respect to some subset U of Ω defined by a set of constraints. Note that in most cases, U is *not* open. In fact, U is usually closed.

Definition 4.3. If $J\colon \Omega \to \mathbb{R}$ is a real-valued function defined on some open subset Ω of a normed vector space E and if U is some subset of Ω, we say that J has a *local minimum* (or *relative minimum*) at the point $u \in U$ *with respect to* U if there is some open subset $W \subseteq \Omega$ containing u such that

$$J(u) \leq J(w) \quad \text{for all } w \in U \cap W.$$

Similarly, we say that J has a *local maximum* (or *relative maximum*) at the point $u \in U$ *with respect to* U if there is some open subset $W \subseteq \Omega$ containing u such that

$$J(u) \geq J(w) \quad \text{for all } w \in U \cap W.$$

In either case, we say that J has a *local extremum* at u *with respect to* U.

In order to find necessary conditions for a function $J\colon \Omega \to \mathbb{R}$ to have a local extremum with respect to a subset U of Ω (where Ω is open), we need to somehow incorporate the definition of U into these conditions. This can be done in two cases:

(1) The set U is defined by a set of equations,

$$U = \{x \in \Omega \mid \varphi_i(x) = 0,\ \ 1 \leq i \leq m\},$$

where the functions $\varphi_i\colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

(2) The set U is defined by a set of inequalities,

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ \ 1 \leq i \leq m\},$$

where the functions $\varphi_i\colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

In (1), the equations $\varphi_i(x) = 0$ are called *equality constraints*, and in (2), the inequalities $\varphi_i(x) \leq 0$ are called *inequality constraints*.

An inequality constraint of the form $\varphi_i(x) \geq 0$ is equivalent to the inequality constraint $-\varphi_x(x) \leq 0$. An equality constraint $\varphi_i(x) = 0$ is equivalent to the conjunction of the two inequality constraints $\varphi_i(x) \leq 0$ and $-\varphi_i(x) \leq 0$, so the case of inequality constraints subsumes the case of equality constraints. However, the case of equality constraints is easier to deal with, and in this chapter we will restrict our attention to this case.

If the functions φ_i are convex and Ω is convex, then U is convex. This is a very important case that we will discuss later. In particular, if the functions φ_i are affine, then the equality constraints can be written as $Ax = b$, and the inequality constraints as $Ax \leq b$, for some $m \times n$ matrix A and some vector $b \in \mathbb{R}^m$. We will also discuss the case of affine constraints later.

In the case of equality constraints, a necessary condition for a local extremum with respect to U can be given in terms of *Lagrange multipliers*. In the case of inequality constraints, there is also a necessary condition for a local extremum with respect to U in terms of generalized Lagrange multipliers and the *Karush–Kuhn–Tucker* conditions. This will be discussed in Chapter 14.

We begin by considering the case where $\Omega \subseteq E_1 \times E_2$ is an open subset of a product of normed vector spaces and where U is the zero locus of some continuous function $\varphi\colon \Omega \to E_2$, which means that

$$U = \{(u_1, u_2) \in \Omega \mid \varphi(u_1, u_2) = 0\}.$$

For the sake of brevity, we say that J has a *constrained local extremum* at u instead of saying that J has a *local extremum* at the point $u \in U$ *with respect to* U.

In most applications, we have $E_1 = \mathbb{R}^{n-m}$ and $E_2 = \mathbb{R}^m$ for some integers m, n such that $1 \leq m < n$, Ω is an open subset of $\mathbb{R}^n$, $J\colon \Omega \to \mathbb{R}$, and we have m functions $\varphi_i\colon \Omega \to \mathbb{R}$ defining the subset

$$U = \{v \in \Omega \mid \varphi_i(v) = 0,\ 1 \leq i \leq m\}.$$

Fortunately, there is a necessary condition for constrained local extrema in terms of *Lagrange multipliers*.

Theorem 4.2. *(Necessary condition for a constrained extremum in terms of Lagrange multipliers) Let Ω be an open subset of $\mathbb{R}^n$, consider m C^1-functions $\varphi_i\colon \Omega \to \mathbb{R}$ (with $1 \leq m < n$), let*

$$U = \{v \in \Omega \mid \varphi_i(v) = 0,\ 1 \leq i \leq m\},$$

and let $u \in U$ be a point such that the derivatives $d\varphi_i(u) \in \mathcal{L}(\mathbb{R}^n; \mathbb{R})$ are linearly independent; equivalently, assume that the $m \times n$ matrix $\big((\partial\varphi_i/\partial x_j)(u)\big)$ has rank m. If $J\colon \Omega \to \mathbb{R}$ is a function which is differentiable at $u \in U$ and if J has a local constrained extremum at u, then there exist m numbers $\lambda_i(u) \in \mathbb{R}$, uniquely defined, such that

$$dJ(u) + \lambda_1(u)d\varphi_1(u) + \cdots + \lambda_m(u)d\varphi_m(u) = 0;$$

equivalently,

$$\nabla J(u) + \lambda_1(u)\nabla\varphi_1(u) + \cdots + \lambda_m(u)\nabla\varphi_m(u) = 0.$$

Theorem 4.2 will be proven as a corollary of Theorem 4.4, which gives a more general formulation that applies to the situation where E_1 is an infinite-dimensional Banach space. To simplify the exposition we postpone a discussion of this theorem until we have presented several examples illustrating the method of Lagrange multipliers.

Definition 4.4. The numbers $\lambda_i(u)$ involved in Theorem 4.2 are called the *Lagrange multipliers* associated with the constrained extremum u (again, with some minor abuse of language).

The linear independence of the linear forms $d\varphi_i(u)$ is equivalent to the fact that the Jacobian matrix $\big((\partial\varphi_i/\partial x_j)(u)\big)$ of $\varphi = (\varphi_1, \ldots, \varphi_m)$ at u has rank m. If $m = 1$, the linear independence of the $d\varphi_i(u)$ reduces to the condition $\nabla\varphi_1(u) \neq 0$.

A fruitful way to reformulate the use of Lagrange multipliers is to introduce the notion of the Lagrangian associated with our constrained extremum problem.

Definition 4.5. The *Lagrangian* associated with our constrained extremum problem is the function $L\colon \Omega \times \mathbb{R}^m \to \mathbb{R}$ given by

$$L(v, \lambda) = J(v) + \lambda_1\varphi_1(v) + \cdots + \lambda_m\varphi_m(v),$$

with $\lambda = (\lambda_1, \ldots, \lambda_m)$.

We have the following simple but important proposition.

Proposition 4.3. *There exists some $\mu = (\mu_1, \ldots, \mu_m)$ and some $u \in U$ such that*

$$dJ(u) + \mu_1 d\varphi_1(u) + \cdots + \mu_m d\varphi_m(u) = 0$$

if and only if

$$dL(u, \mu) = 0,$$

or equivalently

$$\nabla L(u, \mu) = 0;$$

that is, iff (u, μ) is a critical point of the Lagrangian L.

Proof. Indeed $dL(u, \mu) = 0$ is equivalent to

$$\begin{aligned} \frac{\partial L}{\partial v}(u, \mu) &= 0 \\ \frac{\partial L}{\partial \lambda_1}(u, \mu) &= 0 \\ &\vdots \\ \frac{\partial L}{\partial \lambda_m}(u, \mu) &= 0, \end{aligned}$$

and since

$$\frac{\partial L}{\partial v}(u, \mu) = dJ(u) + \mu_1 d\varphi_1(u) + \cdots + \mu_m d\varphi_m(u)$$

and

$$\frac{\partial L}{\partial \lambda_i}(u, \mu) = \varphi_i(u),$$

we get

$$dJ(u) + \mu_1 d\varphi_1(u) + \cdots + \mu_m d\varphi_m(u) = 0$$

and

$$\varphi_1(u) = \cdots = \varphi_m(u) = 0,$$

that is, $u \in U$. The converse is proven in a similar fashion (essentially by reversing the argument). □

If we write out explicitly the condition

$$dJ(u) + \lambda_1 d\varphi_1(u) + \cdots + \lambda_m d\varphi_m(u) = 0,$$

we get the $n \times m$ system

$$\begin{aligned}
\frac{\partial J}{\partial x_1}(u) + \lambda_1 \frac{\partial \varphi_1}{\partial x_1}(u) + \cdots + \lambda_m \frac{\partial \varphi_m}{\partial x_1}(u) &= 0 \\
&\vdots \\
\frac{\partial J}{\partial x_n}(u) + \lambda_1 \frac{\partial \varphi_1}{\partial x_n}(u) + \cdots + \lambda_m \frac{\partial \varphi_m}{\partial x_n}(u) &= 0,
\end{aligned}$$

and it is important to note that the matrix of this system is the *transpose* of the Jacobian matrix of φ at u. If we write $\mathrm{Jac}(\varphi)(u) = \big((\partial \varphi_i/\partial x_j)(u)\big)$ for the Jacobian matrix of φ (at u), then the above system is written in matrix form as

$$\nabla J(u) + (\mathrm{Jac}(\varphi)(u))^\top \lambda = 0,$$

where λ is viewed as a column vector, and the Lagrangian is equal to

$$L(u, \lambda) = J(u) + (\varphi_1(u), \ldots, \varphi_m(u))\lambda.$$

The beauty of the Lagrangian is that the constraints $\{\varphi_i(v) = 0\}$ have been incorporated into the function $L(v, \lambda)$, and that the necessary condition for the existence of a constrained local extremum of J is reduced to the necessary condition for the existence of a local extremum of the *unconstrained* L.

However, one should be careful to check that the assumptions of Theorem 4.2 are satisfied (in particular, the linear independence of the linear forms $d\varphi_i$).

Example 4.1. For example, let $J\colon \mathbb{R}^3 \to \mathbb{R}$ be given by

$$J(x, y, z) = x + y + z^2$$

and $g\colon \mathbb{R}^3 \to \mathbb{R}$ by

$$g(x, y, z) = x^2 + y^2.$$

Since $g(x, y, z) = 0$ iff $x = y = 0$, we have $U = \{(0, 0, z) \mid z \in \mathbb{R}\}$ and the restriction of J to U is given by

$$J(0, 0, z) = z^2,$$

which has a minimum for $z = 0$. However, a "blind" use of Lagrange multipliers would require that there is some λ so that

$$\frac{\partial J}{\partial x}(0, 0, z) = -\lambda \frac{\partial g}{\partial x}(0, 0, z), \quad \frac{\partial J}{\partial y}(0, 0, z) = -\lambda \frac{\partial g}{\partial y}(0, 0, z), \quad \frac{\partial J}{\partial z}(0, 0, z) = -\lambda \frac{\partial g}{\partial z}(0, 0, z),$$

and since

$$\frac{\partial g}{\partial x}(x, y, z) = 2x, \quad \frac{\partial g}{\partial y}(x, y, z) = 2y, \quad \frac{\partial g}{\partial z}(0, 0, z) = 0,$$

the partial derivatives above all vanish for $x = y = 0$, so at a local extremum we should also have

$$\frac{\partial J}{\partial x}(0,0,z) = 0, \quad \frac{\partial J}{\partial y}(0,0,z) = 0, \quad \frac{\partial J}{\partial z}(0,0,z) = 0,$$

but this is absurd since

$$\frac{\partial J}{\partial x}(x,y,z) = 1, \quad \frac{\partial J}{\partial y}(x,y,z) = 1, \quad \frac{\partial J}{\partial z}(x,y,z) = 2z.$$

The reader should enjoy finding the reason for the flaw in the argument.

One should also keep in mind that Theorem 4.2 gives only a necessary condition. The (u, λ) may *not* correspond to local extrema! Thus, it is always necessary to analyze the local behavior of J near a critical point u. This is generally difficult, but in the case where J is affine or quadratic and the constraints are affine or quadratic, this is possible (although not always easy).

Example 4.2. Let us apply the above method to the following example in which $E_1 = \mathbb{R}$, $E_2 = \mathbb{R}$, $\Omega = \mathbb{R}^2$, and

$$\begin{aligned} J(x_1, x_2) &= -x_2 \\ \varphi(x_1, x_2) &= x_1^2 + x_2^2 - 1. \end{aligned}$$

Observe that

$$U = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1^2 + x_2^2 = 1\}$$

is the unit circle, and since

$$\nabla\varphi(x_1, x_2) = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix},$$

it is clear that $\nabla\varphi(x_1, x_2) \neq 0$ for every point $= (x_1, x_2)$ on the unit circle. If we form the Lagrangian

$$L(x_1, x_2, \lambda) = -x_2 + \lambda(x_1^2 + x_2^2 - 1),$$

Theorem 4.2 says that a necessary condition for J to have a constrained local extremum is that $\nabla L(x_1, x_2, \lambda) = 0$, so the following equations must hold:

$$\begin{aligned} 2\lambda x_1 &= 0 \\ -1 + 2\lambda x_2 &= 0 \\ x_1^2 + x_2^2 &= 1. \end{aligned}$$

The second equation implies that $\lambda \neq 0$, and then the first yields $x_1 = 0$, so the third yields $x_2 = \pm 1$, and we get two solutions:

$$\begin{aligned} \lambda &= \frac{1}{2}, & (x_1, x_2) &= (0, 1) \\ \lambda &= -\frac{1}{2}, & (x_1', x_2') &= (0, -1). \end{aligned}$$

We can check immediately that the first solution is a minimum and the second is a maximum. The reader should look for a geometric interpretation of this problem.

Example 4.3. Let us now consider the case in which J is a quadratic function of the form

$$J(v) = \frac{1}{2}v^\top Av - v^\top b,$$

where A is an $n \times n$ symmetric matrix, $b \in \mathbb{R}^n$, and the constraints are given by a linear system of the form

$$Cv = d,$$

where C is an $m \times n$ matrix with $m < n$ and $d \in \mathbb{R}^m$. We also assume that C has rank m. In this case, the function φ is given by

$$\varphi(v) = (Cv - d)^\top,$$

because we view $\varphi(v)$ as a row vector (and v as a column vector), and by Example 3.3,

$$d\varphi(u)(w) = (Cw)^\top,$$

so the condition that the Jacobian matrix of φ at u have rank m is satisfied because the range of $d\varphi(u)$ is spanned by the row vectors transpose of the columns of C, which form a matrix of rank m. The Lagrangian of this problem is

$$L(v, \lambda) = \frac{1}{2}v^\top Av - v^\top b + (Cv - d)^\top \lambda = \frac{1}{2}v^\top Av - v^\top b + v^\top C^\top \lambda - d^\top \lambda,$$

where λ is viewed as a column vector. Now recall that because A is a symmetric matrix, it was shown in Example 3.9 that

$$\nabla L_{(v,\lambda)} = \begin{pmatrix} Av - b + C^\top \lambda \\ Cv - d \end{pmatrix}.$$

Therefore, the necessary condition for constrained local extrema is

$$\begin{aligned} Av + C^\top \lambda &= b \\ Cv &= d, \end{aligned}$$

which can be expressed in matrix form as

$$\begin{pmatrix} A & C^\top \\ C & 0 \end{pmatrix} \begin{pmatrix} v \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix},$$

where the matrix of the system is a symmetric matrix. We should not be surprised to find the system discussed later in Chapter 6, except for some renaming of the matrices and vectors involved. As we will show in Section 6.2, the function J has a minimum iff A is positive definite, so in general, if A is only a symmetric matrix, the critical points of the Lagrangian do *not* correspond to extrema of J.

Remark: If the Jacobian matrix $\mathrm{Jac}(\varphi)(v) = \big((\partial\varphi_i/\partial x_j)(v)\big)$ has rank m for all $v \in U$ (which is equivalent to the linear independence of the linear forms $d\varphi_i(v)$), then we say that $0 \in \mathbb{R}^m$ is a *regular value* of φ. In this case, it is known that

$$U = \{v \in \Omega \mid \varphi(v) = 0\}$$

is a *smooth submanifold of dimension* $n - m$ *of* $\mathbb{R}^n$. Furthermore, the set

$$T_vU = \{w \in \mathbb{R}^n \mid d\varphi_i(v)(w) = 0,\ 1 \leq i \leq m\} = \bigcap_{i=1}^{m} \mathrm{Ker}\, d\varphi_i(v)$$

is the *tangent space* to U at v (a vector space of dimension $n - m$). Then, the condition

$$dJ(v) + \mu_1 d\varphi_1(v) + \cdots + \mu_m d\varphi_m(v) = 0$$

implies that $dJ(v)$ vanishes on the tangent space T_vU. Conversely, if $dJ(v)(w) = 0$ for all $w \in T_vU$, this means that $dJ(v)$ is orthogonal (in the sense of Definition 10.3 (Vol. I)) to T_vU. Since (by Theorem 10.4(b) (Vol. I)) the orthogonal of T_vU is the space of linear forms spanned by $d\varphi_1(v), \ldots, d\varphi_m(v)$, it follows that $dJ(v)$ must be a linear combination of the $d\varphi_i(v)$. Therefore, when 0 is a regular value of φ, Theorem 4.2 asserts that if $u \in U$ is a local extremum of J, then $dJ(u)$ must vanish on the tangent space T_uU. We can say even more. The subset $Z(J(u))$ of Ω given by

$$Z(J(u)) = \{v \in \Omega \mid J(v) = J(u)\}$$

(the *level set of level* $J(u)$) is a hypersurface in Ω, and if $dJ(u) \neq 0$, the zero locus of $dJ(u)$ is the tangent space $T_uZ(J(u))$ to $Z(J(u))$ at u (a vector space of dimension $n - 1$), where

$$T_uZ(J(u)) = \{w \in \mathbb{R}^n \mid dJ(u)(w) = 0\}.$$

Consequently, Theorem 4.2 asserts that

$$T_uU \subseteq T_uZ(J(u));$$

this is a geometric condition.

We now return to the general situation where E_1 and E_2 may be infinite-dimensional normed vector spaces (with E_1 a Banach space) and we state and prove the following general result about the method of Lagrange multipliers.

Theorem 4.4. *(Necessary condition for a constrained extremum) Let $\Omega \subseteq E_1 \times E_2$ be an open subset of a product of normed vector spaces, with E_1 a Banach space (E_1 is complete), let $\varphi\colon \Omega \to E_2$ be a C^1-function (which means that $d\varphi(\omega)$ exists and is continuous for all $\omega \in \Omega$), and let*

$$U = \{(u_1, u_2) \in \Omega \mid \varphi(u_1, u_2) = 0\}.$$

Moreover, let $u = (u_1, u_2) \in U$ *be a point such that*

$$\frac{\partial \varphi}{\partial x_2}(u_1, u_2) \in \mathcal{L}(E_2; E_2) \quad \text{and} \quad \left(\frac{\partial \varphi}{\partial x_2}(u_1, u_2)\right)^{-1} \in \mathcal{L}(E_2; E_2),$$

and let $J\colon \Omega \to \mathbb{R}$ *be a function which is differentiable at* u*. If* J *has a constrained local extremum at* u*, then there is a continuous linear form* $\Lambda(u) \in \mathcal{L}(E_2; \mathbb{R})$ *such that*

$$dJ(u) + \Lambda(u) \circ d\varphi(u) = 0.$$

Proof. The plan of attack is to use the implicit function theorem; Theorem 3.13. Observe that the assumptions of Theorem 3.13 are indeed met. Therefore, there exist some open subsets $U_1 \subseteq E_1$, $U_2 \subseteq E_2$, and a continuous function $g\colon U_1 \to U_2$ with $(u_1, u_2) \in U_1 \times U_2 \subseteq \Omega$ and such that

$$\varphi(v_1, g(v_1)) = 0$$

for all $v_1 \in U_1$. Moreover, g is differentiable at $u_1 \in U_1$ and

$$dg(u_1) = -\left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_1}(u).$$

It follows that the restriction of J to $(U_1 \times U_2) \cap U$ yields a function G of a single variable, with

$$G(v_1) = J(v_1, g(v_1))$$

for all $v_1 \in U_1$. Now the function G is differentiable at u_1 and it has a local extremum at u_1 on U_1, so Proposition 4.1 implies that

$$dG(u_1) = 0.$$

By the chain rule,

$$\begin{aligned} dG(u_1) &= \frac{\partial J}{\partial x_1}(u) + \frac{\partial J}{\partial x_2}(u) \circ dg(u_1) \\ &= \frac{\partial J}{\partial x_1}(u) - \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_1}(u). \end{aligned}$$

From $dG(u_1) = 0$, we deduce

$$\frac{\partial J}{\partial x_1}(u) = \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_1}(u),$$

and since we also have

$$\frac{\partial J}{\partial x_2}(u) = \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_2}(u),$$

if we let

$$\Lambda(u) = -\frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1},$$

then we get

$$\begin{aligned} dJ(u) &= \frac{\partial J}{\partial x_1}(u) + \frac{\partial J}{\partial x_2}(u) \\ &= \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \left(\frac{\partial \varphi}{\partial x_1}(u) + \frac{\partial \varphi}{\partial x_2}(u)\right) \\ &= -\Lambda(u) \circ d\varphi(u), \end{aligned}$$

which yields $dJ(u) + \Lambda(u) \circ d\varphi(u) = 0$, as claimed. □

Finally, we prove Theorem 4.2.

Proof of Theorem 4.2. The linear independence of the m linear forms $d\varphi_i(u)$ is equivalent to the fact that the $m \times n$ matrix $A = \big((\partial \varphi_i/\partial x_j)(u)\big)$ has rank m. By reordering the columns, we may assume that the first m columns are linearly independent. To conform to the set-up of Theorem 4.4 we define E_1 and E_2 as

$$E_1 = \left\{ \sum_{i=m+1}^{n} v_i e_i \mid (v_{m+1}, \ldots, v_n) \in \mathbb{R}^{n-m} \right\}, \quad E_2 = \left\{ \sum_{i=1}^{m} v_i e_i \mid (v_1, \ldots, v_m) \in \mathbb{R}^m \right\}.$$

If we let $\psi \colon \Omega \to \mathbb{R}^m$ be the function defined by

$$\psi(v_{m+1}, \ldots, v_n, v_1, \ldots, v_m) = (\varphi_1(v), \ldots, \varphi_m(v))$$

for all $(v_{m+1}, \ldots, v_n, v_1, \ldots, v_m) \in \Omega$, with $v = (v_1, \ldots, v_n)$, then we see that $\partial \psi/\partial x_2(u)$ is invertible and both $\partial \psi/\partial x_2(u)$ and its inverse are continuous, so that Theorem 4.4 applies, and there is some (continuous) linear form $\Lambda(u) \in \mathcal{L}(\mathbb{R}^m; \mathbb{R})$ such that

$$dJ(u) + \Lambda(u) \circ d\psi(u_{m+1}, \ldots, u_n, u_1, \ldots, u_m) = 0,$$

namely

$$dJ(u) + \Lambda(u) \circ d\varphi(u) = 0.$$

However, $\Lambda(u)$ is defined by some m-tuple $(\lambda_1(u), \ldots, \lambda_m(u)) \in \mathbb{R}^m$, and in view of the definition of φ, the above equation is equivalent to

$$dJ(u) + \lambda_1(u) d\varphi_1(u) + \cdots + \lambda_m(u) d\varphi_m(u) = 0.$$

The uniqueness of the $\lambda_i(u)$ is a consequence of the linear independence of the $d\varphi_i(u)$. □

We now investigate conditions for the existence of extrema involving the second derivative of J.

4.2 Using Second Derivatives to Find Extrema

For the sake of brevity, we consider only the case of local minima; analogous results are obtained for local maxima (replace J by $-J$, since $\max_u J(u) = -\min_u -J(u)$). We begin with a necessary condition for an unconstrained local minimum.

Proposition 4.5. *Let E be a normed vector space and let $J: \Omega \to \mathbb{R}$ be a function, with Ω some open subset of E. If the function J is differentiable in Ω, if J has a second derivative $\mathrm{D}^2J(u)$ at some point $u \in \Omega$, and if J has a local minimum at u, then*

$$\mathrm{D}^2J(u)(w, w) \geq 0 \quad \textit{for all } w \in E.$$

Proof. Pick any nonzero vector $w \in E$. Since Ω is open, for t small enough, $u + tw \in \Omega$ and $J(u + tw) \geq J(u)$, so there is some open interval $I \subseteq \mathbb{R}$ such that

$$u + tw \in \Omega \quad \text{and} \quad J(u + tw) \geq J(u)$$

for all $t \in I$. Using the Taylor–Young formula and the fact that we must have $dJ(u) = 0$ since J has a local minimum at u, we get

$$0 \leq J(u + tw) - J(u) = \frac{t^2}{2}\mathrm{D}^2J(u)(w, w) + t^2 \|w\|^2 \epsilon(tw),$$

with $\lim_{t \mapsto 0} \epsilon(tw) = 0$, which implies that

$$\mathrm{D}^2J(u)(w, w) \geq 0.$$

Since the argument holds for all $w \in E$ (trivially if $w = 0$), the proposition is proven. □

One should be cautioned that there is no converse to the previous proposition. For example, the function $f: x \mapsto x^3$ has no local minimum at 0, yet $df(0) = 0$ and $\mathrm{D}^2f(0)(u, v) = 0$. Similarly, the reader should check that the function $f: \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x, y) = x^2 - 3y^3$$

has no local minimum at $(0, 0)$; yet $df(0, 0) = 0$ since $df(x, y) = (2x, -9y^2)$, and for $u = (u_1, u_2)$, $\mathrm{D}^2f(0, 0)(u, u) = 2u_1^2 \geq 0$ since

$$D^2f(x, y)(u, u) = \begin{pmatrix} u_1 & u_2 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & -18y \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}.$$

See Figure 4.3.

When $E = \mathbb{R}^n$, Proposition 4.5 says that a necessary condition for having a local minimum is that the Hessian $\nabla^2 J(u)$ be positive semidefinite (it is always symmetric).

We now give sufficient conditions for the existence of a local minimum.

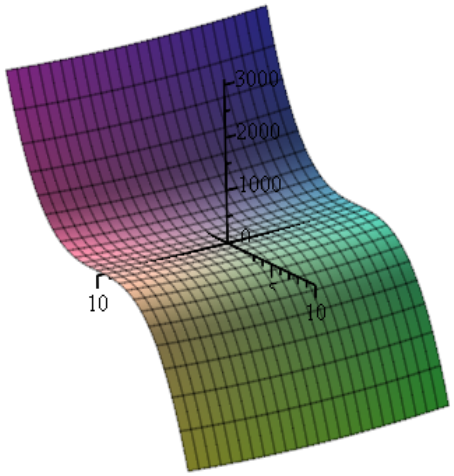


Figure 4.3: The graph of $f(x,y) = x^2 - 3y^3$. Note that $(0,0)$ not a local extremum despite the fact that $df(0,0) = 0$.

Theorem 4.6. *Let E be a normed vector space, let $J\colon \Omega \to \mathbb{R}$ be a function with Ω some open subset of E, and assume that J is differentiable in Ω and that $dJ(u) = 0$ at some point $u \in \Omega$. The following properties hold:*

(1) If $\mathrm{D}^2 J(u)$ exists and if there is some number $\alpha \in \mathbb{R}$ such that $\alpha > 0$ and

$$\mathrm{D}^2 J(u)(w,w) \geq \alpha \|w\|^2 \quad \text{for all } w \in E,$$

then J has a strict local minimum at u.

(2) If $\mathrm{D}^2 J(v)$ exists for all $v \in \Omega$ and if there is a ball $B \subseteq \Omega$ centered at u such that

$$\mathrm{D}^2 J(v)(w,w) \geq 0 \quad \text{for all } v \in B \text{ and all } w \in E,$$

then J has a local minimum at u.

Proof. (1) Using the formula of Taylor–Young, for every vector w small enough, we can write

$$\begin{aligned} J(u+w) - J(u) &= \frac{1}{2}\mathrm{D}^2 J(u)(w,w) + \|w\|^2 \epsilon(w) \\ &\geq \left(\frac{1}{2}\alpha + \epsilon(w)\right) \|w\|^2 \end{aligned}$$

with $\lim_{w \mapsto 0} \epsilon(w) = 0$. Consequently if we pick $r > 0$ small enough that $|\epsilon(w)| < \alpha/2$ for all w with $\|w\| < r$, then $J(u+w) > J(u)$ for all $u + w \in B$, where B is the open ball of center u and radius r. This proves that J has a local strict minimum at u.

(2) The formula of Taylor–Maclaurin shows that for all $u + w \in B$, we have

$$J(u+w) = J(u) + \frac{1}{2}\mathrm{D}^2 J(v)(w,w) \geq J(u),$$

for some $v \in (u, u+w)$ (recall that $(u, u+w) = \{(1-\lambda)(u+w)+\lambda(u+w) \mid 0 < \lambda < 1\}$). □

There are no converses of the two assertions of Theorem 4.6. However, there is a condition on $\mathrm{D}^2 J(u)$ that implies the condition of Part (1). Since this condition is easier to state when $E = \mathbb{R}^n$, we begin with this case.

Recall that a $n \times n$ symmetric matrix A is *positive definite* if $x^\top Ax > 0$ for all $x \in \mathbb{R}^n - \{0\}$. In particular, A must be invertible.

Proposition 4.7. *For any symmetric matrix A, if A is positive definite, then there is some $\alpha > 0$ such that*

$$x^\top Ax \geq \alpha \|x\|^2 \quad \textit{for all } x \in \mathbb{R}^n.$$

Proof. Pick any norm in $\mathbb{R}^n$ (recall that all norms on $\mathbb{R}^n$ are equivalent). Since the unit sphere $S^{n-1} = \{x \in \mathbb{R}^n \mid \|x\| = 1\}$ is compact and since the function $f(x) = x^\top Ax$ is never zero on S^{n-1}, the function f has a minimum $\alpha > 0$ on S^{n-1}. Using the usual trick that $x = \|x\|\,(x/\|x\|)$ for every nonzero vector $x \in \mathbb{R}^n$ and the fact that the inequality of the proposition is trivial for $x = 0$, from

$$x^\top Ax \geq \alpha \quad \text{for all } x \text{ with } \|x\| = 1,$$

we get

$$x^\top Ax \geq \alpha \|x\|^2 \quad \text{for all } x \in \mathbb{R}^n,$$

as claimed. □

We can combine Theorem 4.6 and Proposition 4.7 to obtain a useful sufficient condition for the existence of a strict local minimum. First let us introduce some terminology.

Definition 4.6. Given a function $J\colon \Omega \to \mathbb{R}$ as before, say that a point $u \in \Omega$ is a *nondegenerate critical point* if $dJ(u) = 0$ and if the Hessian matrix $\nabla^2 J(u)$ is invertible.

Proposition 4.8. *Let $J\colon \Omega \to \mathbb{R}$ be a function defined on some open subset $\Omega \subseteq \mathbb{R}^n$. If J is differentiable in Ω and if some point $u \in \Omega$ is a nondegenerate critical point such that $\nabla^2 J(u)$ is positive definite, then J has a strict local minimum at u.*

Remark: It is possible to generalize Proposition 4.8 to infinite-dimensional spaces by finding a suitable generalization of the notion of a nondegenerate critical point. Firstly, we assume that E is a Banach space (a complete normed vector space). Then we define the dual E' of E as the set of continuous linear forms on E, so that $E' = \mathcal{L}(E; \mathbb{R})$. Following Lang, we use

the notation E' for the space of continuous linear forms to avoid confusion with the space $E^* = \mathrm{Hom}(E, \mathbb{R})$ of all linear maps from E to $\mathbb{R}$. A continuous bilinear map $\varphi\colon E \times E \to \mathbb{R}$ in $\mathcal{L}_2(E, E; \mathbb{R})$ yields a map Φ from E to E' given by

$$\Phi(u) = \varphi_u,$$

where $\varphi_u \in E'$ is the linear form defined by

$$\varphi_u(v) = \varphi(u, v).$$

It is easy to check that φ_u is continuous and that the map Φ is continuous. Then we say that φ is *nondegenerate* iff $\Phi\colon E \to E'$ is an isomorphism of Banach spaces, which means that Φ is invertible and that both Φ and Φ^{-1} are continuous linear maps. Given a function $J\colon \Omega \to \mathbb{R}$ differentiable on Ω as before (where Ω is an open subset of E), if $\mathrm{D}^2 J(u)$ exists for some $u \in \Omega$, we say that u is a *nondegenerate critical point* if $dJ(u) = 0$ and if $\mathrm{D}^2 J(u)$ is nondegenerate. Of course, $\mathrm{D}^2 J(u)$ is positive definite if $\mathrm{D}^2 J(u)(w, w) > 0$ for all $w \in E - \{0\}$.

Using the above definition, Proposition 4.7 can be generalized to a nondegenerate positive definite bilinear form (on a Banach space) and Theorem 4.8 can also be generalized to the situation where $J\colon \Omega \to \mathbb{R}$ is defined on an open subset of a Banach space. For details and proofs, see Cartan [21] (Part I Chapter 8) and Avez [5] (Chapter 8 and Chapter 10).

In the next section we make use of convexity; both on the domain Ω and on the function J itself.

4.3 Using Convexity to Find Extrema

We begin by reviewing the definition of a convex set and of a convex function.

Definition 4.7. Given any real vector space E, we say that a subset C of E is *convex* if either $C = \emptyset$ or if for every pair of points $u, v \in C$, the line segment connecting u and v is contained in C, i.e.,

$$(1 - \lambda)u + \lambda v \in C \quad \text{for all } \lambda \in \mathbb{R} \text{ such that } 0 \leq \lambda \leq 1.$$

Given any two points $u, v \in E$, the *line segment* $[u, v]$ is the set

$$[u, v] = \{(1 - \lambda)u + \lambda v \in E \mid \lambda \in \mathbb{R},\ 0 \leq \lambda \leq 1\}.$$

Clearly, a nonempty set C is convex iff $[u, v] \subseteq C$ whenever $u, v \in C$. See Figure 4.4 for an example of a convex set.

Definition 4.8. If C is a nonempty convex subset of E, a function $f\colon C \to \mathbb{R}$ is *convex* (on C) if for every pair of points $u, v \in C$,

$$f((1 - \lambda)u + \lambda v) \leq (1 - \lambda)f(u) + \lambda f(v) \quad \text{for all } \lambda \in \mathbb{R} \text{ such that } 0 \leq \lambda \leq 1;$$

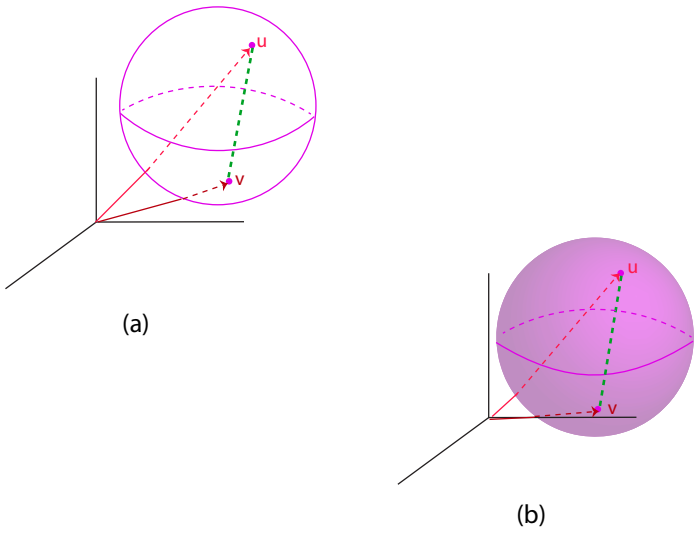


Figure 4.4: Figure (a) shows that a sphere is not convex in $\mathbb{R}^3$ since the dashed green line does not lie on its surface. Figure (b) shows that a solid ball is convex in $\mathbb{R}^3$.

the function f is *strictly convex* (on C) if for every pair of distinct points $u, v \in C$ $(u \neq v)$,

$$f((1-\lambda)u + \lambda v) < (1-\lambda)f(u) + \lambda f(v) \quad \text{for all } \lambda \in \mathbb{R} \text{ such that } 0 < \lambda < 1;$$

see Figure 4.5. The *epigraph*[1] $\mathbf{epi}(f)$ of a function $f\colon A \to \mathbb{R}$ defined on some subset A of $\mathbb{R}^n$ is the subset of $\mathbb{R}^{n+1}$ defined as

$$\mathbf{epi}(f) = \{(x, y) \in \mathbb{R}^{n+1} \mid f(x) \leq y,\ x \in A\}.$$

A function $f\colon C \to \mathbb{R}$ defined on a convex subset C is *concave* (resp. *strictly concave*) if $(-f)$ is convex (resp. strictly convex).

It is obvious that a function f is convex iff its epigraph $\mathbf{epi}(f)$ is a convex subset of $\mathbb{R}^{n+1}$.

Example 4.4. Here are some common examples of convex sets.

- Subspaces $V \subseteq E$ of a vector space E are convex.
- *Affine subspaces*, that is, sets of the form $u + V$, where V is a subspace of E and $u \in E$, are convex.
- Balls (open or closed) are convex. Given any linear form $\varphi\colon E \to \mathbb{R}$, for any scalar $c \in \mathbb{R}$, the *closed half–spaces*

$$H^+_{\varphi,c} = \{u \in E \mid \varphi(u) \geq c\}, \qquad H^-_{\varphi,c} = \{u \in E \mid \varphi(u) \leq c\},$$

are convex.

[1] "Epi" means above.

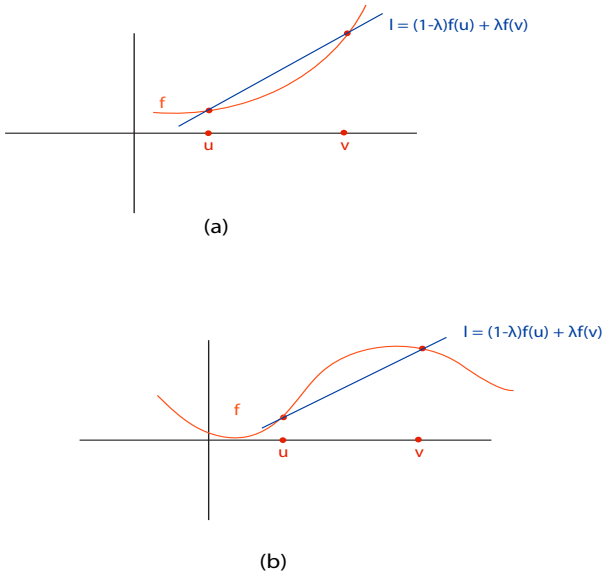


Figure 4.5: Figures (a) and (b) are the graphs of real valued functions. Figure (a) is the graph of convex function since the blue line lies above the graph of f. Figure (b) shows the graph of a function which is not convex.

- Any intersection of half–spaces is convex.
- More generally, any intersection of convex sets is convex.

Example 4.5. Here are some common examples of convex and concave functions.

- Linear forms are convex functions (but not strictly convex).
- Any norm $\| \| : E \to \mathbb{R}_+$ is a convex function.
- The max function,
$$\max(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$$
is convex on $\mathbb{R}^n$.
- The exponential $x \mapsto e^{cx}$ is strictly convex for any $c \neq 0$ $(c \in \mathbb{R})$.
- The logarithm function is concave on $\mathbb{R}_+ - \{0\}$.
- The *log-determinant function* log det is concave on the set of symmetric positive definite matrices. This function plays an important role in convex optimization.

An excellent exposition of convexity and its applications to optimization can be found in Boyd [18].

Here is a necessary condition for a function to have a local minimum with respect to a convex subset U.

Theorem 4.9. *(Necessary condition for a local minimum on a convex subset) Let $J: \Omega \to \mathbb{R}$ be a function defined on some open subset Ω of a normed vector space E and let $U \subseteq \Omega$ be a nonempty convex subset. Given any $u \in U$, if $dJ(u)$ exists and if J has a local minimum in u with respect to U, then*

$$dJ(u)(v-u) \geq 0 \quad \textit{for all } v \in U.$$

Proof. Let $v = u + w$ be an arbitrary point in U. Since U is convex, we have $u + tw \in U$ for all t such that $0 \leq t \leq 1$. Since $dJ(u)$ exists, we can write

$$J(u+tw) - J(u) = dJ(u)(tw) + \|tw\|\,\epsilon(tw)$$

with $\lim_{t \mapsto 0} \epsilon(tw) = 0$. However, because $0 \leq t$,

$$J(u+tw) - J(u) = t(dJ(u)(w) + \|w\|\,\epsilon(tw))$$

and since u is a local minimum with respect to U, we have $J(u+tw) - J(u) \geq 0$, so we get

$$t(dJ(u)(w) + \|w\|\,\epsilon(tw)) \geq 0.$$

The above implies that $dJ(u)(w) \geq 0$, because otherwise we could pick $t > 0$ small enough so that

$$dJ(u)(w) + \|w\|\,\epsilon(tw) < 0,$$

a contradiction. Since the argument holds for all $v = u + w \in U$, the theorem is proven. □

Observe that the convexity of U is a substitute for the use of Lagrange multipliers, but we now have to deal with an *inequality* instead of an equality.

In the special case where U is a subspace of E we have the following result.

Corollary 4.10. *With the same assumptions as in Theorem 4.9, if U is a subspace of E, if $dJ(u)$ exists and if J has a local minimum in u with respect to U, then*

$$dJ(u)(w) = 0 \quad \textit{for all } w \in U.$$

Proof. In this case since $u \in U$ we have $2u \in U$, and for any $u + w \in U$, we must have $2u - (u+w) = u - w \in U$. The previous theorem implies that $dJ(u)(w) \geq 0$ and $dJ(u)(-w) \geq 0$, that is, $dJ(u)(w) \leq 0$, so $dJ(u) = 0$. Since the argument holds for $w \in U$ (because U is a subspace, if $u, w \in U$, then $u + w \in U$), we conclude that

$$dJ(u)(w) = 0 \quad \text{for all } w \in U.$$ □

We will now characterize convex functions when they have a first derivative or a second derivative.

Proposition 4.11. *(Convexity and first derivative) Let $f\colon \Omega \to \mathbb{R}$ be a function differentiable on some open subset Ω of a normed vector space E and let $U \subseteq \Omega$ be a nonempty convex subset.*

(1) The function f is convex on U iff

$$f(v) \geq f(u) + df(u)(v-u) \quad \textit{for all } u, v \in U.$$

(2) The function f is strictly convex on U iff

$$f(v) > f(u) + df(u)(v-u) \quad \textit{for all } u, v \in U \textit{ with } u \neq v.$$

See Figure 4.6.

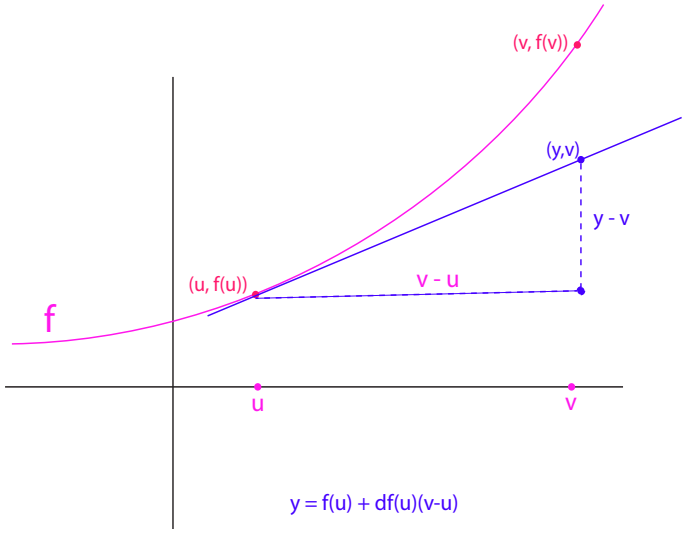


Figure 4.6: An illustration of a convex valued function f. Since f is convex it always lies above its tangent line.

Proof. Let $u, v \in U$ be any two distinct points and pick $\lambda \in \mathbb{R}$ with $0 < \lambda < 1$. If the function f is convex, then

$$f((1-\lambda)u + \lambda v) \leq (1-\lambda)f(u) + \lambda f(v),$$

which yields

$$\frac{f((1-\lambda)u + \lambda v) - f(u)}{\lambda} \leq f(v) - f(u).$$

It follows that

$$df(u)(v-u) = \lim_{\lambda \mapsto 0} \frac{f((1-\lambda)u + \lambda v) - f(u)}{\lambda} \leq f(v) - f(u).$$

If f is strictly convex, the above reasoning does not work, because a strict inequality is not necessarily preserved by "passing to the limit." We have recourse to the following trick: for any ω such that $0 < \omega < 1$, observe that

$$(1-\lambda)u + \lambda v = u + \lambda(v-u) = \frac{\omega - \lambda}{\omega}u + \frac{\lambda}{\omega}(u + \omega(v-u)).$$

If we assume that $0 < \lambda \leq \omega$, the convexity of f yields

$$f(u+\lambda(v-u)) = f\left(\left(1-\frac{\lambda}{\omega}\right)u + \frac{\lambda}{\omega}(u+\omega(v-u))\right) \leq \frac{\omega-\lambda}{\omega}f(u) + \frac{\lambda}{\omega}f(u+\omega(v-u)).$$

If we subtract $f(u)$ to both sides, we get

$$\frac{f(u+\lambda(v-u)) - f(u)}{\lambda} \leq \frac{f(u+\omega(v-u)) - f(u)}{\omega}.$$

Now since $0 < \omega < 1$ and f is strictly convex,

$$f(u+\omega(v-u)) = f((1-\omega)u + \omega v) < (1-\omega)f(u) + \omega f(v),$$

which implies that

$$\frac{f(u+\omega(v-u)) - f(u)}{\omega} < f(v) - f(u),$$

and thus we get

$$\frac{f(u+\lambda(v-u)) - f(u)}{\lambda} \leq \frac{f(u+\omega(v-u)) - f(u)}{\omega} < f(v) - f(u).$$

If we let λ go to 0, by passing to the limit we get

$$df(u)(v-u) \leq \frac{f(u+\omega(v-u)) - f(u)}{\omega} < f(v) - f(u),$$

which yields the desired strict inequality.

Let us now consider the converse of (1); that is, assume that

$$f(v) \geq f(u) + df(u)(v-u) \quad \text{for all } u, v \in U.$$

For any two distinct points $u, v \in U$ and for any λ with $0 < \lambda < 1$, we get

$$\begin{aligned} f(v) &\geq f(v+\lambda(u-v)) - \lambda df(v+\lambda(u-v))(u-v) \\ f(u) &\geq f(v+\lambda(u-v)) + (1-\lambda)df(v+\lambda(u-v))(u-v), \end{aligned}$$

and if we multiply the first inequality by $1-\lambda$ and the second inequality by λ and them add up the resulting inequalities, we get

$$(1-\lambda)f(v) + \lambda f(u) \geq f(v+\lambda(u-v)) = f((1-\lambda)v + \lambda u),$$

which proves that f is convex.

The proof of the converse of (2) is similar, except that the inequalities are replaced by strict inequalities. □

We now establish a convexity criterion using the second derivative of f. This criterion is often easier to check than the previous one.

Proposition 4.12. *(Convexity and second derivative) Let $f\colon \Omega \to \mathbb{R}$ be a function twice differentiable on some open subset Ω of a normed vector space E and let $U \subseteq \Omega$ be a nonempty convex subset.*

(1) The function f is convex on U iff

$$\mathrm{D}^2 f(u)(v-u, v-u) \geq 0 \quad \textit{for all } u, v \in U.$$

(2) If

$$\mathrm{D}^2 f(u)(v-u, v-u) > 0 \quad \textit{for all } u, v \in U \textit{ with } u \neq v,$$

then f is strictly convex.

Proof. First assume that the inequality in Condition (1) is satisfied. For any two distinct points $u, v \in U$, the formula of Taylor–Maclaurin yields

$$\begin{aligned} f(v) - f(u) - df(u)(v-u) &= \frac{1}{2}\mathrm{D}^2 f(w)(v-u, v-u) \\ &= \frac{\rho^2}{2}\mathrm{D}^2 f(w)(v-w, v-w), \end{aligned}$$

for some $w = (1-\lambda)u + \lambda v = u + \lambda(v-u)$ with $0 < \lambda < 1$, and with $\rho = 1/(1-\lambda) > 0$, so that $v - u = \rho(v-w)$. Since $\mathrm{D}^2 f(w)(v-w, v-w) \geq 0$ for all $u, w \in U$, we conclude by applying Proposition 4.11(1).

Similarly, if (2) holds, the above reasoning and Proposition 4.11(2) imply that f is strictly convex.

To prove the necessary condition in (1), define $g\colon \Omega \to \mathbb{R}$ by

$$g(v) = f(v) - df(u)(v),$$

where $u \in U$ is any point considered fixed. If f is convex, since

$$g(v) - g(u) = f(v) - f(u) - df(u)(v-u),$$

Proposition 4.11 implies that $f(v) - f(u) - df(u)(v-u) \geq 0$, which implies that g has a local minimum at u with respect to all $v \in U$. Therefore, we have $dg(u) = 0$. Observe that g is twice differentiable in Ω and $\mathrm{D}^2 g(u) = \mathrm{D}^2 f(u)$, so the formula of Taylor–Young yields for every $v = u + w \in U$ and all t with $0 \leq t \leq 1$,

$$\begin{aligned} 0 \leq g(u+tw) - g(u) &= \frac{t^2}{2}\mathrm{D}^2 f(u)(tw, tw) + \|tw\|^2 \epsilon(tw) \\ &= \frac{t^2}{2}(\mathrm{D}^2 f(u)(w, w) + 2\|w\|^2 \epsilon(wt)), \end{aligned}$$

with $\lim_{t \mapsto 0} \epsilon(wt) = 0$, and for t small enough, we must have $\mathrm{D}^2 f(u)(w, w) \geq 0$, as claimed. □

The converse of Proposition 4.12 (2) is false as we see by considering the strictly convex function f given by $f(x) = x^4$ and its second derivative at $x = 0$.

Example 4.6. On the other hand, if f is a quadratic function of the form

$$f(u) = \frac{1}{2}u^\top Au - u^\top b$$

where A is a symmetric matrix, we know that

$$df(u)(v) = v^\top(Au - b),$$

so

$$\begin{aligned} f(v) - f(u) - df(u)(v-u) &= \frac{1}{2}v^\top Av - v^\top b - \frac{1}{2}u^\top Au + u^\top b - (v-u)^\top(Au - b) \\ &= \frac{1}{2}v^\top Av - \frac{1}{2}u^\top Au - (v-u)^\top Au \\ &= \frac{1}{2}v^\top Av + \frac{1}{2}u^\top Au - v^\top Au \\ &= \frac{1}{2}(v-u)^\top A(v-u). \end{aligned}$$

Therefore, Proposition 4.11 implies that A is positive semidefinite iff f is convex and A is positive definite iff f is strictly convex.

We conclude this section by applying our previous theorems to convex functions defined on convex subsets. In this case local minima (resp. local maxima) are global minima (resp. global maxima). The next definition is the special case of Definition 4.1 in which $W = E$ but it does not hurt to state it explicitly.

Definition 4.9. Let $f\colon E \to \mathbb{R}$ be any function defined on some normed vector space (or more generally, any set). For any $u \in E$, we say that f has a *minimum* in u (resp. *maximum* in u) if

$$f(u) \leq f(v) \text{ (resp. } f(u) \geq f(v)) \quad \text{for all } v \in E.$$

We say that f has a *strict minimum* in u (resp. *strict maximum* in u) if

$$f(u) < f(v) \text{ (resp. } f(u) > f(v)) \quad \text{for all } v \in E - \{u\}.$$

If $U \subseteq E$ is a subset of E and $u \in U$, we say that f has a *minimum* in u (resp. *strict minimum* in u) *with respect to* U if

$$f(u) \leq f(v) \quad \text{for all } v \in U \quad (\text{resp. } f(u) < f(v) \quad \text{for all } v \in U - \{u\}),$$

and similarly for a *maximum* in u (resp. *strict maximum* in u) *with respect to* U with $\leq$ changed to $\geq$ and $<$ to $>$.

Sometimes, we say *global* maximum (or minimum) to stress that a maximum (or a minimum) is not simply a local maximum (or minimum).

Theorem 4.13. *Given any normed vector space E, let U be any nonempty convex subset of E.*

(1) For any convex function $J: U \to \mathbb{R}$, for any $u \in U$, if J has a local minimum at u in U, then J has a (global) minimum at u in U.

(2) Any strict convex function $J: U \to \mathbb{R}$ has at most one minimum (in U), and if it does, then it is a strict minimum (in U).

(3) Let $J: \Omega \to \mathbb{R}$ be any function defined on some open subset Ω of E with $U \subseteq \Omega$ and assume that J is convex on U. For any point $u \in U$, if $dJ(u)$ exists, then J has a minimum in u with respect to U iff

$$dJ(u)(v-u) \geq 0 \quad \textit{for all } v \in U.$$

(4) If the convex subset U in (3) is open, then the above condition is equivalent to

$$dJ(u) = 0.$$

Proof. (1) Let $v = u + w$ be any arbitrary point in U. Since J is convex, for all t with $0 \leq t \leq 1$, we have

$$J(u+tw) = J(u+t(v-u)) = J((1-t)u+tv) \leq (1-t)J(u) + tJ(v),$$

which yields

$$J(u+tw) - J(u) \leq t(J(v) - J(u)).$$

Because J has a local minimum at u, there is some t_0 with $0 < t_0 < 1$ such that

$$0 \leq J(u+t_0w) - J(u) \leq t_0(J(v) - J(u)),$$

which implies that $J(v) - J(u) \geq 0$.

(2) If J is strictly convex, the above reasoning with $w \neq 0$ shows that there is some t_0 with $0 < t_0 < 1$ such that

$$0 \leq J(u+t_0w) - J(u) < t_0(J(v) - J(u)),$$

which shows that u is a strict global minimum (in U), and thus that it is unique.

(3) We already know from Theorem 4.9 that the condition $dJ(u)(v-u) \geq 0$ for all $v \in U$ is necessary (even if J is not convex). Conversely, because J is convex, careful inspection of the proof of Part (1) of Proposition 4.11 shows that only the fact that $dJ(u)$ exists is needed to prove that

$$J(v) - J(u) \geq dJ(u)(v-u) \quad \text{for all } v \in U,$$

and if

$$dJ(u)(v-u) \geq 0 \quad \text{for all } v \in U,$$

then

$$J(v) - J(u) \geq 0 \quad \text{for all } v \in U,$$

as claimed.

(4) If U is open, then for every $u \in U$ we can find an open ball B centered at u of radius ϵ small enough so that $B \subseteq U$. Then for any $w \neq 0$ such that $\|w\| < \epsilon$, we have both $v = u + w \in B$ and $v' = u - w \in B$, so Condition (3) implies that

$$dJ(u)(w) \geq 0 \quad \text{and} \quad dJ(u)(-w) \geq 0,$$

which yields

$$dJ(u)(w) = 0.$$

Since the above holds for all $w \neq 0$ such such that $\|w\| < \epsilon$ and since $dJ(u)$ is linear, we leave it to the reader to fill in the details of the proof that $dJ(u) = 0$. □

Example 4.7. Theorem 4.13 can be used to rederive the fact that the least squares solutions of a linear system $Ax = b$ (where A is an $m \times n$ matrix) are given by the normal equation

$$A^\top Ax = A^\top b.$$

For this, we consider the quadratic function

$$J(v) = \frac{1}{2}\|Av - b\|_2^2 - \frac{1}{2}\|b\|_2^2,$$

and our least squares problem is equivalent to finding the minima of J on $\mathbb{R}^n$. A computation reveals that

$$\begin{aligned} J(v) &= \frac{1}{2}\|Av - b\|_2^2 - \frac{1}{2}\|b\|_2^2 \\ &= \frac{1}{2}(Av - b)^\top (Av - b) - \frac{1}{2}b^\top b \\ &= \frac{1}{2}(v^\top A^\top - b^\top)(Av - b) - \frac{1}{2}b^\top b \\ &= \frac{1}{2}v^\top A^\top Av - v^\top A^\top b, \end{aligned}$$

and so

$$dJ(u) = A^\top Au - A^\top b.$$

Since $A^\top A$ is positive semidefinite, the function J is convex, and Theorem 4.13(4) implies that the minima of J are the solutions of the equation

$$A^\top Au - A^\top b = 0.$$

The considerations in this chapter reveal the need to find methods for finding the zeros of the derivative map

$$dJ : \Omega \to E',$$

where Ω is some open subset of a normed vector space E and E' is the space of all continuous linear forms on E (a subspace of E^*). Generalizations of *Newton's method* yield such methods and they are the object of the next chapter.

4.4 Summary

The main concepts and results of this chapter are listed below:

- Local minimum, local maximum, local extremum, strict local minimum, strict local maximum.
- Necessary condition for a local extremum involving the derivative; critical point.
- Local minimum with respect to a subset U, local maximum with respect to a subset U, local extremum with respect to a subset U.
- Constrained local extremum.
- Necessary condition for a constrained extremum.
- Necessary condition for a constrained extremum in terms of Lagrange multipliers.
- Lagrangian.
- Critical points of a Lagrangian.
- Necessary condition of an unconstrained local minimum involving the second-order derivative.
- Sufficient condition for a local minimum involving the second-order derivative.
- A sufficient condition involving nondegenerate critical points.
- Convex sets, convex functions, concave functions, strictly convex functions, strictly concave functions.
- Necessary condition for a local minimum on a convex set involving the derivative.
- Convexity of a function involving a condition on its first derivative.
- Convexity of a function involving a condition on its second derivative.
- Minima of convex functions on convex sets.

4.5 Problems

Problem 4.1. Find the extrema of the function $J(v_1, v_2) = v_2^2$ on the subset U given by

$$U = \{(v_1, v_2) \in \mathbb{R}^2 \mid v_1^2 + v_2^2 - 1 = 0\}.$$

Problem 4.2. Find the extrema of the function $J(v_1, v_2) = v_1 + (v_2 - 1)^2$ on the subset U given by

$$U = \{(v_1, v_2) \in \mathbb{R}^2 \mid v_1^2 = 0\}.$$

Problem 4.3. Let A be an $n \times n$ real symmetric matrix, B an $n \times n$ symmetric positive definite matrix, and let $b \in \mathbb{R}^n$.

(1) Prove that a necessary condition for the function J given by

$$J(v) = \frac{1}{2} v^\top A v - b^\top v$$

to have an extremum at $u \in U$, with U defined by

$$U = \{v \in \mathbb{R}^n \mid v^\top B v = 1\},$$

is that there is some $\lambda \in \mathbb{R}$ such that

$$Au - b = \lambda B u.$$

(2) Prove that there is a symmetric positive definite matrix S such that $B = S^2$. Prove that if $b = 0$, then λ is an eigenvalue of the symmetric matrix $S^{-1} A S^{-1}$.

(3) Prove that for all $(u, \lambda) \in U \times \mathbb{R}$, if $Au - b = \lambda B u$, then

$$J(v) - J(u) = \frac{1}{2}(v - u)^\top (A - \lambda B)(v - u)$$

for all $v \in U$. Deduce that without additional assumptions, it is not possible to conclude that u is an extremum of J on U.

Problem 4.4. Let E be a normed vector space, and let U be a subset of E such that for all $u, v \in U$, we have $(1/2)(u + v) \in U$.

(1) Prove that if U is closed, then U is convex.

Hint. Every real $\theta \in (0, 1)$ can be written as

$$\theta = \sum_{n \geq 1} \alpha_n 2^{-n},$$

with $\alpha_n \in \{0, 1\}$.

(2) Does the result in (1) hold if U is not closed?

Problem 4.5. Prove that the function f with domain $\mathrm{dom}(f) = \mathbb{R} - \{0\}$ given by $f(x) = 1/x^2$ has the property that $f''(x) > 0$ for all $x \in \mathrm{dom}(f)$, but it is not convex. Why isn't Proposition 4.12 applicable?

Problem 4.6. (1) Prove that the function $x \mapsto e^{ax}$ (on $\mathbb{R}$) is convex for any $a \in \mathbb{R}$.

(2) Prove that the function $x \mapsto x^a$ is convex on $\{x \in \mathbb{R} \mid x > 0\}$, for all $a \in \mathbb{R}$ such that $a \leq 0$ or $a \geq 1$.

Problem 4.7. (1) Prove that the function $x \mapsto |x|^p$ is convex on $\mathbb{R}$ for all $p \geq 1$.

(2) Prove that the function $x \mapsto \log x$ is concave on $\{x \in \mathbb{R} \mid x > 0\}$.

(3) Prove that the function $x \mapsto x \log x$ is convex on $\{x \in \mathbb{R} \mid x > 0\}$.

Problem 4.8. (1) Prove that the function f given by $f(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$ is convex on $\mathbb{R}^n$.

(2) Prove that the function g given by $g(x_1, \ldots, x_n) = \log(e^{x_1} + \cdots + e^{x_n})$ is convex on $\mathbb{R}^n$.

Prove that

$$\max\{x_1, \ldots, x_n\} \leq g(x_1, \ldots, x_n) \leq \max\{x_1, \ldots, x_n\} + \log n.$$

Problem 4.9. In Problem 3.6, it was shown that

$$df_A(X) = \mathrm{tr}(A^{-1}X)$$
$$\mathrm{D}^2 f(A)(X_1, X_2) = -\mathrm{tr}(A^{-1}X_1A^{-1}X_2),$$

for all $n \times n$ real matrices X, X_1, X_2, where f is the function defined on $\mathbf{GL}^+(n, \mathbb{R})$ (the $n \times n$ real invertible matrices of positive determinants), given by

$$f(A) = \log \det(A).$$

Assume that A is symmetric positive definite and that X is symmetric.

(1) Prove that the eigenvalues of $A^{-1}X$ are real (even though $A^{-1}X$ may **not** be symmetric).

Hint. Since A is symmetric positive definite, then so is A^{-1}, so we can write $A^{-1} = S^2$ for some symmetric positive definite matrix S, and then

$$A^{-1}X = S^2X = S(SXS)S^{-1}.$$

(2) Prove that the eigenvalues of $(A^{-1}X)^2$ are nonnegative. Deduce that

$$\mathrm{D}^2 f(A)(X, X) = -\mathrm{tr}((A^{-1}X)^2) < 0$$

for all nonzero symmetric matrices X and SPD matrices A. Conclude that the function $X \mapsto \log \det X$ is strictly concave on the set of symmetric positive definite matrices.

Chapter 5

Newton's Method and Its Generalizations

In Chapter 4 we investigated the problem of determining when a function $J: \Omega \to \mathbb{R}$ defined on some open subset Ω of a normed vector space E has a local extremum. Proposition 4.1 gives a necessary condition when J is differentiable: if J has a local extremum at $u \in \Omega$, then we must have

$$J'(u) = 0.$$

Thus we are led to the problem of finding the zeros of the derivative

$$J': \Omega \to E',$$

where $E' = \mathcal{L}(E; \mathbb{R})$ is the set of linear continuous functions from E to $\mathbb{R}$; that is, the *dual* of E, as defined in the remark after Proposition 4.8.

This leads us to consider the problem in a more general form, namely, given a function $f: \Omega \to Y$ from an open subset Ω of a normed vector space X to a normed vector space Y, find

(i) Sufficient conditions which guarantee the *existence of a zero* of the function f; that is, an element $a \in \Omega$ such that $f(a) = 0$.

(ii) An *algorithm* for approximating such an a, that is, a sequence (x_k) of points of Ω whose limit is a.

In this chapter we discuss Newton's method and some of it generalizations to give (partial) answers to Problems (i) and (ii).

5.1 Newton's Method for Real Functions of a Real Argument

When $X = Y = \mathbb{R}$, we can use *Newton's method* to find a zero of a function $f: \Omega \to \mathbb{R}$. We pick some initial element $x_0 \in \mathbb{R}$ "close enough" to a zero a of f, and we define the sequence

(x_k) by

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},$$

for all $k \geq 0$, provided that $f'(x_k) \neq 0$. The idea is to define x_{k+1} as the intersection of the x-axis with the tangent line to the graph of the function $x \mapsto f(x)$ at the point $(x_k, f(x_k))$. Indeed, the equation of this tangent line is

$$y - f(x_k) = f'(x_k)(x - x_k),$$

and its intersection with the x-axis is obtained for $y = 0$, which yields

$$x = x_k - \frac{f(x_k)}{f'(x_k)},$$

as claimed. See Figure 5.1.

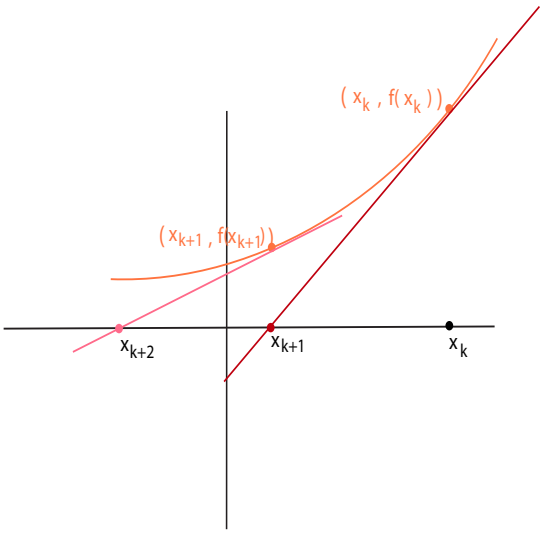


Figure 5.1: The construction of two stages in Newton's method.

Example 5.1. If $\alpha > 0$ and $f(x) = x^2 - \alpha$, Newton's method yields the sequence

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{\alpha}{x_k}\right)$$

to compute the square root $\sqrt{\alpha}$ of α. It can be shown that the method converges to $\sqrt{\alpha}$ for any $x_0 > 0$; see Problem 5.1. Actually, the method also converges when $x_0 < 0$! Find out what is the limit.

The case of a real function suggests the following method for finding the zeros of a function $f\colon \Omega \to Y$, with $\Omega \subseteq X$: given a starting point $x_0 \in \Omega$, the sequence (x_k) is defined by

$$x_{k+1} = x_k - (f'(x_k))^{-1}(f(x_k)) \tag{$*$}$$

for all $k \geq 0$.

For the above to make sense, it must be ensured that

(1) All the points x_k remain within Ω.

(2) The function f is differentiable within Ω.

(3) The derivative $f'(x)$ is a bijection from X to Y for all $x \in \Omega$.

These are rather demanding conditions but there are sufficient conditions that guarantee that they are met. Another practical issue is that it may be very costly to compute $(f'(x_k))^{-1}$ at every iteration step. In the next section we investigate generalizations of Newton's method which address the issues that we just discussed.

5.2 Generalizations of Newton's Method

Suppose that $f\colon \Omega \to \mathbb{R}^n$ is given by n functions $f_i\colon \Omega \to \mathbb{R}$, where $\Omega \subseteq \mathbb{R}^n$. In this case, finding a zero a of f is equivalent to solving the system

$$\begin{aligned} f_1(a_1 \ldots, a_n) &= 0 \\ f_2(a_1 \ldots, a_n) &= 0 \\ &\vdots \\ f_n(a_1 \ldots, a_n) &= 0. \end{aligned}$$

In the standard Newton method, the iteration step is given by $(*)$, namely

$$x_{k+1} = x_k - (f'(x_k))^{-1}(f(x_k)),$$

and if we define Δx_k as $\Delta x_k = x_{k+1} - x_k$, we see that $\Delta x_k = -(f'(x_k))^{-1}(f(x_k))$, so Δx_k is obtained by solving the equation

$$f'(x_k)\Delta x_k = -f(x_k),$$

and then we set $x_{k+1} = x_k + \Delta x_k$.

The generalization is as follows.

Variant 1. A single iteration of Newton's method consists in solving the linear system

$$(J(f)(x_k))\Delta x_k = -f(x_k),$$

and then setting

$$x_{k+1} = x_k + \Delta x_k,$$

where $J(f)(x_k) = \left(\frac{\partial f_i}{\partial x_j}(x_k)\right)$ is the Jacobian matrix of f at x_k.

In general it is very costly to compute $J(f)(x_k)$ at each iteration and then to solve the corresponding linear system. If the method converges, the consecutive vectors x_k should differ only a little, as also the corresponding matrices $J(f)(x_k)$. Thus, we are led to several variants of Newton's method.

Variant 2. This variant consists in keeping the same matrix for p consecutive steps (where p is some fixed integer ≥ 2):

$$\begin{aligned}
x_{k+1} &= x_k - (f'(x_0))^{-1}(f(x_k)), && 0 \leq k \leq p-1\\
x_{k+1} &= x_k - (f'(x_p))^{-1}(f(x_k)), && p \leq k \leq 2p-1\\
&\vdots\\
x_{k+1} &= x_k - (f'(x_{rp}))^{-1}(f(x_k)), && rp \leq k \leq (r+1)p-1\\
&\vdots
\end{aligned}$$

Variant 3. Set $p = \infty$, that is, use *the same matrix $f'(x_0)$ for all iterations*, which leads to iterations of the form

$$x_{k+1} = x_k - (f'(x_0))^{-1}(f(x_k)), \quad k \geq 0,$$

Variant 4. Replace $f'(x_0)$ by a particular matrix A_0 which is easy to invert:

$$x_{k+1} = x_k - A_0^{-1} f(x_k), \quad k \geq 0.$$

In the last two cases, if possible, we use an LU factorization of $f'(x_0)$ or A_0 to speed up the method. In some cases, it may even possible to set $A_0 = I$.

The above considerations lead us to the definition of a *generalized Newton method*, as in Ciarlet [25] (Chapter 7). Recall that a linear map $f \in \mathcal{L}(E; F)$ is called an *isomorphism* iff f is continuous, bijective, and f^{-1} is also continuous.

Definition 5.1. If X and Y are two normed vector spaces and if $f\colon \Omega \to Y$ is a function from some open subset Ω of X, a *generalized Newton method* for finding zeros of f consists of

(1) A sequence of families $(A_k(x))$ of linear isomorphisms from X to Y, for all $x \in \Omega$ and all integers $k \geq 0$;

(2) Some starting point $x_0 \in \Omega$;

(3) A sequence (x_k) of points of Ω defined by

$$x_{k+1} = x_k - (A_k(x_\ell))^{-1}(f(x_k)), \quad k \geq 0, \tag{$**$}$$

where for every integer $k \geq 0$, the integer ℓ satisfies the condition

$$0 \leq \ell \leq k.$$

With $\Delta x_k = x_{k+1} - x_k$, Equation $(**)$ is equivalent to solving the equation

$$A_k(x_\ell)(\Delta x_k) = -f(x_k)$$

and setting $x_{k+1} = x_k + \Delta x_k$. The function $A_k(x)$ usually depends on f'.

Definition 5.1 gives us enough flexibility to capture all the situations that we have previously discussed:

	Function	Index
Variant 1:	$A_k(x) = f'(x)$,	$\ell = k$
Variant 2:	$A_k(x) = f'(x)$,	$\ell = \min\{rp, k\}$, if $rp \leq k \leq (r+1)p - 1, r \geq 0$
Variant 3:	$A_k(x) = f'(x)$,	$\ell = 0$
Variant 4:	$A_k(x) = A_0$,	

where A_0 is a linear isomorphism from X to Y. Note that in Variant 2, ℓ is defined more concisely as $\ell = \lfloor k/p \rfloor p$. The first case corresponds to Newton's original method and the others to the variants that we just discussed. We could also have $A_k(x) = A_k$, a fixed linear isomorphism independent of $x \in \Omega$.

Example 5.2. Consider the matrix function f given by

$$f(X) = A - X^{-1},$$

with A and X invertible $n \times n$ matrices. If we apply Variant 1 of Newton's method starting with any $n \times n$ matrix X_0, since the derivative of the function g given by $g(X) = X^{-1}$ is $dg_X(Y) = -X^{-1}YX^{-1}$, we have

$$f'_X(Y) = X^{-1}YX^{-1},$$

so

$$(f'_X)^{-1}(Y) = XYX$$

and the Newton step is

$$X_{k+1} = X_k - (f'_{X_k})^{-1}(f(X_k)) = X_k - X_k(A - X_k^{-1})X_k,$$

which yields the sequence (X_k) with

$$X_{k+1} = X_k(2I - AX_k), \quad k \geq 0.$$

In Problem 5.5, it is shown that Newton's method converges to A^{-1} iff the spectral radius of $I - X_0A$ is strictly smaller than 1, that is, $\rho(I - X_0A) < 1$.

The following theorem inspired by the *Newton–Kantorovich theorem* gives sufficient conditions that guarantee that the sequence (x_k) constructed by a generalized Newton method converges to a zero of f close to x_0. Although quite technical, these conditions are not very surprising.

Theorem 5.1. *Let X be a Banach space, let $f\colon \Omega \to Y$ be differentiable on the open subset $\Omega \subseteq X$, and assume that there are constants $r, M, \beta > 0$ such that if we let*

$$B = \{x \in X \mid \|x - x_0\| \leq r\} \subseteq \Omega,$$

then

(1)

$$\sup_{k\geq 0} \sup_{x\in B} \left\|A_k^{-1}(x)\right\|_{\mathcal{L}(Y;X)} \leq M,$$

(2) $\beta < 1$ and

$$\sup_{k\geq 0} \sup_{x,x'\in B} \|f'(x) - A_k(x')\|_{\mathcal{L}(X;Y)} \leq \frac{\beta}{M}$$

(3)

$$\|f(x_0)\| \leq \frac{r}{M}(1-\beta).$$

Then the sequence (x_k) defined by

$$x_{k+1} = x_k - A_k^{-1}(x_\ell)(f(x_k)), \quad 0 \leq \ell \leq k$$

is entirely contained within B and converges to a zero a of f, which is the only zero of f in B. Furthermore, the convergence is geometric, which means that

$$\|x_k - a\| \leq \frac{\|x_1 - x_0\|}{1-\beta}\beta^k.$$

Proof. We follow Ciarlet [25] (Theorem 7.5.1, Section 7.5). The proof has three steps.

Step 1. We establish the following inequalities for all $k \geq 1$.

$$\|x_k - x_{k-1}\| \leq M\,\|f(x_{k-1})\| \tag{a}$$
$$\|x_k - x_0\| \leq r \quad (x_k \in B) \tag{b}$$
$$\|f(x_k)\| \leq \frac{\beta}{M}\,\|x_k - x_{k-1}\|\,. \tag{c}$$

We proceed by induction on k, starting with the base case $k = 1$. Since

$$x_1 = x_0 - A_0^{-1}(x_0)(f(x_0)),$$

we have $x_1 - x_0 = -A_0^{-1}(x_0)(f(x_0))$, so by (1) and (3) and since $0 < \beta < 1$, we have

$$\|x_1 - x_0\| \leq M \|f(x_0)\| \leq r(1-\beta) \leq r,$$

establishing (a) and (b) for $k = 1$. We also have $f(x_0) = -A_0(x_0)(x_1 - x_0)$, so $-f(x_0) - A_0(x_0)(x_1 - x_0) = 0$ and thus

$$f(x_1) = f(x_1) - f(x_0) - A_0(x_0)(x_1 - x_0).$$

By the mean value theorem (Proposition 3.11) applied to the function $x \mapsto f(x) - A_0(x_0)(x)$, by (2), we get

$$\|f(x_1)\| \leq \sup_{x \in B} \|f'(x) - A_0(x_0)\| \|x_1 - x_0\| \leq \frac{\beta}{M} \|x_1 - x_0\|,$$

which is (c) for $k = 1$. We now establish the induction step.

Since by definition

$$x_k - x_{k-1} = -A_{k-1}^{-1}(x_\ell)(f(x_{k-1})), \quad 0 \leq \ell \leq k-1,$$

by (1) and the fact that by the induction hypothesis for (b), $x_\ell \in B$, we get

$$\|x_k - x_{k-1}\| \leq M \|f(x_{k-1})\|,$$

which proves (a) for k. As a consequence, since by the induction hypothesis for (c),

$$\|f(x_{k-1})\| \leq \frac{\beta}{M} \|x_{k-1} - x_{k-2}\|,$$

we get

$$\|x_k - x_{k-1}\| \leq M \|f(x_{k-1})\| \leq \beta \|x_{k-1} - x_{k-2}\|, \tag{$*_1$}$$

and by repeating this step,

$$\|x_k - x_{k-1}\| \leq \beta^{k-1} \|x_1 - x_0\|. \tag{$*_2$}$$

Using ($*_2$) and (3), we obtain

$$\begin{aligned}\|x_k - x_0\| &\leq \sum_{j=1}^{k} \|x_j - x_{j-1}\| \leq \left(\sum_{j=1}^{k} \beta^{j-1}\right) \|x_1 - x_0\| \\ &\leq \frac{\|x_1 - x_0\|}{1-\beta} \leq \frac{M}{1-\beta} \|f(x_0)\| \leq r,\end{aligned}$$

which proves that $x_k \in B$, which is (b) for k.

Since

$$x_k - x_{k-1} = -A_{k-1}^{-1}(x_\ell)(f(x_{k-1}))$$

we also have $f(x_{k-1}) = -A_{k-1}(x_\ell)(x_k - x_{k-1})$, so we have

$$f(x_k) = f(x_k) - f(x_{k-1}) - A_{k-1}(x_\ell)(x_k - x_{k-1}),$$

and as in the base case, applying the mean value theorem (Proposition 3.11) to the function $x \mapsto f(x) - A_{k-1}(x_\ell)(x)$, by (2), we obtain

$$\|f(x_k)\| \leq \sup_{x \in B} \|f'(x) - A_{k-1}(x_\ell)\| \, \|x_k - x_{k-1}\| \leq \frac{\beta}{M} \|x_k - x_{k-1}\|,$$

proving (c) for k.

Step 2. Prove that f has a zero in B.

To do this we prove that (x_k) is a Cauchy sequence. This is because, using $(*_2)$, we have

$$\begin{aligned}\|x_{k+j} - x_k\| &\leq \sum_{i=0}^{j-1} \|x_{k+i+1} - x_{k+i}\| \leq \beta^k \left(\sum_{i=0}^{j-1} \beta^i \right) \|x_1 - x_0\| \\ &\leq \frac{\beta^k}{1-\beta} \|x_1 - x_0\|,\end{aligned}$$

for all $k \geq 0$ and all $j \geq 0$, proving that (x_k) is a Cauchy sequence. Since B is a closed ball in a complete normed vector space, B is complete and the Cauchy sequence (x_k) converges to a limit $a \in B$. Since f is continuous on Ω (because it is differentiable), by (c) we obtain

$$\|f(a)\| = \lim_{k \mapsto \infty} \|f(x_k)\| \leq \frac{\beta}{M} \lim_{k \mapsto \infty} \|x_k - x_{k-1}\| = 0,$$

which yields $f(a) = 0$.

Since

$$\|x_{k+j} - x_k\| \leq \frac{\beta^k}{1-\beta} \|x_1 - x_0\|,$$

if we let j tend to infinity, we obtain the inequality

$$\|x_k - a\| = \|a - x_k\| \leq \frac{\beta^k}{1-\beta} \|x_1 - x_0\|,$$

which is the last statement of the theorem.

Step 3. Prove that f has a unique zero in B.

Suppose $f(a) = f(b) = 0$ with $a, b \in B$. Since $A_0^{-1}(x_0)(A_0(x_0)(b-a)) = b - a$, we have

$$b - a = -A_0^{-1}(x_0)(f(b) - f(a) - A_0(x_0)(b-a)),$$

which by (1) and (2) and the mean value theorem implies that

$$\|b - a\| \leq \left\|A_0^{-1}(x_0)\right\| \sup_{x \in B} \|f'(x) - A_0(x_0)\| \, \|b - a\| \leq \beta \|b - a\|.$$

Since $0 < \beta < 1$, the inequality $\|b - a\| \leq \beta \|b - a\|$ is only possible if $a = b$. □

It should be observed that the conditions of Theorem 5.1 are typically quite stringent. It can be shown that Theorem 5.1 applies to the function f of Example 5.1 given by $f(x) = x^2 - \alpha$ with $\alpha > 0$, for any $x_0 > 0$ such that

$$\frac{6}{7}\alpha \leq x_0^2 \leq \frac{6}{5}\alpha,$$

with $\beta = 2/5$, $r = (1/6)x_0$, $M = 3/(5x_0)$. Such values of x_0 are quite close to $\sqrt{\alpha}$.

If we assume that we already know that some element $a \in \Omega$ is a zero of f, the next theorem gives sufficient conditions for a special version of a generalized Newton method to converge. For this special method the linear isomorphisms $A_k(x)$ are independent of $x \in \Omega$.

Theorem 5.2. *Let X be a Banach space and let $f\colon \Omega \to Y$ be differentiable on the open subset $\Omega \subseteq X$. If $a \in \Omega$ is a point such that $f(a) = 0$, if $f'(a)$ is a linear isomorphism, and if there is some λ with $0 < \lambda < 1/2$ such that*

$$\sup_{k \geq 0} \|A_k - f'(a)\|_{\mathcal{L}(X;Y)} \leq \frac{\lambda}{\|(f'(a))^{-1}\|_{\mathcal{L}(Y;X)}},$$

then there is a closed ball B of center a such that for every $x_0 \in B$, the sequence (x_k) defined by

$$x_{k+1} = x_k - A_k^{-1}(f(x_k)), \quad k \geq 0,$$

is entirely contained within B and converges to a, which is the only zero of f in B. Furthermore, the convergence is geometric, which means that

$$\|x_k - a\| \leq \beta^k \|x_0 - a\|,$$

for some $\beta < 1$.

A proof of Theorem 5.2 can be found in Ciarlet [25] (Section 7.5).

For the sake of completeness, we state a version of the Newton–Kantorovich theorem which corresponds to the case where $A_k(x) = f'(x)$. In this instance, a stronger result can be obtained especially regarding upper bounds, and we state a version due to Gragg and Tapia which appears in Problem 7.5-4 of Ciarlet [25].

Theorem 5.3. *(Newton–Kantorovich) Let X be a Banach space, and let $f\colon \Omega \to Y$ be differentiable on the open subset $\Omega \subseteq X$. Assume that there exist three positive constants λ, μ, ν and a point $x_0 \in \Omega$ such that*

$$0 < \lambda\mu\nu \leq \frac{1}{2},$$

and if we let

$$\begin{aligned}\rho^- &= \frac{1-\sqrt{1-2\lambda\mu\nu}}{\mu\nu}\\ \rho^+ &= \frac{1+\sqrt{1-2\lambda\mu\nu}}{\mu\nu}\\ B &= \{x \in X \mid \|x - x_0\| < \rho^-\}\\ \Omega^+ &= \{x \in \Omega \mid \|x - x_0\| < \rho^+\},\end{aligned}$$

then $\overline{B} \subseteq \Omega$, $f'(x_0)$ *is an isomorphism of* $\mathcal{L}(X;Y)$, *and*

$$\begin{aligned}\left\|(f'(x_0))^{-1}\right\| &\leq \mu,\\ \left\|(f'(x_0))^{-1} f(x_0)\right\| &\leq \lambda,\\ \sup_{x,y\in\Omega^+} \|f'(x) - f'(y)\| &\leq \nu \|x - y\|.\end{aligned}$$

Then $f'(x)$ *is isomorphism of* $\mathcal{L}(X;Y)$ *for all* $x \in B$, *and the sequence defined by*

$$x_{k+1} = x_k - (f'(x_k))^{-1}(f(x_k)), \quad k \geq 0$$

is entirely contained within the ball B *and converges to a zero* a *of* f *which is the only zero of* f *in* Ω^+. *Finally, if we write* $\theta = \rho^-/\rho^+$, *then we have the following bounds:*

$$\begin{aligned}\|x_k - a\| &\leq \frac{2\sqrt{1-2\lambda\mu\nu}}{\lambda\mu\nu}\frac{\theta^{2k}}{1-\theta^{2k}}\|x_1 - x_0\| && \text{if } \lambda\mu\nu < \frac{1}{2}\\ \|x_k - a\| &\leq \frac{\|x_1 - x_0\|}{2^{k-1}} && \text{if } \lambda\mu\nu = \frac{1}{2},\end{aligned}$$

and

$$\frac{2\|x_{k+1} - x_k\|}{1+\sqrt{(1+4\theta^{2k}(1+\theta^{2k})^{-2})}} \leq \|x_k - a\| \leq \theta^{2k-1}\|x_k - x_{k-1}\|.$$

We can now specialize Theorems 5.1 and 5.2 to the search of zeros of the derivative $J'\colon \Omega \to E'$, of a function $J\colon \Omega \to \mathbb{R}$, with $\Omega \subseteq E$. The second derivative J'' of J is a continuous bilinear form $J''\colon E \times E \to \mathbb{R}$, but is is convenient to view it as a linear map in $\mathcal{L}(E, E')$; the continuous linear form $J''(u)$ is given by $J''(u)(v) = J''(u, v)$. In our next theorem, which follows immediately from Theorem 5.1, we assume that the $A_k(x)$ are isomorphisms in $\mathcal{L}(E, E')$.

Theorem 5.4. *Let* E *be a Banach space, let* $J\colon \Omega \to \mathbb{R}$ *be twice differentiable on the open subset* $\Omega \subseteq E$, *and assume that there are constants* $r, M, \beta > 0$ *such that if we let*

$$B = \{x \in E \mid \|x - x_0\| \leq r\} \subseteq \Omega,$$

then

(1)

$$\sup_{k\geq 0}\sup_{x\in B}\left\|A_k^{-1}(x)\right\|_{\mathcal{L}(E';E)}\leq M,$$

(2) $\beta<1$ *and*

$$\sup_{k\geq 0}\sup_{x,x'\in B}\|J''(x)-A_k(x')\|_{\mathcal{L}(E;E')}\leq\frac{\beta}{M}$$

(3)

$$\|J'(x_0)\|\leq\frac{r}{M}(1-\beta).$$

Then the sequence (x_k) *defined by*

$$x_{k+1}=x_k-A_k^{-1}(x_\ell)(J'(x_k)),\quad 0\leq\ell\leq k$$

is entirely contained within B *and converges to a zero* a *of* J'*, which is the only zero of* J' *in* B*. Furthermore, the convergence is geometric, which means that*

$$\|x_k-a\|\leq\frac{\|x_1-x_0\|}{1-\beta}\beta^k.$$

In the next theorem, which follows immediately from Theorem 5.2, we assume that the $A_k(x)$ are isomorphisms in $\mathcal{L}(E,E')$ that are independent of $x\in\Omega$.

Theorem 5.5. *Let* E *be a Banach space and let* $J\colon\Omega\to\mathbb{R}$ *be twice differentiable on the open subset* $\Omega\subseteq E$*. If* $a\in\Omega$ *is a point such that* $J'(a)=0$*, if* $J''(a)$ *is a linear isomorphism, and if there is some* λ *with* $0<\lambda<1/2$ *such that*

$$\sup_{k\geq 0}\|A_k-J''(a)\|_{\mathcal{L}(E;E')}\leq\frac{\lambda}{\|(J''(a))^{-1}\|_{\mathcal{L}(E';E)}},$$

then there is a closed ball B *of center* a *such that for every* $x_0\in B$*, the sequence* (x_k) *defined by*

$$x_{k+1}=x_k-A_k^{-1}(J'(x_k)),\quad k\geq 0,$$

is entirely contained within B *and converges to* a*, which is the only zero of* J' *in* B*. Furthermore, the convergence is geometric, which means that*

$$\|x_k-a\|\leq\beta^k\|x_0-a\|,$$

for some $\beta<1$*.*

When $E=\mathbb{R}^n$, the Newton method given by Theorem 5.4 yields an iteration step of the form

$$x_{k+1}=x_k-A_k^{-1}(x_\ell)\nabla J(x_k),\quad 0\leq\ell\leq k,$$

where $\nabla J(x_k)$ is the gradient of J at x_k (here, we identify E' with $\mathbb{R}^n$). In particular, Newton's original method picks $A_k = J''$, and the iteration step is of the form

$$x_{k+1} = x_k - (\nabla^2 J(x_k))^{-1} \nabla J(x_k), \quad k \geq 0,$$

where $\nabla^2 J(x_k)$ is the Hessian of J at x_k.

Example 5.3. Let us apply Newton's original method to the function J given by $J(x) = \frac{1}{3}x^3 - 4x$. We have $J'(x) = x^2 - 4$ and $J''(x) = 2x$, so the Newton step is given by

$$x_{k+1} = x_k - \frac{x_k^2 - 4}{2x_k} = \frac{1}{2}\left(x_k + \frac{4}{x_k}\right).$$

This is the sequence of Example 5.1 to compute the square root of 4. Starting with any $x_0 > 0$ it converges very quickly to 2.

As remarked in Ciarlet [25] (Section 7.5), generalized Newton methods have a very wide range of applicability. For example, various versions of gradient descent methods can be viewed as instances of Newton method. See Section 13.9 for an example.

Newton's method also plays an important role in convex optimization, in particular, interior-point methods. A variant of Newton's method dealing with equality constraints has been developed. We refer the reader to Boyd and Vandenberghe [18], Chapters 10 and 11, for a comprehensive exposition of these topics.

5.3 Summary

The main concepts and results of this chapter are listed below:

- Newton's method for functions $f\colon \mathbb{R} \to \mathbb{R}$.
- Generalized Newton methods.
- The *Newton-Kantorovich* theorem.

5.4 Problems

Problem 5.1. If $\alpha > 0$ and $f(x) = x^2 - \alpha$, Newton's method yields the sequence

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{\alpha}{x_k}\right)$$

to compute the square root $\sqrt{\alpha}$ of α.

(1) Prove that if $x_0 > 0$, then $x_k > 0$ and

$$x_{k+1} - \sqrt{\alpha} = \frac{1}{2x_k}(x_k - \sqrt{\alpha})^2$$
$$x_{k+2} - x_{k+1} = \frac{1}{2x_{k+1}}(\alpha - x_{k+1}^2)$$

for all $k \geq 0$. Deduce that Newton's method converges to $\sqrt{\alpha}$ for any $x_0 > 0$.

(2) Prove that if $x_0 < 0$, then Newton's method converges to $-\sqrt{\alpha}$.

Problem 5.2. (1) If $\alpha > 0$ and $f(x) = x^2 - \alpha$, show that the conditions of Theorem 5.1 are satisfied by any $\beta \in (0, 1)$ and any x_0 such that

$$|x_0^2 - \alpha| \leq \frac{4\beta(1-\beta)}{(\beta+2)^2} x_0^2,$$

with

$$r = \frac{\beta}{\beta+2} x_0, \qquad M = \frac{\beta+2}{4x_0}.$$

(2) Prove that the maximum of the function defined on $[0, 1]$ by

$$\beta \mapsto \frac{4\beta(1-\beta)}{(\beta+2)^2}$$

has a maximum for $\beta = 2/5$. For this value of β, check that $r = (1/6)x_0$, $M = 3/(5x_0)$, and

$$\frac{6}{7}\alpha \leq x_0^2 \leq \frac{6}{5}\alpha.$$

Problem 5.3. Consider generalizing Problem 5.1 to the matrix function f given by $f(X) = X^2 - C$, where X and C are two real $n \times n$ matrices with C symmetric positive definite. The first step is to determine for which A does the inverse df_A^{-1} exist. Let g be the function given by $g(X) = X^2$. From Problem 3.1 we know that the derivative at A of the function g is $dg_A(X) = AX + XA$, and obviously $df_A = dg_A$. Thus we are led to figure out when the linear matrix map $X \mapsto AX + XA$ is invertible. This can be done using the Kronecker product.

Given an $m \times n$ matrix $A = (a_{ij})$ and a $p \times q$ matrix $B = (b_{ij})$, the *Kronecker product* (or *tensor product*) $A \otimes B$ of A and B is the $mp \times nq$ matrix

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}.$$

It can be shown (and you may use these facts without proof) that $\otimes$ is associative and that

$$(A \otimes B)(C \otimes D) = AC \otimes BD$$
$$(A \otimes B)^\top = A^\top \otimes B^\top,$$

whenever AC and BD are well defined.

Given any $n \times n$ matrix X, let $\mathrm{vec}(X)$ be the vector in $\mathbb{R}^{n^2}$ obtained by concatenating the *rows* of X.

(1) Prove that $AX = Y$ iff

$$(A \otimes I_n)\mathrm{vec}(X) = \mathrm{vec}(Y)$$

and $XA = Y$ iff

$$(I_n \otimes A^\top)\mathrm{vec}(X) = \mathrm{vec}(Y).$$

Deduce that $AX + XA = Y$ iff

$$((A \otimes I_n) + (I_n \otimes A^\top))\mathrm{vec}(X) = \mathrm{vec}(Y).$$

In the case where $n = 2$ and if we write

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

check that

$$A \otimes I_2 + I_2 \otimes A^\top = \begin{pmatrix} 2a & c & b & 0 \\ b & a+d & 0 & b \\ c & 0 & a+d & c \\ 0 & c & b & 2d \end{pmatrix}.$$

The problem is to determine when the matrix $(A \otimes I_n) + (I_n \otimes A^\top)$ is invertible.

Remark: The equation $AX + XA = Y$ is a special case of the equation $AX + XB = C$ (sometimes written $AX - XB = C$), called the *Sylvester equation*, where A is an $m \times m$ matrix, B is an $n \times n$ matrix, and X, C are $m \times n$ matrices; see Higham [42] (Appendix B).

(2) In the case where $n = 2$, prove that

$$\det(A \otimes I_2 + I_2 \otimes A^\top) = 4(a+d)^2(ad - bc).$$

(3) Let A and B be any two $n \times n$ complex matrices. Use Schur factorizations $A = UT_1U^*$ and $B = VT_2V^*$ (where U and V are unitary and T_1, T_2 are upper-triangular) to prove that if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A and $\mu_1, \ldots, \mu_n$ are the eigenvalues of B, then the scalars $\lambda_i\mu_j$ are the eigenvalues of $A \otimes B$, for $i, j = 1, \ldots, n$.

Hint. Check that $U \otimes V$ is unitary and that $T_1 \otimes T_2$ is upper triangular.

(4) Prove that the eigenvalues of $(A \otimes I_n) + (I_n \otimes B)$ are the scalars $\lambda_i + \mu_j$, for $i, j = 1, \ldots, n$. Deduce that the eigenvalues of $(A \otimes I_n) + (I_n \otimes A^\top)$ are $\lambda_i + \lambda_j$, for $i, j = 1, \ldots, n$. Thus $(A \otimes I_n) + (I_n \otimes A^\top)$ is invertible iff $\lambda_i + \lambda_j \neq 0$, for $i, j = 1, \ldots, n$. In particular, prove that if A is symmetric positive definite, then so is $(A \otimes I_n) + (I_n \otimes A^\top)$.

Hint. Use (3).

(5) Prove that if A and B are symmetric and $(A \otimes I_n) + (I_n \otimes A^\top)$ is invertible, then the unique solution X of the equation $AX + XA = B$ is symmetric.

(6) Starting with a symmetric positive definite matrix X_0, the general step of Newton's method is

$$X_{k+1} = X_k - (f'_{X_k})^{-1}(X_k^2 - C) = X_k - (g'_{X_k})^{-1}(X_k^2 - C),$$

and since g'_{X_k} is linear, this is equivalent to

$$X_{k+1} = X_k - (g'_{X_k})^{-1}(X_k^2) + (g'_{X_k})^{-1}(C).$$

But since X_k is SPD, $(g'_{X_k})^{-1}(X_k^2)$ is the unique solution of

$$X_k Y + Y X_k = X_k^2$$

whose solution is obviously $Y = (1/2)X_k$. Therefore the Newton step is

$$X_{k+1} = X_k - (g'_{X_k})^{-1}(X_k^2) + (g'_{X_k})^{-1}(C) = X_k - \frac{1}{2}X_k + (g'_{X_k})^{-1}(C) = \frac{1}{2}X_k + (g'_{X_k})^{-1}(C),$$

so we have

$$X_{k+1} = \frac{1}{2}X_k + (g'_{X_k})^{-1}(C) = (g'_{X_k})^{-1}(X_k^2 + C).$$

Prove that if X_k and C are symmetric positive definite, then $(g'_{X_k})^{-1}(C)$ is symmetric positive definite, and if C is symmetric positive semidefinite, then $(g'_{X_k})^{-1}(C)$ is symmetric positive semidefinite.

Hint. By (5) the unique solution Z of the equation $X_k Z + Z X_k = C$ (where C is symmetric) is symmetric so it can be diagonalized as $Z = QDQ^\top$ with Q orthogonal and D a real diagonal matrix. Prove that

$$Q^\top X_k Q D + D Q^\top X_k Q = Q^\top C Q,$$

and solve the system using the diagonal elements.

Deduce that if X_k and C are SPD, then X_{k+1} is SPD.

Since $C = P\Sigma P^\top$ is SPD, it has an SPD square root (in fact unique) $C^{1/2} = P\Sigma^{1/2}P^\top$. Prove that

$$X_{k+1} - C^{1/2} = (g'_{X_k})^{-1}(X_k - C^{1/2})^2.$$

Prove that

$$\left\|(g'_{X_k})^{-1}\right\|_2 \geq \frac{1}{2\left\|X_k\right\|_2}.$$

Open problem: Does Theorem 5.1 apply for some suitable r, M, β?

(7) Prove that if C and X_0 commute, provided that the equation $X_k Z + Z X_k = C$ has a unique solution for all k, then X_k and C commute for all k and Z is given by

$$Z = \frac{1}{2} X_k^{-1} C = \frac{1}{2} C X_k^{-1}.$$

Deduce that

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-1} C) = \frac{1}{2}(X_k + C X_k^{-1}).$$

This is the matrix analog of the formula given in Problem 5.1(1).

Prove that if C and X_0 have positive eigenvalues and C and X_0 commute, then X_{k+1} has positive eigenvalues for all $k \geq 0$ and thus the sequence (X_k) is defined.

Hint. Because X_k and C commute, X_k^{-1} and C commute, and obviously X_k and X_k^{-1} commute. By Proposition 22.15 of Vol. I, X_k, X_k^{-1}, and C are triangulable in a common basis, so there is some orthogonal matrix P and some upper-triangular matrices T_1, T_2 such that

$$X_k = P T_1 P^\top, \quad X_k^{-1} = P T_1^{-1} P^\top, \quad C = P T_2 P^\top.$$

It follows that

$$X_{k+1} = \frac{1}{2} P(T_1 + T_1^{-1} T_2) P^\top.$$

Also recall that the diagonal entries of an upper-triangular matrix are the eigenvalues of that matrix.

We conjecture that if C has positive eigenvalues, then the Newton sequence converges starting with any X_0 of the form $X_0 = \mu I_n$, with $\mu > 0$.

(8) Implement the above method in `Matlab` (there is a command `kron(A, B)` to form the Kronecker product of A and B). Test your program on diagonalizable and nondiagonalizable matrices, including

$$W = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 5 & 4 & 1 & 1 \\ 4 & 5 & 1 & 1 \\ 1 & 1 & 4 & 2 \\ 1 & 1 & 2 & 4 \end{pmatrix},$$

and

$$A_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 0.01 & 0 & 0 \\ -1 & -1 & 100 & 100 \\ -1 & -1 & -100 & 100 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad A_4 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

What happens with

$$C = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad X_0 = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

The problem of determining when square roots of matrices exist and procedures for finding them are thoroughly investigated in Higham [42] (Chapter 6).

Problem 5.4. (1) Show that Newton's method applied to the function

$$f(x) = \alpha - \frac{1}{x}$$

with $\alpha \neq 0$ and $x \in \mathbb{R} - \{0\}$ yields the sequence (x_k) with

$$x_{k+1} = x_k(2 - \alpha x_k), \quad k \geq 0.$$

(2) If we let $r_k = 1 - \alpha x_k$, prove that $r_{k+1} = r_k^2$ for all $k \geq 0$. Deduce that Newton's method converges to $1/\alpha$ if $0 < \alpha x_0 < 2$.

Problem 5.5. (1) Show that Newton's method applied to the matrix function

$$f(X) = A - X^{-1},$$

with A and X invertible $n \times n$ matrices and started with any $n \times n$ matrix X_0 yields the sequence (X_k) with

$$X_{k+1} = X_k(2I - AX_k), \quad k \geq 0.$$

(2) If we let $R_k = I - AX_k$, prove that

$$R_{k+1} = I - (I - R_k)(I + R_k) = R_k^2$$

for all $k \geq 0$. Deduce that Newton's method converges to A^{-1} iff the spectral radius of $I - AX_0$ is strictly smaller than 1, that is, $\rho(I - AX_0) < 1$.

(3) Assume that A is symmetric positive definite and let $X_0 = \mu I$. Prove that the condition $\rho(I - AX_0) < 1$ is equivalent to

$$0 < \mu < \frac{2}{\rho(A)}.$$

(4) Write a `Matlab` program implementing Newton's method specified in (1). Test your program with the $n \times n$ matrix

$$A_n = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix},$$

and with $X_0 = \mu I_n$, for various values of n, including $n = 8, 10, 20$, and various values of μ such that $0 < \mu \leq 1/2$. Find some $\mu > 1/2$ causing divergence.

Problem 5.6. A method for computing the nth root $x^{1/n}$ of a positive real number $x \in \mathbb{R}$, with $n \in \mathbb{N}$ a positive integer $n \geq 2$, proceeds as follows: define the sequence (x_k), where x_0 is any chosen positive real, and

$$x_{k+1} = \frac{1}{n}\left((n-1)x_k + \frac{x}{x_k^{n-1}}\right), \quad k \geq 0.$$

(1) Implement the above method in `Matlab` and test it for various input values of x, x_0, and of $n \geq 2$, by running successively your program for $m = 2, 3, \ldots, 100$ iterations. Have your program plot the points (i, x_i) to watch how quickly the sequence converges.

Experiment with various choices of x_0. One of these choices should be $x_0 = x$. Compare your answers with the result of applying the of `Matlab` function $x \mapsto x^{1/n}$.

In some case, when x_0 is small, the number of iterations has to be at least 1000. Exhibit this behavior.

Problem 5.7. Refer to Problem 5.6 for the definition of the sequence (x_k).

(1) Define the *relative error* ϵ_k as

$$\epsilon_k = \frac{x_k}{x^{1/n}} - 1, \quad k \geq 0.$$

Prove that

$$\epsilon_{k+1} = \frac{x^{(1-1/n)}}{nx_k^{n-1}}\left(\frac{(n-1)x_k^n}{x} - \frac{nx_k^{n-1}}{x^{(1-1/n)}} + 1\right),$$

and then that

$$\epsilon_{k+1} = \frac{1}{n(\epsilon_k+1)^{n-1}}\left(\epsilon_k(\epsilon_k+1)^{n-2}((n-1)\epsilon_k + (n-2)) + 1 - (\epsilon_k+1)^{n-2}\right),$$

for all $k \geq 0$.

(2) Since

$$\epsilon_k + 1 = \frac{x_k}{x^{1/n}},$$

and since we assumed $x_0, x > 0$, we have $\epsilon_0 + 1 > 0$. We would like to prove that

$$\epsilon_k \geq 0, \quad \text{for all} \quad k \geq 1.$$

For this consider the variations of the function f given by

$$f(u) = (n-1)u^n - nx^{1/n}u^{n-1} + x,$$

for $u \in \mathbb{R}$.

Use the above to prove that $f(u) \geq 0$ for all $u \geq 0$. Conclude that

$$\epsilon_k \geq 0, \quad \text{for all} \quad k \geq 1.$$

(3) Prove that if $n = 2$, then

$$0 \leq \epsilon_{k+1} = \frac{\epsilon_k^2}{2(\epsilon_k + 1)}, \quad \text{for all} \quad k \geq 0,$$

else if $n \geq 3$, then

$$0 \leq \epsilon_{k+1} \leq \frac{(n-1)}{n}\epsilon_k, \quad \text{for all} \quad k \geq 1.$$

Prove that the sequence (x_k) converges to $x^{1/n}$ for every initial value $x_0 > 0$.

(4) When $n = 2$, we saw in Problem 5.7(3) that

$$0 \leq \epsilon_{k+1} = \frac{\epsilon_k^2}{2(\epsilon_k + 1)}, \quad \text{for all} \quad k \geq 0.$$

For $n = 3$, prove that

$$\epsilon_{k+1} = \frac{2\epsilon_k^2(3/2 + \epsilon_k)}{3(\epsilon_k + 1)^2}, \quad \text{for all} \quad k \geq 0,$$

and for $n = 4$, prove that

$$\epsilon_{k+1} = \frac{3\epsilon_k^2}{4(\epsilon_k + 1)^3}\left(2 + (8/3)\epsilon_k + \epsilon_k^2\right), \quad \text{for all} \quad k \geq 0.$$

Let μ_3 and μ_4 be the functions given by

$$\mu_3(a) = \frac{3}{2} + a$$
$$\mu_4(a) = 2 + \frac{8}{3}a + a^2,$$

so that if $n = 3$, then

$$\epsilon_{k+1} = \frac{2\epsilon_k^2\mu_3(\epsilon_k)}{3(\epsilon_k + 1)^2}, \quad \text{for all} \quad k \geq 0,$$

and if $n = 4$, then

$$\epsilon_{k+1} = \frac{3\epsilon_k^2\mu_4(\epsilon_k)}{4(\epsilon_k + 1)^3}, \quad \text{for all} \quad k \geq 0.$$

Prove that

$$a\mu_3(a) \leq (a+1)^2 - 1, \quad \text{for all} \quad a \geq 0,$$

and

$$a\mu_4(a) \leq (a+1)^3 - 1, \quad \text{for all} \quad a \geq 0.$$

Let $\eta_{3,k} = \mu_3(\epsilon_1)\epsilon_k$ when $n = 3$, and $\eta_{4,k} = \mu_4(\epsilon_1)\epsilon_k$ when $n = 4$. Prove that

$$\eta_{3,k+1} \leq \frac{2}{3}\eta_{3,k}^2, \quad \text{for all} \quad k \geq 1,$$

and

$$\eta_{4,k+1} \leq \frac{3}{4}\eta_{4,k}^2, \quad \text{for all} \quad k \geq 1.$$

Deduce from the above that the rate of convergence of $\eta_{i,k}$ is very fast, for $i = 3, 4$ (and $k \geq 1$).

Remark: If we let $\mu_2(a) = a$ for all a and $\eta_{2,k} = \epsilon_k$, we then proved that

$$\eta_{2,k+1} \leq \frac{1}{2}\eta_{2,k}^2, \quad \text{for all} \quad k \geq 1.$$

Problem 5.8. This is a continuation of Problem 5.7.

(1) Prove that for all $n \geq 2$, we have

$$\epsilon_{k+1} = \left(\frac{n-1}{n}\right)\frac{\epsilon_k^2 \mu_n(\epsilon_k)}{(\epsilon_k + 1)^{n-1}}, \quad \text{for all} \quad k \geq 0,$$

where μ_n is given by

$$\mu_n(a) = \frac{1}{2}n + \sum_{j=1}^{n-4} \frac{1}{n-1}\left((n-1)\binom{n-2}{j} + (n-2)\binom{n-2}{j+1} - \binom{n-2}{j+2}\right)a^j + \frac{n(n-2)}{n-1}a^{n-3} + a^{n-2}.$$

Furthermore, prove that μ_n can be expressed as

$$\mu_n(a) = \frac{1}{2}n + \frac{n(n-2)}{3}a + \sum_{j=2}^{n-4} \frac{(j+1)n}{(j+2)(n-1)}\binom{n-1}{j+1}a^j + \frac{n(n-2)}{n-1}a^{n-3} + a^{n-2}.$$

(2) Prove that for every j, with $1 \leq j \leq n-1$, the coefficient of a^j in $a\mu_n(a)$ is less than or equal to the coefficient of a^j in $(a+1)^{n-1} - 1$, and thus

$$a\mu_n(a) \leq (a+1)^{n-1} - 1, \quad \text{for all} \quad a \geq 0,$$

with strict inequality if $n \geq 3$. In fact, prove that if $n \geq 3$, then for every j, with $3 \leq j \leq n-2$, the coefficient of a^j in $a\mu_n(a)$ is strictly less than the coefficient of a^j in $(a+1)^{n-1} - 1$, and if $n \geq 4$, this also holds for $j = 2$.

Let $\eta_{n,k} = \mu_n(\epsilon_1)\epsilon_k$ $(n \geq 2)$. Prove that

$$\eta_{n,k+1} \leq \left(\frac{n-1}{n}\right)\eta_{n,k}^2, \quad \text{for all} \quad k \geq 1.$$

Chapter 6

Quadratic Optimization Problems

In this chapter we consider two classes of quadratic optimization problems that appear frequently in engineering and in computer science (especially in computer vision):

1. Minimizing
$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b$$
over all $x \in \mathbb{R}^n$, or subject to linear or affine constraints.

2. Minimizing
$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b$$
over the unit sphere.

In both cases, A is a symmetric matrix. We also seek necessary and sufficient conditions for Q to have a global minimum.

6.1 Quadratic Optimization: The Positive Definite Case

Many problems in physics and engineering can be stated as the minimization of some energy function, with or without constraints. Indeed, it is a fundamental principle of mechanics that nature acts so as to minimize energy. Furthermore, if a physical system is in a stable state of equilibrium, then the energy in that state should be minimal. For example, a small ball placed on top of a sphere is in an unstable equilibrium position. A small motion causes the ball to roll down. On the other hand, a ball placed inside and at the bottom of a sphere is in a stable equilibrium position because the potential energy is minimal.

The simplest kind of energy function is a quadratic function. Such functions can be conveniently defined in the form
$$Q(x) = x^\top Ax - x^\top b,$$

where A is a symmetric $n \times n$ matrix and x, b, are vectors in $\mathbb{R}^n$, viewed as column vectors. Actually, for reasons that will be clear shortly, it is preferable to put a factor $\frac{1}{2}$ in front of the quadratic term, so that

$$Q(x) = \frac{1}{2} x^\top A x - x^\top b.$$

The question is, under what conditions (on A) does $Q(x)$ have a global minimum, preferably unique?

We give a complete answer to the above question in two stages:

1. In this section we show that if A is symmetric positive definite, then $Q(x)$ has a unique global minimum precisely when
$$Ax = b.$$
2. In Section 6.2 we give necessary and sufficient conditions in the general case, in terms of the pseudo-inverse of A.

We begin with the matrix version of Definition 20.2 (Vol. I).

Definition 6.1. A symmetric *positive definite matrix* is a matrix whose eigenvalues are strictly positive, and a symmetric *positive semidefinite matrix* is a matrix whose eigenvalues are nonnegative.

Equivalent criteria are given in the following proposition.

Proposition 6.1. *Given any Euclidean space E of dimension n, the following properties hold:*

(1) Every self-adjoint linear map $f : E \to E$ is positive definite iff

$$\langle f(x), x\rangle > 0$$

for all $x \in E$ with $x \neq 0$.

(2) Every self-adjoint linear map $f : E \to E$ is positive semidefinite iff

$$\langle f(x), x\rangle \geq 0$$

for all $x \in E$.

Proof. (1) First assume that f is positive definite. Recall that every self-adjoint linear map has an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors, and let $\lambda_1, \ldots, \lambda_n$ be the corresponding eigenvalues. With respect to this basis, for every $x = x_1 e_1 + \cdots + x_n e_n \neq 0$, we have

$$\langle f(x), x\rangle = \Big\langle f\Big(\sum_{i=1}^n x_i e_i\Big), \sum_{i=1}^n x_i e_i\Big\rangle = \Big\langle \sum_{i=1}^n \lambda_i x_i e_i, \sum_{i=1}^n x_i e_i\Big\rangle = \sum_{i=1}^n \lambda_i x_i^2,$$

which is strictly positive, since $\lambda_i > 0$ for $i = 1, \ldots, n$, and $x_i^2 > 0$ for some i, since $x \neq 0$.

Conversely, assume that

$$\langle f(x), x\rangle > 0$$

for all $x \neq 0$. Then for $x = e_i$, we get

$$\langle f(e_i), e_i\rangle = \langle \lambda_i e_i, e_i\rangle = \lambda_i,$$

and thus $\lambda_i > 0$ for all $i = 1, \ldots, n$.

(2) As in (1), we have

$$\langle f(x), x\rangle = \sum_{i=1}^{n} \lambda_i x_i^2,$$

and since $\lambda_i \geq 0$ for $i = 1, \ldots, n$ because f is positive semidefinite, we have $\langle f(x), x\rangle \geq 0$, as claimed. The converse is as in (1) except that we get only $\lambda_i \geq 0$ since $\langle f(e_i), e_i\rangle \geq 0$. □

Some special notation is customary (especially in the field of convex optimization) to express that a symmetric matrix is positive definite or positive semidefinite.

Definition 6.2. Given any $n \times n$ symmetric matrix A we write $A \succeq 0$ if A is positive semidefinite and we write $A \succ 0$ if A is positive definite.

Remark: It should be noted that we can define the relation

$$A \succeq B$$

between any two $n \times n$ matrices (symmetric or not) iff $A - B$ is symmetric positive semidefinite. It is easy to check that this relation is actually a partial order on matrices, called the *positive semidefinite cone ordering*; for details, see Boyd and Vandenberghe [18], Section 2.4.

If A is symmetric positive definite, it is easily checked that A^{-1} is also symmetric positive definite. Also, if C is a symmetric positive definite $m \times m$ matrix and A is an $m \times n$ matrix of rank n (and so $m \geq n$ and the map $x \mapsto Ax$ is injective), then $A^\top CA$ is symmetric positive definite.

We can now prove that

$$Q(x) = \frac{1}{2} x^\top Ax - x^\top b$$

has a global minimum when A is symmetric positive definite.

Proposition 6.2. *Given a quadratic function*

$$Q(x) = \frac{1}{2} x^\top Ax - x^\top b,$$

if A is symmetric positive definite, then $Q(x)$ has a unique global minimum for the solution $x_0 = A^{-1}b$ of the linear system $Ax = b$. The minimum value of $Q(x)$ is

$$Q(A^{-1}b) = -\frac{1}{2} b^\top A^{-1} b.$$

Proof. Since A is positive definite, it is invertible since its eigenvalues are all strictly positive. Let $x_0 = A^{-1}b$, and compute $Q(y) - Q(x_0)$ for any $y \in \mathbb{R}^n$. Since $Ax_0 = b$, we get

$$\begin{aligned} Q(y) - Q(x_0) &= \frac{1}{2}y^\top Ay - y^\top b - \frac{1}{2}x_0^\top Ax_0 + x_0^\top b \\ &= \frac{1}{2}y^\top Ay - y^\top Ax_0 + \frac{1}{2}x_0^\top Ax_0 \\ &= \frac{1}{2}(y - x_0)^\top A(y - x_0). \end{aligned}$$

Since A is positive definite, the last expression is nonnegative, and thus

$$Q(y) \geq Q(x_0)$$

for all $y \in \mathbb{R}^n$, which proves that $x_0 = A^{-1}b$ is a global minimum of $Q(x)$. A simple computation yields

$$Q(A^{-1}b) = -\frac{1}{2}b^\top A^{-1}b.$$

□

Remarks:

(1) The quadratic function $Q(x)$ is also given by

$$Q(x) = \frac{1}{2}x^\top Ax - b^\top x,$$

but the definition using $x^\top b$ is more convenient for the proof of Proposition 6.2.

(2) If $Q(x)$ contains a constant term $c \in \mathbb{R}$, so that

$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b + c,$$

the proof of Proposition 6.2 still shows that $Q(x)$ has a unique global minimum for $x = A^{-1}b$, but the minimal value is

$$Q(A^{-1}b) = -\frac{1}{2}b^\top A^{-1}b + c.$$

Thus when the energy function $Q(x)$ of a system is given by a quadratic function

$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b,$$

where A is symmetric positive definite, finding the global minimum of $Q(x)$ is equivalent to solving the linear system $Ax = b$. Sometimes, it is useful to recast a linear problem $Ax = b$ as a variational problem (finding the minimum of some energy function). However, very often, a minimization problem comes with extra constraints that must be satisfied for all admissible solutions.

Example 6.1. For instance, we may want to minimize the quadratic function

$$Q(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2)$$

subject to the constraint

$$2x_1 - x_2 = 5.$$

The solution for which $Q(x_1, x_2)$ is minimum is no longer $(x_1, x_2) = (0, 0)$, but instead, $(x_1, x_2) = (2, -1)$, as will be shown later.

Geometrically, the graph of the function defined by $z = Q(x_1, x_2)$ in $\mathbb{R}^3$ is a paraboloid of revolution P with axis of revolution Oz. The constraint

$$2x_1 - x_2 = 5$$

corresponds to the vertical plane H parallel to the z-axis and containing the line of equation $2x_1 - x_2 = 5$ in the xy-plane. Thus, as illustrated by Figure 6.1, the constrained minimum of Q is located on the parabola that is the intersection of the paraboloid P with the plane H.

A nice way to solve constrained minimization problems of the above kind is to use the method of *Lagrange multipliers* discussed in Section 4.1. But first let us define precisely what kind of minimization problems we intend to solve.

Definition 6.3. The *quadratic constrained minimization problem* consists in minimizing a quadratic function

$$Q(x) = \frac{1}{2}x^\top A^{-1}x - b^\top x$$

subject to the linear constraints

$$B^\top x = f,$$

where A^{-1} is an $m \times m$ symmetric positive definite matrix, B is an $m \times n$ matrix of rank n (so that $m \geq n$), and where $b, x \in \mathbb{R}^m$ (viewed as column vectors), and $f \in \mathbb{R}^n$ (viewed as a column vector).

The reason for using A^{-1} instead of A is that the constrained minimization problem has an interpretation as a set of equilibrium equations in which the matrix that arises naturally is A (see Strang [76]). Since A and A^{-1} are both symmetric positive definite, this doesn't make any difference, but it seems preferable to stick to Strang's notation.

In Example 6.1 we have $m = 2$, $n = 1$,

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I_2, \quad b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad B = \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \quad f = 5.$$

As explained in Section 4.1, the method of Lagrange multipliers consists in incorporating the n constraints $B^\top x = f$ into the quadratic function $Q(x)$, by introducing extra variables

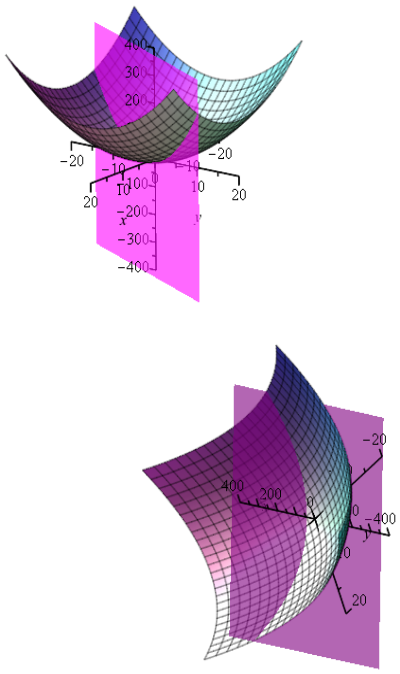


Figure 6.1: Two views of the constrained optimization problem $Q(x_1, x_2) = \frac{1}{2}\big(x_1^2 + x_2^2\big)$ subject to the constraint $2x_1 - x_2 = 5$. The minimum $(x_1, x_2) = (2, -1)$ is the the vertex of the parabolic curve formed the intersection of the magenta planar constraint with the bowl shaped surface.

$\lambda = (\lambda_1, \ldots, \lambda_n)$ called *Lagrange multipliers*, one for each constraint. We form the *Lagrangian*

$$L(x, \lambda) = Q(x) + \lambda^\top (B^\top x - f) = \frac{1}{2} x^\top A^{-1} x - (b - B\lambda)^\top x - \lambda^\top f.$$

We know from Theorem 4.2 that a necessary condition for our constrained optimization problem to have a solution is that $\nabla L(x, \lambda) = 0$. Since

$$\begin{aligned} \frac{\partial L}{\partial x}(x, \lambda) &= A^{-1} x - (b - B\lambda) \\ \frac{\partial L}{\partial \lambda}(x, \lambda) &= B^\top x - f, \end{aligned}$$

we obtain the system of linear equations

$$\begin{aligned} A^{-1} x + B\lambda &= b, \\ B^\top x &= f, \end{aligned}$$

which can be written in matrix form as

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}.$$

We shall prove in Proposition 6.3 below that our constrained minimization problem has a unique solution actually given by the above system.

Note that the matrix of this system is symmetric. We solve it as follows. Eliminating x from the first equation

$$A^{-1}x + B\lambda = b,$$

we get

$$x = A(b - B\lambda),$$

and substituting into the second equation, we get

$$B^\top A(b - B\lambda) = f,$$

that is,

$$B^\top AB\lambda = B^\top Ab - f.$$

However, by a previous remark, since A is symmetric positive definite and the columns of B are linearly independent, $B^\top AB$ is symmetric positive definite, and thus invertible. Thus we obtain the solution

$$\lambda = (B^\top AB)^{-1}(B^\top Ab - f), \qquad x = A(b - B\lambda).$$

Note that this way of solving the system requires solving for the Lagrange multipliers first.

Letting $e = b - B\lambda$, we also note that the system

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}$$

is equivalent to the system

$$\begin{aligned} e &= b - B\lambda, \\ x &= Ae, \\ B^\top x &= f. \end{aligned}$$

The latter system is called the *equilibrium equations* by Strang [76]. Indeed, Strang shows that the equilibrium equations of many physical systems can be put in the above form. This includes spring-mass systems, electrical networks and trusses, which are structures built from elastic bars. In each case, x, e, b, A, λ, f, and $K = B^\top AB$ have a physical interpretation. The matrix $K = B^\top AB$ is usually called the *stiffness matrix*. Again, the reader is referred to Strang [76].

In order to prove that our constrained minimization problem has a unique solution, we proceed to prove that the constrained minimization of $Q(x)$ subject to $B^\top x = f$ is equivalent to the unconstrained maximization of another function $-G(\lambda)$. We get $G(\lambda)$ by minimizing the Lagrangian $L(x, \lambda)$ treated as a function of x alone. The function $-G(\lambda)$ is the *dual function* of the Lagrangian $L(x, \lambda)$. Here we are encountering a special case of the notion of dual function defined in Section 14.7.

Since A^{-1} is symmetric positive definite and

$$L(x, \lambda) = \frac{1}{2} x^\top A^{-1} x - (b - B\lambda)^\top x - \lambda^\top f,$$

by Proposition 6.2 the global minimum (with respect to x) of $L(x, \lambda)$ is obtained for the solution x of

$$A^{-1} x = b - B\lambda,$$

that is, when

$$x = A(b - B\lambda),$$

and the minimum of $L(x, \lambda)$ is

$$\min_x L(x, \lambda) = -\frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) - \lambda^\top f.$$

Letting

$$G(\lambda) = \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f,$$

we will show in Proposition 6.3 that the solution of the constrained minimization of $Q(x)$ subject to $B^\top x = f$ is equivalent to the unconstrained maximization of $-G(\lambda)$. This is a special case of the duality discussed in Section 14.7.

Of course, since we minimized $L(x, \lambda)$ with respect to x, we have

$$L(x, \lambda) \geq -G(\lambda)$$

for all x and all λ. However, when the constraint $B^\top x = f$ holds, $L(x, \lambda) = Q(x)$, and thus for any admissible x, which means that $B^\top x = f$, we have

$$\min_x Q(x) \geq \max_\lambda -G(\lambda).$$

In order to prove that the unique minimum of the constrained problem $Q(x)$ subject to $B^\top x = f$ is the unique maximum of $-G(\lambda)$, we compute $Q(x) + G(\lambda)$.

Proposition 6.3. *The quadratic constrained minimization problem of Definition 6.3 has a unique solution (x, λ) given by the system*

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}.$$

Furthermore, the component λ of the above solution is the unique value for which $-G(\lambda)$ is maximum.

Proof. As we suggested earlier, let us compute $Q(x) + G(\lambda)$, assuming that the constraint $B^\top x = f$ holds. Eliminating f, since $b^\top x = x^\top b$ and $\lambda^\top B^\top x = x^\top B\lambda$, we get

$$\begin{aligned} Q(x) + G(\lambda) &= \frac{1}{2}x^\top A^{-1}x - b^\top x + \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f \\ &= \frac{1}{2}(A^{-1}x + B\lambda - b)^\top A(A^{-1}x + B\lambda - b). \end{aligned}$$

Since A is positive definite, the last expression is nonnegative. In fact, it is null iff

$$A^{-1}x + B\lambda - b = 0,$$

that is,

$$A^{-1}x + B\lambda = b.$$

But then the unique constrained minimum of $Q(x)$ subject to $B^\top x = f$ is equal to the unique maximum of $-G(\lambda)$ exactly when $B^\top x = f$ and $A^{-1}x + B\lambda = b$, which proves the proposition. □

We can confirm that the maximum of $-G(\lambda)$, or equivalently the minimum of

$$G(\lambda) = \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f,$$

corresponds to value of λ obtained by solving the system

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}.$$

Indeed, since

$$G(\lambda) = \frac{1}{2}\lambda^\top B^\top AB\lambda - \lambda^\top B^\top Ab + \lambda^\top f + \frac{1}{2}b^\top b,$$

and $B^\top AB$ is symmetric positive definite, by Proposition 6.2, the global minimum of $G(\lambda)$ is obtained when

$$B^\top AB\lambda - B^\top Ab + f = 0,$$

that is, $\lambda = (B^\top AB)^{-1}(B^\top Ab - f)$, as we found earlier.

Remarks:

(1) There is a form of duality going on in this situation. The constrained minimization of $Q(x)$ subject to $B^\top x = f$ is called the *primal problem*, and the unconstrained maximization of $-G(\lambda)$ is called the *dual problem*. Duality is the fact stated slightly loosely as

$$\min_x Q(x) = \max_\lambda -G(\lambda).$$

A general treatment of duality in constrained minimization problems is given in Section 14.7.

Recalling that $e = b - B\lambda$, since

$$G(\lambda) = \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f,$$

we can also write

$$G(\lambda) = \frac{1}{2}e^\top Ae + \lambda^\top f.$$

This expression often represents the total potential energy of a system. Again, the optimal solution is the one that minimizes the potential energy (and thus maximizes $-G(\lambda)$).

(2) It is immediately verified that the equations of Proposition 6.3 are equivalent to the equations stating that the partial derivatives of the Lagrangian $L(x, \lambda)$ are null:

$$\frac{\partial L}{\partial x_i} = 0, \quad i = 1, \ldots, m,$$
$$\frac{\partial L}{\partial \lambda_j} = 0, \quad j = 1, \ldots, n.$$

Thus, the constrained minimum of $Q(x)$ subject to $B^\top x = f$ is an extremum of the Lagrangian $L(x, \lambda)$. As we showed in Proposition 6.3, this extremum corresponds to simultaneously minimizing $L(x, \lambda)$ with respect to x and maximizing $L(x, \lambda)$ with respect to λ. Geometrically, such a point is a *saddle point* for $L(x, \lambda)$. Saddle points are discussed in Section 14.7.

(3) The Lagrange multipliers sometimes have a natural physical meaning. For example, in the spring-mass system they correspond to node displacements. In some general sense, Lagrange multipliers are correction terms needed to satisfy equilibrium equations and the price paid for the constraints. For more details, see Strang [76].

Going back to the constrained minimization of $Q(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2)$ subject to

$$2x_1 - x_2 = 5,$$

the Lagrangian is

$$L(x_1, x_2, \lambda) = \frac{1}{2}\big(x_1^2 + x_2^2\big) + \lambda(2x_1 - x_2 - 5),$$

and the equations stating that the Lagrangian has a saddle point are

$$\begin{aligned} x_1 + 2\lambda &= 0, \\ x_2 - \lambda &= 0, \\ 2x_1 - x_2 - 5 &= 0. \end{aligned}$$

We obtain the solution $(x_1, x_2, \lambda) = (2, -1, -1)$.

The use of Lagrange multipliers in optimization and variational problems is discussed extensively in Chapter 14.

Least squares methods and Lagrange multipliers are used to tackle many problems in computer graphics and computer vision; see Trucco and Verri [79], Metaxas [55], Jain, Katsuri, and Schunck [44], Faugeras [31], and Foley, van Dam, Feiner, and Hughes [32].

6.2 Quadratic Optimization: The General Case

In this section we complete the study initiated in Section 6.1 and give necessary and sufficient conditions for the quadratic function $\frac{1}{2}x^\top Ax - x^\top b$ to have a global minimum. We begin with the following simple fact:

Proposition 6.4. *If A is an invertible symmetric matrix, then the function*

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b$$

has a minimum value iff $A \succeq 0$, in which case this optimal value is obtained for a unique value of x, namely $x^* = A^{-1}b$, and with

$$f(A^{-1}b) = -\frac{1}{2}b^\top A^{-1}b.$$

Proof. Observe that

$$\frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) = \frac{1}{2}x^\top Ax - x^\top b + \frac{1}{2}b^\top A^{-1}b.$$

Thus,

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b = \frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) - \frac{1}{2}b^\top A^{-1}b.$$

If A has some negative eigenvalue, say $-\lambda$ (with $\lambda > 0$), if we pick any eigenvector u of A associated with λ, then for any $\alpha \in \mathbb{R}$ with $\alpha \neq 0$, if we let $x = \alpha u + A^{-1}b$, then since $Au = -\lambda u$, we get

$$\begin{aligned} f(x) &= \frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) - \frac{1}{2}b^\top A^{-1}b \\ &= \frac{1}{2}\alpha u^\top A\alpha u - \frac{1}{2}b^\top A^{-1}b \\ &= -\frac{1}{2}\alpha^2\lambda \left\|u\right\|_2^2 - \frac{1}{2}b^\top A^{-1}b, \end{aligned}$$

and since α can be made as large as we want and $\lambda > 0$, we see that f has no minimum. Consequently, in order for f to have a minimum, we must have $A \succeq 0$. If $A \succeq 0$, since A is invertible, it is positive definite, so $(x - A^{-1}b)^\top A(x - A^{-1}b) > 0$ iff $x - A^{-1}b \neq 0$, and it is clear that the minimum value of f is achieved when $x - A^{-1}b = 0$, that is, $x = A^{-1}b$. □

Let us now consider the case of an arbitrary symmetric matrix A.

Proposition 6.5. *If A is an $n \times n$ symmetric matrix, then the function*

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b$$

has a minimum value iff $A \succeq 0$ and $(I - AA^+)b = 0$, in which case this minimum value is

$$p^* = -\frac{1}{2}b^\top A^+ b.$$

Furthermore, if A is diagonalized as $A = U^\top \Sigma U$ (with U orthogonal), then the optimal value is achieved by all $x \in \mathbb{R}^n$ of the form

$$x = A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix},$$

for any $z \in \mathbb{R}^{n-r}$, where r is the rank of A.

Proof. The case that A is invertible is taken care of by Proposition 6.4, so we may assume that A is singular. If A has rank $r < n$, then we can diagonalize A as

$$A = U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} U,$$

where U is an orthogonal matrix and where Σ_r is an $r \times r$ diagonal invertible matrix. Then we have

$$\begin{aligned} f(x) &= \frac{1}{2}x^\top U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} Ux - x^\top U^\top Ub \\ &= \frac{1}{2}(Ux)^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} Ux - (Ux)^\top Ub. \end{aligned}$$

If we write

$$Ux = \begin{pmatrix} y \\ z \end{pmatrix} \quad \text{and} \quad Ub = \begin{pmatrix} c \\ d \end{pmatrix},$$

with $y, c \in \mathbb{R}^r$ and $z, d \in \mathbb{R}^{n-r}$, we get

$$\begin{aligned} f(x) &= \frac{1}{2}(Ux)^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} Ux - (Ux)^\top Ub \\ &= \frac{1}{2}(y^\top \; z^\top) \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} - (y^\top \; z^\top) \begin{pmatrix} c \\ d \end{pmatrix} \\ &= \frac{1}{2}y^\top \Sigma_r y - y^\top c - z^\top d. \end{aligned}$$

For $y = 0$, we get

$$f(x) = -z^\top d,$$

so if $d \neq 0$, the function f has no minimum. Therefore, if f has a minimum, then $d = 0$. However, $d = 0$ means that

$$Ub = \begin{pmatrix} c \\ 0 \end{pmatrix},$$

and we know from Proposition 21.5 (Vol. I) that b is in the range of A (here, U is $V^\top$), which is equivalent to $(I - AA^+)b = 0$. If $d = 0$, then

$$f(x) = \frac{1}{2} y^\top \Sigma_r y - y^\top c.$$

Consider the function $g\colon \mathbb{R}^r \to \mathbb{R}$ given by

$$g(y) = \frac{1}{2} y^\top \Sigma_r y - y^\top c, \quad y \in \mathbb{R}^r.$$

Since

$$\begin{pmatrix} y \\ z \end{pmatrix} = U^\top x$$

and $U^\top$ is invertible (with inverse U), when x ranges over $\mathbb{R}^n$, y ranges over the whole of $\mathbb{R}^r$, and since $f(x) = g(y)$, the function f has a minimum iff g has a minimum. Since Σ_r is invertible, by Proposition 6.4, the function g has a minimum iff $\Sigma_r \succeq 0$, which is equivalent to $A \succeq 0$.

Therefore, we have proven that if f has a minimum, then $(I - AA^+)b = 0$ and $A \succeq 0$. Conversely, if $(I - AA^+)b = 0$, then

$$\begin{aligned}\left(\begin{pmatrix} I_r & 0 \\ 0 & I_{n-r} \end{pmatrix} - U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} UU^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U\right) b &= \left(\begin{pmatrix} I_r & 0 \\ 0 & I_{n-r} \end{pmatrix} - U^\top \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} U\right) b \\ &= U^\top \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix} Ub = 0,\end{aligned}$$

which implies that if

$$Ub = \begin{pmatrix} c \\ d \end{pmatrix},$$

then $d = 0$, so as above

$$f(x) = g(y) = \frac{1}{2} y^\top \Sigma_r y - y^\top c,$$

and because $A \succeq 0$, we also have $\Sigma_r \succeq 0$, so g and f have a minimum.

When the above conditions hold, since

$$A = U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} U$$

is positive semidefinite, the pseudo-inverse A^+ of A is given by

$$A^+ = U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U,$$

and since

$$f(x) = g(y) = \frac{1}{2} y^\top \Sigma_r y - y^\top c,$$

by Proposition 6.4 the minimum of g is achieved iff $y^* = \Sigma_r^{-1} c$. Since $f(x)$ is independent of z, we can choose $z = 0$, and since $d = 0$, for x^* given by

$$Ux^* = \begin{pmatrix} \Sigma_r^{-1} c \\ 0 \end{pmatrix} \quad \text{and} \quad Ub = \begin{pmatrix} c \\ 0 \end{pmatrix},$$

we deduce that

$$x^* = U^\top \begin{pmatrix} \Sigma_r^{-1} c \\ 0 \end{pmatrix} = U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} c \\ 0 \end{pmatrix} = U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} Ub = A^+ b, \tag{$*$}$$

and the minimum value of f is

$$f(x^*) = \frac{1}{2}(A^+ b)^\top A A^+ b - b^\top A^+ b = \frac{1}{2} b^\top A^+ A A^+ b - b^\top A^+ b = -\frac{1}{2} b^\top A^+ b,$$

since A^+ is symmetric and $A^+ A A^+ = A^+$. For any $x \in \mathbb{R}^n$ of the form

$$x = A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix}, \quad z \in \mathbb{R}^{n-r},$$

since

$$x = A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix} = U^\top \begin{pmatrix} \Sigma_r^{-1} c \\ 0 \end{pmatrix} + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix} = U^\top \begin{pmatrix} \Sigma_r^{-1} c \\ z \end{pmatrix},$$

and since $f(x)$ is independent of z (because $f(x) = g(y)$), we have

$$f(x) = f(x^*) = -\frac{1}{2} b^\top A^+ b. \qquad \square$$

The problem of minimizing the function

$$f(x) = \frac{1}{2} x^\top A x - x^\top b$$

in the case where we add either linear constraints of the form $C^\top x = 0$ or affine constraints of the form $C^\top x = t$ (where $t \in \mathbb{R}^m$ and $t \neq 0$) where C is an $n \times m$ matrix can be reduced to the unconstrained case using a QR-decomposition of C. Let us show how to do this for linear constraints of the form $C^\top x = 0$.

If we use a QR decomposition of C, by permuting the columns of C to make sure that the first r columns of C are linearly independent (where $r = \text{rank}(C)$), we may assume that

$$C = Q^\top \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix} \Pi,$$

where Q is an $n \times n$ orthogonal matrix, R is an $r \times r$ invertible upper triangular matrix, S is an $r \times (m-r)$ matrix, and Π is a permutation matrix (C has rank r). Then if we let

$$x = Q^\top \begin{pmatrix} y \\ z \end{pmatrix},$$

where $y \in \mathbb{R}^r$ and $z \in \mathbb{R}^{n-r}$, then $C^\top x = 0$ becomes

$$C^\top x = \Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} Qx = \Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = 0,$$

which implies $y = 0$, and every solution of $C^\top x = 0$ is of the form

$$x = Q^\top \begin{pmatrix} 0 \\ z \end{pmatrix}.$$

Our original problem becomes

$$\begin{array}{ll} \text{minimize} & \dfrac{1}{2}(y^\top \; z^\top) QAQ^\top \begin{pmatrix} y \\ z \end{pmatrix} + (y^\top \; z^\top) Qb \\ \text{subject to} & y = 0,\ y \in \mathbb{R}^r,\ z \in \mathbb{R}^{n-r}. \end{array}$$

Thus, the constraint $C^\top x = 0$ has been simplified to $y = 0$, and if we write

$$QAQ^\top = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix},$$

where G_{11} is an $r \times r$ matrix and G_{22} is an $(n-r) \times (n-r)$ matrix and

$$Qb = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \quad b_1 \in \mathbb{R}^r,\ b_2 \in \mathbb{R}^{n-r},$$

our problem becomes

$$\text{minimize } \frac{1}{2} z^\top G_{22} z + z^\top b_2, \quad z \in \mathbb{R}^{n-r},$$

the problem solved in Proposition 6.5.

Constraints of the form $C^\top x = t$ (where $t \neq 0$) can be handled in a similar fashion. In this case, we may assume that C is an $n \times m$ matrix with full rank (so that $m \leq n$) and $t \in \mathbb{R}^m$. Then we use a QR-decomposition of the form

$$C = P \begin{pmatrix} R \\ 0 \end{pmatrix},$$

where P is an orthogonal $n \times n$ matrix and R is an $m \times m$ invertible upper triangular matrix. If we write

$$x = P \begin{pmatrix} y \\ z \end{pmatrix},$$

where $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^{n-m}$, the equation $C^\top x = t$ becomes

$$(R^\top \, 0) P^\top x = t,$$

that is,

$$(R^\top \, 0) \begin{pmatrix} y \\ z \end{pmatrix} = t,$$

which yields

$$R^\top y = t.$$

Since R is invertible, we get $y = (R^\top)^{-1} t$, and then it is easy to see that our original problem reduces to an unconstrained problem in terms of the matrix $P^\top A P$; the details are left as an exercise.

6.3 Maximizing a Quadratic Function on the Unit Sphere

In this section we discuss various quadratic optimization problems mostly arising from computer vision (image segmentation and contour grouping). These problems can be reduced to the following basic optimization problem: given an $n \times n$ real symmetric matrix A

$$\begin{array}{ll} \text{maximize} & x^\top A x \\ \text{subject to} & x^\top x = 1, \; x \in \mathbb{R}^n. \end{array}$$

In view of Proposition 21.10 (Vol. I), the maximum value of $x^\top A x$ on the unit sphere is equal to the largest eigenvalue λ_1 of the matrix A, and it is achieved for any unit eigenvector u_1 associated with λ_1. Similarly, the minimum value of $x^\top A x$ on the unit sphere is equal to the smallest eigenvalue λ_n of the matrix A, and it is achieved for any unit eigenvector u_n associated with λ_n.

A variant of the above problem often encountered in computer vision consists in minimizing $x^\top A x$ on the ellipsoid given by an equation of the form

$$x^\top B x = 1,$$

where B is a symmetric positive definite matrix. Since B is positive definite, it can be diagonalized as

$$B = Q D Q^\top,$$

where Q is an orthogonal matrix and D is a diagonal matrix,

$$D = \mathrm{diag}(d_1, \ldots, d_n),$$

with $d_i > 0$, for $i = 1, \ldots, n$. If we define the matrices $B^{1/2}$ and $B^{-1/2}$ by

$$B^{1/2} = Q\,\mathrm{diag}\left(\sqrt{d_1}, \ldots, \sqrt{d_n}\right) Q^\top$$

and

$$B^{-1/2} = Q\,\mathrm{diag}\left(1/\sqrt{d_1}, \ldots, 1/\sqrt{d_n}\right) Q^\top,$$

it is clear that these matrices are symmetric, that $B^{-1/2}BB^{-1/2} = I$, and that $B^{1/2}$ and $B^{-1/2}$ are mutual inverses. Then if we make the change of variable

$$x = B^{-1/2}y,$$

the equation $x^\top Bx = 1$ becomes $y^\top y = 1$, and the optimization problem

$$\begin{array}{ll} \text{minimize} & x^\top Ax \\ \text{subject to} & x^\top Bx = 1,\ x \in \mathbb{R}^n, \end{array}$$

is equivalent to the problem

$$\begin{array}{ll} \text{minimize} & y^\top B^{-1/2}AB^{-1/2}y \\ \text{subject to} & y^\top y = 1,\ y \in \mathbb{R}^n, \end{array}$$

where $y = B^{1/2}x$ and $B^{-1/2}AB^{-1/2}$ are symmetric.

The complex version of our basic optimization problem in which A is a Hermitian matrix also arises in computer vision. Namely, given an $n \times n$ complex Hermitian matrix A,

$$\begin{array}{ll} \text{maximize} & x^*Ax \\ \text{subject to} & x^*x = 1,\ x \in \mathbb{C}^n. \end{array}$$

Again by Proposition 21.10 (Vol. I), the maximum value of x^*Ax on the unit sphere is equal to the largest eigenvalue λ_1 of the matrix A, and it is achieved for any unit eigenvector u_1 associated with λ_1.

Remark: It is worth pointing out that if A is a *skew-Hermitian* matrix, that is, if $A^* = -A$, then x^*Ax is *pure imaginary or zero*.

Indeed, since $z = x^*Ax$ is a scalar, we have $z^* = \overline{z}$ (the conjugate of z), so we have

$$\overline{x^*Ax} = (x^*Ax)^* = x^*A^*x = -x^*Ax,$$

so $\overline{x^*Ax} + x^*Ax = 2\mathrm{Re}(x^*Ax) = 0$, which means that x^*Ax is pure imaginary or zero.

In particular, if A is a real matrix and if A is *skew-symmetric*, then

$$x^\top Ax = 0.$$

Thus, for any real matrix (symmetric or not),

$$x^\top Ax = x^\top H(A)x,$$

where $H(A) = (A + A^\top)/2$, the symmetric part of A.

There are situations in which it is necessary to add linear constraints to the problem of maximizing a quadratic function on the sphere. This problem was completely solved by Golub [37] (1973). The problem is the following: given an $n \times n$ real symmetric matrix A and an $n \times p$ matrix C,

$$\begin{array}{ll} \text{minimize} & x^\top Ax \\ \text{subject to} & x^\top x = 1,\ C^\top x = 0,\ x \in \mathbb{R}^n. \end{array}$$

As in Section 6.2, Golub shows that the linear constraint $C^\top x = 0$ can be eliminated as follows: if we use a QR decomposition of C, by permuting the columns, we may assume that

$$C = Q^\top \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix} \Pi,$$

where Q is an orthogonal $n \times n$ matrix, R is an $r \times r$ invertible upper triangular matrix, and S is an $r \times (p - r)$ matrix (assuming C has rank r). If we let

$$x = Q^\top \begin{pmatrix} y \\ z \end{pmatrix},$$

where $y \in \mathbb{R}^r$ and $z \in \mathbb{R}^{n-r}$, then $C^\top x = 0$ becomes

$$\Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} Qx = \Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = 0,$$

which implies $y = 0$, and every solution of $C^\top x = 0$ is of the form

$$x = Q^\top \begin{pmatrix} 0 \\ z \end{pmatrix}.$$

Our original problem becomes

$$\begin{array}{ll} \text{minimize} & (y^\top\ z^\top) QAQ^\top \begin{pmatrix} y \\ z \end{pmatrix} \\ \text{subject to} & z^\top z = 1,\ z \in \mathbb{R}^{n-r}, \\ & y = 0,\ y \in \mathbb{R}^r. \end{array}$$

Thus the constraint $C^\top x = 0$ has been simplified to $y = 0$, and if we write

$$QAQ^\top = \begin{pmatrix} G_{11} & G_{12} \\ G_{12}^\top & G_{22} \end{pmatrix},$$

our problem becomes

$$\begin{array}{ll} \text{minimize} & z^\top G_{22} z \\ \text{subject to} & z^\top z = 1,\ z \in \mathbb{R}^{n-r}, \end{array}$$

a standard eigenvalue problem.

Remark: There is a way of finding the eigenvalues of G_{22} which does not require the QR-factorization of C. Observe that if we let

$$J = \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix},$$

then

$$JQAQ^\top J = \begin{pmatrix} 0 & 0 \\ 0 & G_{22} \end{pmatrix},$$

and if we set

$$P = Q^\top JQ,$$

then

$$PAP = Q^\top JQAQ^\top JQ.$$

Now, $Q^\top JQAQ^\top JQ$ and $JQAQ^\top J$ have the same eigenvalues, so PAP and $JQAQ^\top J$ also have the same eigenvalues. It follows that the solutions of our optimization problem are among the eigenvalues of $K = PAP$, and at least r of those are 0. Using the fact that CC^+ is the projection onto the range of C, where C^+ is the pseudo-inverse of C, it can also be shown that

$$P = I - CC^+,$$

the projection onto the kernel of $C^\top$. So P can be computed directly in terms of C. In particular, when $n \geq p$ and C has full rank (the columns of C are linearly independent), then we know that $C^+ = (C^\top C)^{-1} C^\top$ and

$$P = I - C(C^\top C)^{-1} C^\top.$$

This fact is used by Cour and Shi [26] and implicitly by Yu and Shi [83].

The problem of adding affine constraints of the form $N^\top x = t$, where $t \neq 0$, also comes up in practice. At first glance, this problem may not seem harder than the linear problem in which $t = 0$, but it is. This problem was extensively studied in a paper by Gander, Golub, and von Matt [36] (1989).

Gander, Golub, and von Matt consider the following problem: Given an $(n+m)\times(n+m)$ real symmetric matrix A (with $n > 0$), an $(n+m)\times m$ matrix N with full rank, and a nonzero vector $t \in \mathbb{R}^m$ with $\left\|(N^\top)^+ t\right\| < 1$ (where $(N^\top)^+$ denotes the pseudo-inverse of $N^\top$),

$$\begin{array}{ll} \text{minimize} & x^\top Ax \\ \text{subject to} & x^\top x = 1,\ N^\top x = t,\ x \in \mathbb{R}^{n+m}. \end{array}$$

The condition $\left\|(N^\top)^+ t\right\| < 1$ ensures that the problem has a solution and is not trivial. The authors begin by proving that the affine constraint $N^\top x = t$ can be eliminated. One way to do so is to use a QR decomposition of N. If

$$N = P\begin{pmatrix} R \\ 0 \end{pmatrix},$$

where P is an orthogonal $(n+m)\times(n+m)$ matrix and R is an $m \times m$ invertible upper triangular matrix, then if we observe that

$$\begin{aligned} x^\top Ax &= x^\top PP^\top APP^\top x, \\ N^\top x &= (R^\top\ 0)P^\top x = t, \\ x^\top x &= x^\top PP^\top x = 1, \end{aligned}$$

and if we write

$$P^\top AP = \begin{pmatrix} B & \Gamma^\top \\ \Gamma & C \end{pmatrix},$$

where B is an $m \times m$ symmetric matrix, C is an $n \times n$ symmetric matrix, Γ is an $m \times n$ matrix, and

$$P^\top x = \begin{pmatrix} y \\ z \end{pmatrix},$$

with $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^n$, we then get

$$\begin{aligned} x^\top Ax &= y^\top By + 2z^\top \Gamma y + z^\top Cz, \\ R^\top y &= t, \\ y^\top y + z^\top z &= 1. \end{aligned}$$

Thus

$$y = (R^\top)^{-1}t,$$

and if we write

$$s^2 = 1 - y^\top y > 0$$

and

$$b = \Gamma y,$$

we get the simplified problem

$$\begin{array}{ll} \text{minimize} & z^\top C z + 2 z^\top b \\ \text{subject to} & z^\top z = s^2,\ z \in \mathbb{R}^m. \end{array}$$

Unfortunately, if $b \neq 0$, Proposition 21.10 (Vol. I) is no longer applicable. It is still possible to find the minimum of the function $z^\top C z + 2 z^\top b$ using Lagrange multipliers, but such a solution is too involved to be presented here. Interested readers will find a thorough discussion in Gander, Golub, and von Matt [36].

6.4 Summary

The main concepts and results of this chapter are listed below:

- Quadratic optimization problems; *quadratic functions*.
- Symmetric *positive definite* and *positive semidefinite* matrices.
- The *positive semidefinite cone ordering*.
- Existence of a global minimum when A is symmetric positive definite.
- Constrained quadratic optimization problems.
- *Lagrange multipliers*; *Lagrangian*.
- *Primal* and *dual* problems.
- Quadratic optimization problems: the case of a symmetric invertible matrix A.
- Quadratic optimization problems: the general case of a symmetric matrix A.
- Adding linear constraints of the form $C^\top x = 0$.
- Adding affine constraints of the form $C^\top x = t$, with $t \neq 0$.
- Maximizing a quadratic function over the unit sphere.
- Maximizing a quadratic function over an ellipsoid.
- Maximizing a Hermitian quadratic form.
- Adding linear constraints of the form $C^\top x = 0$.
- Adding affine constraints of the form $N^\top x = t$, with $t \neq 0$.

6.5 Problems

Problem 6.1. Prove that the relation

$$A \succeq B$$

between any two $n \times n$ matrices (symmetric or not) iff $A - B$ is symmetric positive semidefinite is indeed a partial order.

Problem 6.2. (1) Prove that if A is symmetric positive definite, then so is A^{-1}.

(2) Prove that if C is a symmetric positive definite $m \times m$ matrix and A is an $m \times n$ matrix of rank n (and so $m \geq n$ and the map $x \mapsto Ax$ is injective), then $A^\top CA$ is symmetric positive definite.

Problem 6.3. Find the minimum of the function

$$Q(x_1, x_2) = \frac{1}{2}(2x_1^2 + x_2^2)$$

subject to the constraint

$$x_1 - x_2 = 3.$$

Problem 6.4. Consider the problem of minimizing the function

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b$$

in the case where we add an affine constraint of the form $C^\top x = t$, with $t \in \mathbb{R}^m$ and $t \neq 0$, and where C is an $n \times m$ matrix of rank $m \leq n$. As in Section 6.2, use a QR-decomposition

$$C = P \begin{pmatrix} R \\ 0 \end{pmatrix},$$

where P is an orthogonal $n \times n$ matrix and R is an $m \times m$ invertible upper triangular matrix, and write

$$x = P \begin{pmatrix} y \\ z \end{pmatrix},$$

to deduce that

$$R^\top y = t.$$

Give the details of the reduction of this constrained minimization problem to an unconstrained minimization problem involving the matrix $P^\top AP$.

Problem 6.5. Find the maximum and the minimum of the function

$$Q(x, y) = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

on the unit circle $x^2 + y^2 = 1$.

Chapter 7

Schur Complements and Applications

Schur complements arise naturally in the process of inverting block matrices of the form

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

and in characterizing when symmetric versions of these matrices are positive definite or positive semidefinite. These characterizations come up in various quadratic optimization problems; see Boyd and Vandenberghe [18], especially Appendix B. In the most general case, pseudo-inverses are also needed.

In this chapter we introduce Schur complements and describe several interesting ways in which they are used. Along the way we provide some details and proofs of some results from Appendix A.5 (especially Section A.5.5) of Boyd and Vandenberghe [18].

7.1 Schur Complements

Let M be an $n \times n$ matrix written as a 2×2 block matrix

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where A is a $p \times p$ matrix and D is a $q \times q$ matrix, with $n = p + q$ (so B is a $p \times q$ matrix and C is a $q \times p$ matrix). We can try to solve the linear system

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c \\ d \end{pmatrix},$$

that is,

$$\begin{aligned} Ax + By &= c, \\ Cx + Dy &= d, \end{aligned}$$

by mimicking Gaussian elimination. If we assume that D is invertible, then we first solve for y, getting

$$y = D^{-1}(d - Cx),$$

and after substituting this expression for y in the first equation, we get

$$Ax + B(D^{-1}(d - Cx)) = c,$$

that is,

$$(A - BD^{-1}C)x = c - BD^{-1}d.$$

If the matrix $A - BD^{-1}C$ is invertible, then we obtain the solution to our system

$$\begin{aligned} x &= (A - BD^{-1}C)^{-1}(c - BD^{-1}d), \\ y &= D^{-1}(d - C(A - BD^{-1}C)^{-1}(c - BD^{-1}d)). \end{aligned}$$

If A is invertible, then by eliminating x first using the first equation, we obtain analogous formulas involving the matrix $D - CA^{-1}B$. The above formulas suggest that the matrices $A - BD^{-1}C$ and $D - CA^{-1}B$ play a special role and suggest the following definition:

Definition 7.1. Given any $n \times n$ block matrix of the form

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where A is a $p \times p$ matrix and D is a $q \times q$ matrix, with $n = p + q$ (so B is a $p \times q$ matrix and C is a $q \times p$ matrix), if D is invertible, then the matrix $A - BD^{-1}C$ is called the *Schur complement* of D in M. If A is invertible, then the matrix $D - CA^{-1}B$ is called the *Schur complement* of A in M.

The above equations written as

$$\begin{aligned} x &= (A - BD^{-1}C)^{-1}c - (A - BD^{-1}C)^{-1}BD^{-1}d, \\ y &= -D^{-1}C(A - BD^{-1}C)^{-1}c \\ &\quad + (D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1})d, \end{aligned}$$

yield a formula for the inverse of M in terms of the Schur complement of D in M, namely

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}.$$

A moment of reflection reveals that

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & 0 \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} \end{pmatrix} \begin{pmatrix} I & -BD^{-1} \\ 0 & I \end{pmatrix},$$

and then

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ -D^{-1}C & I \end{pmatrix} \begin{pmatrix} (A - BD^{-1}C)^{-1} & 0 \\ 0 & D^{-1} \end{pmatrix} \begin{pmatrix} I & -BD^{-1} \\ 0 & I \end{pmatrix}.$$

By taking inverses, we obtain the following result.

Proposition 7.1. *If the matrix D is invertible, then*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I & 0 \\ D^{-1}C & I \end{pmatrix}.$$

The above expression can be checked directly and has the advantage of requiring only the invertibility of D.

Remark: If A is invertible, then we can use the Schur complement $D - CA^{-1}B$ of A to obtain the following factorization of M:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix}.$$

If $D - CA^{-1}B$ is invertible, we can invert all three matrices above, and we get another formula for the inverse of M in terms of $(D - CA^{-1}B)$, namely,

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}.$$

If A, D and both Schur complements $A - BD^{-1}C$ and $D - CA^{-1}B$ are all invertible, by comparing the two expressions for M^{-1}, we get the (non-obvious) formula

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}.$$

Using this formula, we obtain another expression for the inverse of M involving the Schur complements of A and D (see Horn and Johnson [43]):

Proposition 7.2. *If A, D and both Schur complements $A - BD^{-1}C$ and $D - CA^{-1}B$ are all invertible, then*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}.$$

If we set $D = I$ and change B to $-B$, we get

$$(A + BC)^{-1} = A^{-1} - A^{-1}B(I - CA^{-1}B)^{-1}CA^{-1},$$

a formula known as the *matrix inversion lemma* (see Boyd and Vandenberghe [18], Appendix C.4, especially C.4.3).

7.2 Symmetric Positive Definite Matrices and Schur Complements

If we assume that our block matrix M is symmetric, so that A, D are symmetric and $C = B^\top$, then we see by Proposition 7.1 that M is expressed as

$$M = \begin{pmatrix} A & B \\ B^\top & D \end{pmatrix} = \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BD^{-1}B^\top & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix}^\top,$$

which shows that M is similar to a block diagonal matrix (obviously, the Schur complement, $A - BD^{-1}B^\top$, is symmetric). As a consequence, we have the following version of "Schur's trick" to check whether $M \succ 0$ for a symmetric matrix.

Proposition 7.3. *For any symmetric matrix M of the form*

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix},$$

if C is invertible, then the following properties hold:

(1) $M \succ 0$ iff $C \succ 0$ and $A - BC^{-1}B^\top \succ 0$.

(2) If $C \succ 0$, then $M \succeq 0$ iff $A - BC^{-1}B^\top \succeq 0$.

Proof. (1) Since C is invertible, we have

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} = \begin{pmatrix} I & BC^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BC^{-1}B^\top & 0 \\ 0 & C \end{pmatrix} \begin{pmatrix} I & BC^{-1} \\ 0 & I \end{pmatrix}^\top. \tag{$*$}$$

Observe that

$$\begin{pmatrix} I & BC^{-1} \\ 0 & I \end{pmatrix}^{-1} = \begin{pmatrix} I & -BC^{-1} \\ 0 & I \end{pmatrix},$$

so $(*)$ yields

$$\begin{pmatrix} I & -BC^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} \begin{pmatrix} I & -BC^{-1} \\ 0 & I \end{pmatrix}^\top = \begin{pmatrix} A - BC^{-1}B^\top & 0 \\ 0 & C \end{pmatrix},$$

and we know that for any symmetric matrix T, here $T = M$, and any invertible matrix N, here

$$N = \begin{pmatrix} I & -BC^{-1} \\ 0 & I \end{pmatrix},$$

the matrix T is positive definite ($T \succ 0$) iff $NTN^\top$ (which is obviously symmetric) is positive definite ($NTN^\top \succ 0$). But a block diagonal matrix is positive definite iff each diagonal block is positive definite, which concludes the proof.

(2) This is because for any symmetric matrix T and any invertible matrix N, we have $T \succeq 0$ iff $NTN^\top \succeq 0$. □

Another version of Proposition 7.3 using the Schur complement of A instead of the Schur complement of C also holds. The proof uses the factorization of M using the Schur complement of A (see Section 7.1).

Proposition 7.4. *For any symmetric matrix M of the form*

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix},$$

if A is invertible then the following properties hold:

(1) $M \succ 0$ iff $A \succ 0$ and $C - B^\top A^{-1} B \succ 0$.

(2) If $A \succ 0$, then $M \succeq 0$ iff $C - B^\top A^{-1} B \succeq 0$.

Here is an illustration of Proposition 7.4(2). Consider the nonlinear quadratic constraint

$$(Ax + b)^\top (Ax + b) \leq c^\top x + d,$$

were $A \in \mathrm{M}_n(\mathbb{R}), x, b, c \in \mathbb{R}^n$ and $d \in \mathbb{R}$. Since obviously $I = I_n$ is invertible and $I \succ 0$, we have

$$\begin{pmatrix} I & Ax + b \\ (Ax + b)^\top & c^\top x + d \end{pmatrix} \succeq 0$$

iff $c^\top x + d - (Ax + b)^\top (Ax + b) \succeq 0$ iff $(Ax + b)^\top (Ax + b) \leq c^\top x + d$, since the matrix (a scalar) $c^\top x + d - (Ax + b)^\top (Ax + b)$ is the Schur complement of I in the above matrix.

The trick of using Schur complements to convert nonlinear inequality constraints into linear constraints on symmetric matrices involving the semidefinite ordering $\succeq$ is used extensively to convert nonlinear problems into semidefinite programs; see Boyd and Vandenberghe [18].

When C is singular (or A is singular), it is still possible to characterize when a symmetric matrix M as above is positive semidefinite, but this requires using a version of the Schur complement involving the pseudo-inverse of C, namely $A - BC^+B^\top$ (or the Schur complement, $C - B^\top A^+ B$, of A). We use the criterion of Proposition 6.5, which tells us when a quadratic function of the form $\frac{1}{2} x^\top P x - x^\top b$ has a minimum and what this optimum value is (where P is a symmetric matrix).

7.3 Symmetric Positive Semidefinite Matrices and Schur Complements

We now return to our original problem, characterizing when a symmetric matrix

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix}$$

is positive semidefinite. Thus, we want to know when the function

$$f(x,y) = \begin{pmatrix} x^\top & y^\top \end{pmatrix} \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = x^\top Ax + 2x^\top By + y^\top Cy$$

has a minimum with respect to both x and y. If we hold y constant, Proposition 6.5 implies that $f(x,y)$ has a minimum iff $A \succeq 0$ and $(I - AA^+)By = 0$, and then the minimum value is

$$f(x^*,y) = -y^\top B^\top A^+ By + y^\top Cy = y^\top (C - B^\top A^+ B)y.$$

Since we want $f(x,y)$ to be uniformly bounded from below for all x, y, we must have $(I - AA^+)B = 0$. Now $f(x^*,y)$ has a minimum iff $C - B^\top A^+ B \succeq 0$. Therefore, we have established that $f(x,y)$ has a minimum over all x, y iff

$$A \succeq 0, \quad (I - AA^+)B = 0, \quad C - B^\top A^+ B \succeq 0.$$

Similar reasoning applies if we first minimize with respect to y and then with respect to x, but this time, the Schur complement $A - BC^+B^\top$ of C is involved. Putting all these facts together, we get our main result:

Theorem 7.5. *Given any symmetric matrix*

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix}$$

the following conditions are equivalent:

(1) $M \succeq 0$ *(*M *is positive semidefinite).*

(2) $A \succeq 0, \quad (I - AA^+)B = 0, \quad C - B^\top A^+ B \succeq 0.$

(3) $C \succeq 0, \quad (I - CC^+)B^\top = 0, \quad A - BC^+B^\top \succeq 0.$

If $M \succeq 0$ as in Theorem 7.5, then it is easy to check that we have the following factorizations (using the fact that $A^+AA^+ = A^+$ and $C^+CC^+ = C^+$):

$$\begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} = \begin{pmatrix} I & BC^+ \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BC^+B^\top & 0 \\ 0 & C \end{pmatrix} \begin{pmatrix} I & 0 \\ C^+B^\top & I \end{pmatrix}$$

and

$$\begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} = \begin{pmatrix} I & 0 \\ B^\top A^+ & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & C - B^\top A^+ B \end{pmatrix} \begin{pmatrix} I & A^+B \\ 0 & I \end{pmatrix}.$$

7.4 Summary

The main concepts and results of this chapter are listed below:

- Schur complements.
- The matrix inversion lemma.
- Symmetric positive definite matrices and Schur complements.
- Symmetric positive semidefinite matrices and Schur complements.

7.5 Problems

Problem 7.1. Prove that maximizing the function $g(\lambda)$ given by

$$g(\lambda) = c_0 + \lambda c_1 - (b_0 + \lambda b_1)^\top (A_0 + \lambda A_1)^+ (b_0 + \lambda b_1),$$

subject to

$$A_0 + \lambda A_1 \succeq 0, \quad b_0 + \lambda b_1 \in \mathrm{range}(A_0 + \lambda A_1),$$

with A_0, A_1 some $n \times n$ symmetric positive semidefinite matrices, $b_0, b_1 \in \mathbb{R}^n$, and $c_0, c_1 \in \mathbb{R}$, is equivalent to maximizing γ subject to the constraints

$$\begin{aligned} &\lambda \geq 0 \\ &\begin{pmatrix} A_0 + \lambda A_1 & b_0 + \lambda b_1 \\ (b_0 + \lambda b_1)^\top & c_0 + \lambda c_1 - \gamma \end{pmatrix} \succeq 0. \end{aligned}$$

Problem 7.2. Let $a_1, \ldots, a_m$ be m vectors in $\mathbb{R}^n$ and assume that they span $\mathbb{R}^n$.

(1) Prove that the matrix

$$\sum_{k=1}^{m} a_k a_k^\top$$

is symmetric positive definite.

(2) Define the matrix X by

$$X = \left(\sum_{k=1}^{m} a_k a_k^\top \right)^{-1}.$$

Prove that

$$\begin{pmatrix} \sum_{k=1}^{m} a_k a_k^\top & a_i \\ a_i^\top & 1 \end{pmatrix} \succeq 0, \quad i = 1, \ldots, m.$$

Deduce that

$$a_i^\top X a_i \leq 1, \quad 1 \leq i \leq m.$$

Problem 7.3. Consider the function g of Example 3.14 defined by

$$g(a, b, c) = \log(ac - b^2),$$

where $ac - b^2 > 0$. We found that the Hessian matrix of g is given by

$$Hg(a, b, c) = \frac{1}{(ac - b^2)^2} \begin{pmatrix} -c^2 & 2bc & -b^2 \\ 2bc & -2(b^2 + ac) & 2ab \\ -b^2 & 2ab & -a^2 \end{pmatrix}.$$

Use the Schur complement (of a^2) to prove that the matrix $-Hg(a, b, c)$ is symmetric positive definite if $ac - b^2 > 0$ and $a, c > 0$.

Part II

Linear Optimization

Chapter 8

Convex Sets, Cones, $\mathcal{H}$-Polyhedra

8.1 What is Linear Programming?

What is *linear programming*? At first glance, one might think that this is some style of computer programming. After all, there is imperative programming, functional programming, object-oriented programming, *etc.* The term linear programming is somewhat misleading, because it really refers to a method for *planning* with linear constraints, or more accurately, an *optimization method* where both the objective function and the constraints are linear.[1]

Linear programming was created in the late 1940's, one of the key players being George Dantzing, who invented the simplex algorithm. Kantorovitch also did some pioneering work on linear programming as early as 1939. The term *linear programming* has a military connotation because in the early 1950's it was used as a synonym for plans or schedules for training troops, logistical supply, resource allocation, *etc.* Unfortunately the term linear programming is well established and we are stuck with it.

Interestingly, even though originally most applications of linear programming were in the field of economics and industrial engineering, linear programming has become an important tool in theoretical computer science and in the theory of algorithms. Indeed, linear programming is often an effective tool for designing approximation algorithms to solve hard problems (typically NP-hard problems). Linear programming is also the "baby version" of convex programming, a very effective methodology which has received much attention in recent years.

Our goal is to present the mathematical underpinnings of linear programming, in particular the existence of an optimal solution if a linear program is feasible and bounded, and the duality theorem in linear programming, one of the deepest results in this field. The duality theorem in linear programming also has significant algorithmic implications but we do not discuss this here. We present the simplex algorithm, the dual simplex algorithm, and the primal dual algorithm. We also describe the tableau formalism for running the simplex

[1] Again, we witness another unfortunate abuse of terminology; the constraints are in fact *affine*.

algorithm and its variants. A particularly nice feature of the tableau formalism is that the update of a tableau can be performed using elementary row operations identical to the operations used during the reduction of a matrix to row reduced echelon form (rref). What differs is the criterion for the choice of the pivot.

However, we do not discuss other methods such as the ellipsoid method or interior points methods. For these more algorithmic issues, we refer the reader to standard texts on linear programming. In our opinion, one of the clearest (and among the most concise!) is Matousek and Gardner [54]; Chvatal [24] and Schrijver [67] are classics. Papadimitriou and Steiglitz [60] offers a very crisp presentation in the broader context of combinatorial optimization, and Bertsimas and Tsitsiklis [14] and Vanderbei [80] are very complete.

Linear programming has to do with maximizing a linear cost function $c_1x_1 + \cdots + c_nx_n$ with respect to m "linear" inequalities of the form

$$a_{i1}x_1 + \cdots + a_{in}x_n \leq b_i.$$

These constraints can be put together into an $m \times n$ matrix $A = (a_{ij})$, and written more concisely as

$$Ax \leq b.$$

For technical reasons that will appear clearer later on, it is often preferable to add the nonnegativity constaints $x_i \geq 0$ for $i = 1, \ldots, n$. We write $x \geq 0$. It is easy to show that every linear program is equivalent to another one satisfying the constraints $x \geq 0$, at the expense of adding new variables that are also constrained to be nonnegative. Let $\mathcal{P}(A, b)$ be the set of *feasible solutions* of our linear program given by

$$\mathcal{P}(A, b) = \{x \in \mathbb{R}^n \mid Ax \leq b,\ x \geq 0\}.$$

Then there are two basic questions:

(1) Is $\mathcal{P}(A, b)$ nonempty, that is, does our linear program have a chance to have a solution?

(2) Does the objective function $c_1x_1 + \cdots + c_nx_n$ have a maximum value on $\mathcal{P}(A, b)$?

The answer to both questions can be **no**. But if $\mathcal{P}(A, b)$ is nonempty and if the objective function is bounded above (on $\mathcal{P}(A, b)$), then it can be shown that the maximum of $c_1x_1 + \cdots + c_nx_n$ is achieved by some $x \in \mathcal{P}(A, b)$. Such a solution is called an *optimal solution*. Perhaps surprisingly, this result is not so easy to prove (unless one has the simplex method at his disposal). We will prove this result in full detail (see Proposition 9.1).

The reason why linear constraints are so important is that the domain of potential optimal solutions $\mathcal{P}(A, b)$ is *convex*. In fact, $\mathcal{P}(A, b)$ is a convex polyhedron which is the intersection of half-spaces cut out by affine hyperplanes. The objective function being linear is convex, and this is also a crucial fact. Thus, we are led to study convex sets, in particular those that arise from solutions of inequalities defined by affine forms, but also convex cones.

We give a brief introduction to these topics. As a reward, we provide several criteria for testing whether a system of inequalities

$$Ax \leq b,\ x \geq 0$$

has a solution or not in terms of versions of the *Farkas lemma* (see Proposition 14.3 and Proposition 11.4). Then we give a complete proof of the strong duality theorem for linear programming (see Theorem 11.7). We also discuss the complementary slackness conditions and show that they can be exploited to design an algorithm for solving a linear program that uses both the primal problem and its dual. This algorithm known as the *primal dual algorithm*, although not used much nowadays, has been the source of inspiration for a whole class of approximation algorithms also known as primal dual algorithms.

We hope that this chapter and the next three will be a motivation for learning more about linear programming, convex optimization, but also convex geometry. The "bible" in convex optimization is Boyd and Vandenberghe [18], and one of the best sources for convex geometry is Ziegler [84]. This is a rather advanced text, so the reader may want to begin with Gallier [35].

8.2 Affine Subsets, Convex Sets, Affine Hyperplanes, Half-Spaces

We view $\mathbb{R}^n$ as consisting of *column vectors* ($n \times 1$ matrices). As usual, row vectors represent *linear forms*, that is linear maps $\varphi\colon \mathbb{R}^n \to \mathbb{R}$, in the sense that the row vector y (a $1 \times n$ matrix) represents the linear form φ if $\varphi(x) = yx$ for all $x \in \mathbb{R}^n$. We denote the space of linear forms (row vectors) by $(\mathbb{R}^n)^*$.

Recall that a *linear combination* of vectors in $\mathbb{R}^n$ is an expression

$$\lambda_1 x_1 + \cdots + \lambda_m x_m$$

where $x_1, \ldots, x_m \in \mathbb{R}^n$ and where $\lambda_1, \ldots, \lambda_m$ are *arbitrary* scalars in $\mathbb{R}$. Given a sequence of vectors $S = (x_1, \ldots, x_m)$ with $x_i \in \mathbb{R}^n$, the set of all linear combinations of the vectors in S is the smallest (linear) subspace containing S called the *linear span* of S, and denoted $\mathrm{span}(S)$. A *linear subspace* of $\mathbb{R}^n$ is any nonempty subset of $\mathbb{R}^n$ closed under linear combinations.

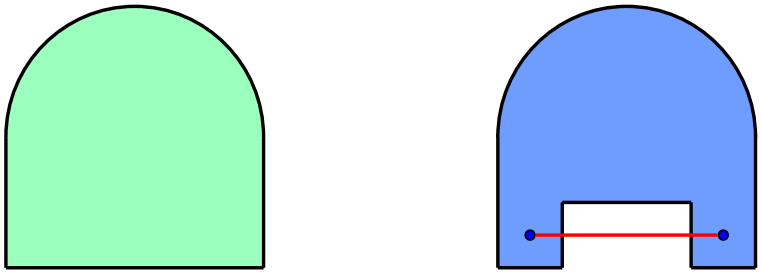
Definition 8.1. An *affine combination* of vectors in $\mathbb{R}^n$ is an expression

$$\lambda_1 x_1 + \cdots + \lambda_m x_m$$

where $x_1, \ldots, x_m \in \mathbb{R}^n$ and where $\lambda_1, \ldots, \lambda_m$ are scalars in $\mathbb{R}$ *satisfying the condition*

$$\lambda_1 + \cdots + \lambda_m = 1.$$

Given a sequence of vectors $S = (x_1, \ldots, x_m)$ with $x_i \in \mathbb{R}^n$, the set of all affine combinations of the vectors in S is the smallest affine subspace containing S called the *affine hull* of S and denoted $\mathrm{aff}(S)$.



(a) (b)

Figure 8.1: (a) A convex set; (b) A nonconvex set

Definition 8.2. An *affine subspace* A of $\mathbb{R}^n$ is any subset of $\mathbb{R}^n$ closed under affine combinations.

If A is a nonempty affine subspace of $\mathbb{R}^n$, then it can be shown that $V_A = \{a-b \mid a, b \in A\}$ is a linear subspace of $\mathbb{R}^n$ and that

$$A = a + V_A = \{a + v \mid v \in V_A\}$$

for *any* $a \in A$; see Gallier [34] (Section 2.5).

Definition 8.3. Given an affine subspace A, the linear space $V_A = \{a - b \mid a, b \in A\}$ is called the *direction of* A. The *dimension* of the nonempty affine subspace A is the dimension of its direction V_A.

Definition 8.4. *Convex combinations* are affine combinations $\lambda_1 x_1 + \cdots + \lambda_m x_m$ satisfying the extra condition that $\lambda_i \geq 0$ for $i = 1, \ldots, m$.

A convex set is defined as follows.

Definition 8.5. A subset V of $\mathbb{R}^n$ is *convex* if for any two points $a, b \in V$, we have $c \in V$ for every point $c = (1 - \lambda)a + \lambda b$, with $0 \leq \lambda \leq 1$ ($\lambda \in \mathbb{R}$). Given any two points a, b, the notation $[a, b]$ is often used to denote the line segment between a and b, that is,

$$[a, b] = \{c \in \mathbb{R}^n \mid c = (1 - \lambda)a + \lambda b,\ 0 \leq \lambda \leq 1\},$$

and thus a set V is convex if $[a, b] \subseteq V$ for any two points $a, b \in V$ ($a = b$ is allowed). The *dimension* of a convex set V is the dimension of its affine hull $\mathrm{aff}(A)$.

The empty set is trivially convex, every one-point set $\{a\}$ is convex, and the entire affine space $\mathbb{R}^n$ is convex.

It is obvious that the intersection of any family (finite or infinite) of convex sets is convex.

Definition 8.6. Given any (nonempty) subset S of $\mathbb{R}^n$, the smallest convex set containing S is denoted by $\mathrm{conv}(S)$ and called the *convex hull of* S (it is the intersection of all convex sets containing S).

It is essential not only to have a good understanding of $\mathrm{conv}(S)$, but to also have good methods for computing it. We have the following simple but crucial result.

Proposition 8.1. *For any family $S = (a_i)_{i\in I}$ of points in $\mathbb{R}^n$, the set V of convex combinations $\sum_{i\in I} \lambda_i a_i$ (where $\sum_{i\in I} \lambda_i = 1$ and $\lambda_i \geq 0$) is the convex hull* $\mathrm{conv}(S)$ *of $S = (a_i)_{i\in I}$.*

It is natural to wonder whether Proposition 8.1 can be sharpened in two directions: (1) Is it possible to have a fixed bound on the number of points involved in the convex combinations? (2) Is it necessary to consider convex combinations of all points, or is it possible to consider only a subset with special properties?

The answer is yes in both cases. In Case 1, Carathéodory's theorem asserts that it is enough to consider convex combinations of $n+1$ points. For example, in the plane $\mathbb{R}^2$, the convex hull of a set S of points is the union of all triangles (interior points included) with vertices in S. In Case 2, the theorem of Krein and Milman asserts that a convex set that is also compact is the convex hull of its extremal points (given a convex set S, a point $a \in S$ is *extremal* if $S - \{a\}$ is also convex).

We will not prove these theorems here, but we invite the reader to consult Gallier [35] or Berger [7].

Convex sets also arise as half-spaces cut out by affine hyperplanes.

Definition 8.7. An *affine form* $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ is defined by some linear form $c \in (\mathbb{R}^n)^*$ and some scalar $\beta \in \mathbb{R}$ so that

$$\varphi(x) = cx + \beta \quad \text{for all } x \in \mathbb{R}^n.$$

If $c \neq 0$, the affine form φ specified by (c, β) defines the *affine hyperplane* (for short *hyperplane*) $H(\varphi)$ given by

$$H(\varphi) = \{x \in \mathbb{R}^n \mid \varphi(x) = 0\} = \{x \in \mathbb{R}^n \mid cx + \beta = 0\},$$

and the two *(closed) half-spaces*

$$H_+(\varphi) = \{x \in \mathbb{R}^n \mid \varphi(x) \geq 0\} = \{x \in \mathbb{R}^n \mid cx + \beta \geq 0\},$$
$$H_-(\varphi) = \{x \in \mathbb{R}^n \mid \varphi(x) \leq 0\} = \{x \in \mathbb{R}^n \mid cx + \beta \leq 0\}.$$

When $\beta = 0$, we call H a *linear hyperplane*.

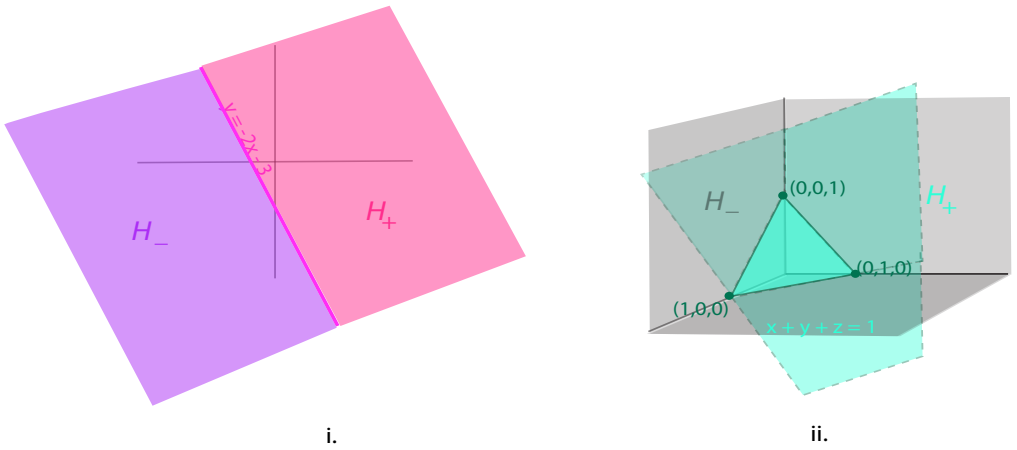


Figure 8.2: Figure i. illustrates the hyperplane $H(\varphi)$ for $\varphi(x, y) = 2x + y + 3$, while Figure ii. illustrates the hyperplane $H(\varphi)$ for $\varphi(x, y, z) = x + y + z - 1$.

Both $H_+(\varphi)$ and $H_-(\varphi)$ are convex and $H = H_+(\varphi) \cap H_-(\varphi)$.

For example, $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$ with $\varphi(x, y) = 2x + y + 3$ is an affine form defining the line given by the equation $y = -2x - 3$. Another example of an affine form is $\varphi\colon \mathbb{R}^3 \to \mathbb{R}$ with $\varphi(x, y, z) = x + y + z - 1$; this affine form defines the plane given by the equation $x + y + z = 1$, which is the plane through the points $(0, 0, 1), (0, 1, 0)$, and $(1, 0, 0)$. Both of these hyperplanes are illustrated in Figure 8.2.

Definition 8.8. For any two vector $x, y \in \mathbb{R}^n$ with $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ we write $x \leq y$ iff $x_i \leq y_i$ for $i = 1, \ldots, n$, and $x \geq y$ iff $y \leq x$. In particular $x \geq 0$ iff $x_i \geq 0$ for $i = 1, \ldots, n$.

Certain special types of convex sets called cones and $\mathcal{H}$-polyhedra play an important role. The set of feasible solutions of a linear program is an $\mathcal{H}$-polyhedron, and cones play a crucial role in the proof of Proposition 9.1 and in the Farkas–Minkowski proposition (Proposition 11.2).

8.3 Cones, Polyhedral Cones, and $\mathcal{H}$-Polyhedra

Cones and polyhedral cones are defined as follows.

Definition 8.9. Given a nonempty subset $S \subseteq \mathbb{R}^n$, the *cone* $C = \text{cone}(S)$ spanned by S is the convex set

$$\text{cone}(S) = \left\{ \sum_{i=1}^{k} \lambda_i u_i,\ u_i \in S, \lambda_i \in \mathbb{R},\ \lambda_i \geq 0 \right\},$$

of positive combinations of vectors from S. If S consists of a finite set of vectors, the cone $C = \text{cone}(S)$ is called a *polyhedral cone*. Figure 8.3 illustrates a polyhedral cone.

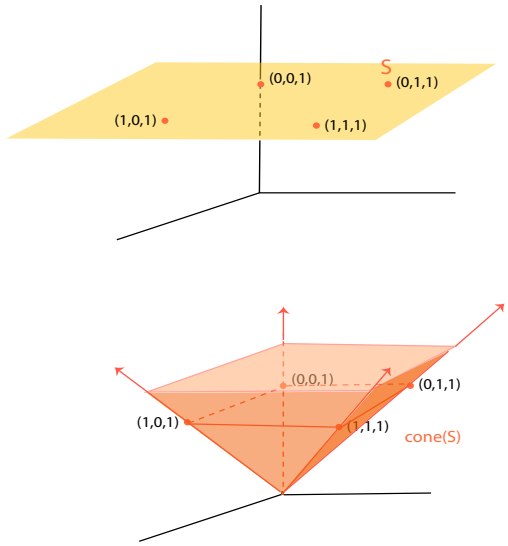


Figure 8.3: Let $S = \{(0,0,1),(1,0,1),(1,1,1),(0,1,1)\}$. The polyhedral cone, cone(S), is the solid "pyramid" with apex at the origin and square cross sections.

Note that if some nonzero vector u belongs to a cone C, then $\lambda u \in C$ for all $\lambda \geq 0$, that is, the *ray* $\{\lambda u \mid \lambda \geq 0\}$ belongs to C.

Remark: The cones (and polyhedral cones) of Definition 8.9 are *always convex*. For this reason, we use the simpler terminology cone instead of convex cone. However, there are more general kinds of cones (see Definition 14.1) that are not convex (for example, a union of polyhedral cones or the linear cone generated by the curve in Figure 8.4), and if we were dealing with those we would refer to the cones of Definition 8.9 as convex cones.

Definition 8.10. An *$\mathcal{H}$-polyhedron*, for short a *polyhedron*, is any subset $\mathcal{P} = \bigcap_{i=1}^{s} C_i$ of $\mathbb{R}^n$ defined as the intersection of a finite number s of closed half-spaces C_i. An example of an $\mathcal{H}$-polyhedron is shown in Figure 8.6. An *$\mathcal{H}$-polytope* is a bounded $\mathcal{H}$-polyhedron, which means that there is a closed ball $B_r(x)$ of center x and radius $r > 0$ such that $\mathcal{P} \subseteq B_r(x)$. An example of a $\mathcal{H}$-polytope is shown in Figure 8.5.

By convention, we agree that $\mathbb{R}^n$ itself is an $\mathcal{H}$-polyhedron.

Remark: The $\mathcal{H}$-polyhedra of Definition 8.10 are always convex. For this reason, as in the case of cones we use the simpler terminology $\mathcal{H}$-polyhedron instead of convex $\mathcal{H}$-polyhedron. In algebraic topology, there are more general polyhedra that are not convex.

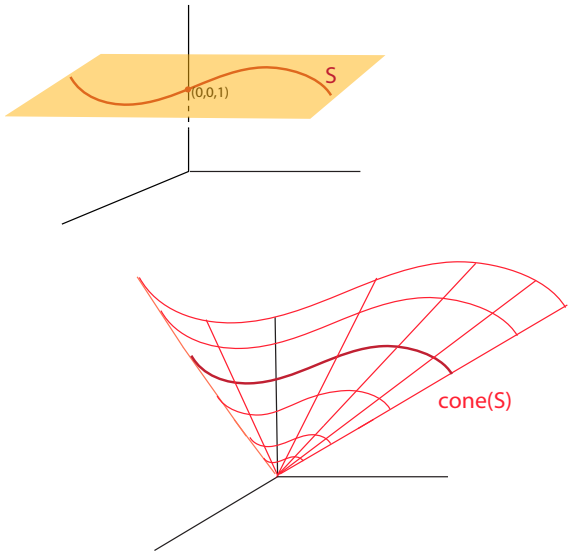


Figure 8.4: Let S be a planar curve in $z = 1$. The linear cone of S, consisting of all half rays connecting S to the origin, is not convex.

It can be shown that an $\mathcal{H}$-polytope $\mathcal{P}$ is equal to the convex hull of finitely many points (the extreme points of $\mathcal{P}$). This is a nontrivial result whose proof takes a significant amount of work; see Gallier [35] and Ziegler [84].

An unbounded $\mathcal{H}$-polyhedron is not equal to the convex hull of finite set of points. To obtain an equivalent notion we introduce the notion of a $\mathcal{V}$-polyhedron.

Definition 8.11. A *$\mathcal{V}$-polyhedron* is any convex subset $A \subseteq \mathbb{R}^n$ of the form

$$A = \mathrm{conv}(Y) + \mathrm{cone}(V) = \{a + v \mid a \in \mathrm{conv}(Y),\ v \in \mathrm{cone}(V)\},$$

where $Y \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^n$ are *finite* (possibly empty).

When $V = \emptyset$ we simply have a *polytope*, and when $Y = \emptyset$ or $Y = \{0\}$, we simply have a cone.

It can be shown that every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron and conversely. This is one of the major theorems in the theory of polyhedra, and its proof is nontrivial. For a complete proof, see Gallier [35] and Ziegler [84].

Every polyhedral cone is closed. This is an important fact that is used in the proof of several other key results such as Proposition 9.1 and the Farkas–Minkowski proposition (Proposition 11.2).

Although it seems obvious that a polyhedral cone should be closed, a rigorous proof is not entirely trivial.

Figure 8.5: An icosahedron is an example of an $\mathcal{H}$-polytope.

Indeed, the fact that a polyhedral cone is closed relies crucially on the fact that C is spanned by a *finite* number of vectors, because the cone generated by an infinite set may not be closed. For example, consider the closed disk $D \subseteq \mathbb{R}^2$ of center $(0, 1)$ and radius 1, which is tangent to the x-axis at the origin. Then the cone(D) consists of the open upper half-plane *plus* the origin $(0, 0)$, but this set is not closed.

Proposition 8.2. *Every polyhedral cone C is closed.*

Proof. This is proven by showing that

1. Every primitive cone is closed, where a *primitive cone* is a polyhedral cone spanned by linearly independent vectors.

2. A polyhedral cone C is the union of finitely many primitive cones.

Assume that $(a_1, \ldots, a_m)$ are linearly independent vectors in $\mathbb{R}^n$, and consider any sequence $(x^{(k)})_{k\geq 0}$

$$x^{(k)} = \sum_{i=1}^{m} \lambda_i^{(k)} a_i$$

of vectors in the primitive cone cone$(\{a_1, \ldots, a_m\})$, which means that $\lambda_j^{(k)} \geq 0$ for $i = 1, \ldots, m$ and all $k \geq 0$. The vectors $x^{(k)}$ belong to the subspace U spanned by $(a_1, \ldots, a_m)$, and U is closed. Assume that the sequence $(x^{(k)})_{k\geq 0}$ converges to a limit $x \in \mathbb{R}^n$. Since U is closed and $x^{(k)} \in U$ for all $k \geq 0$, we have $x \in U$. If we write $x = x_1 a_1 + \cdots + x_m a_m$, we would like to prove that $x_i \geq 0$ for $i = 1, \ldots, m$. The sequence the $(x^{(k)})_{k\geq 0}$ converges to x iff

$$\lim_{k \mapsto \infty} \left\| x^{(k)} - x \right\| = 0,$$

iff

$$\lim_{k \mapsto \infty} \Big(\sum_{i=1}^{m} |\lambda_i^{(k)} - x_i|^2 \Big)^{1/2} = 0$$

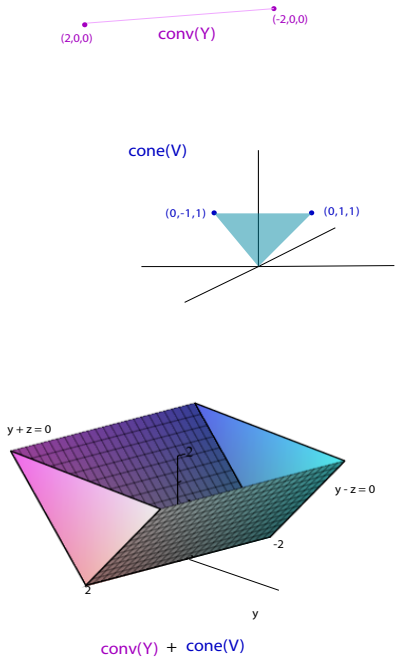


Figure 8.6: The "triangular trough" determined by the inequalities $y - z \leq 0$, $y + z \geq 0$, and $-2 \leq x \leq 2$ is an $\mathcal{H}$-polyhedron and an $\mathcal{V}$-polyhedron, where $Y = \{(2, 0, 0), (-2, 0, 0)\}$ and $V = \{(0, 1, 1), (0, -1, 1)\}$.

iff

$$\lim_{k \mapsto \infty} \lambda_i^{(k)} = x_i, \quad i = 1, \ldots, m.$$

Since $\lambda_i^{(k)} \geq 0$ for $i = 1, \ldots, m$ and all $k \geq 0$, we have $x_i \geq 0$ for $i = 1, \ldots, m$, so $x \in \text{cone}(\{a_1, \ldots, a_m\})$.

Next, assume that x belongs to the polyhedral cone C. Consider a positive combination

$$x = \lambda_1 a_1 + \cdots + \lambda_k a_k, \tag{$*_1$}$$

for some nonzero $a_1, \ldots, a_k \in C$, with $\lambda_i \geq 0$ and with k *minimal*. Since k is minimal, we must have $\lambda_i > 0$ for $i = 1, \ldots, k$. We claim that $(a_1, \ldots, a_k)$ are linearly independent.

If not, there is some nontrivial linear combination

$$\mu_1 a_1 + \cdots + \mu_k a_k = 0, \tag{$*_2$}$$

and since the a_i are nonzero, $\mu_j \neq 0$ for some at least some j. We may assume that $\mu_j < 0$ for some j (otherwise, we consider the family $(-\mu_i)_{1 \leq i \leq k}$), so let

$$J = \{j \in \{1, \ldots, k\} \mid \mu_j < 0\}.$$

For any $t \in \mathbb{R}$, since $x = \lambda_1 a_1 + \cdots + \lambda_k a_k$, using $(*_2)$ we get

$$x = (\lambda_1 + t\mu_1)a_1 + \cdots + (\lambda_k + t\mu_k)a_k, \tag{$*_3$}$$

and if we pick

$$t = \min_{j \in J}\left(-\frac{\lambda_j}{\mu_j}\right) \geq 0,$$

we have $(\lambda_i + t\mu_i) \geq 0$ for $i = 1, \ldots, k$, but $\lambda_j + t\mu_j = 0$ for some $j \in J$, so $(*_3)$ is an expression of x with less that k nonzero coefficients, contradicting the minimality of k in $(*_1)$. Therefore, $(a_1, \ldots, a_k)$ are linearly independent.

Since a polyhedral cone C is spanned by finitely many vectors, there are finitely many primitive cones (corresponding to linearly independent subfamilies), and since every $x \in C$, belongs to some primitive cone, C is the union of a finite number of primitive cones. Since every primitive cone is closed, as a union of finitely many closed sets, C itself is closed.

The above facts are also proven in Matousek and Gardner [54] (Chapter 6, Section 5, Lemma 6.5.3, 6.5.4, and 6.5.5). □

Another way to prove that a polyhedral cone C is closed is to show that C is also a $\mathcal{H}$-polyhedron. This takes even more work; see Gallier [35] (Chapter 4, Section 4, Proposition 4.16). Yet another proof is given in Lax [51] (Chapter 13, Theorem 1).

8.4 Summary

The main concepts and results of this chapter are listed below:

- Affine combination.
- Affine hull.
- Affine subspace; direction of an affine subspace, dimension of an affine subspace.
- Convex combination.
- Convex set, dimension of a convex set.
- Convex hull.
- Affine form.
- Affine hyperplane, half-spaces.
- Cone, polyhedral cone.
- $\mathcal{H}$-polyhedron, $\mathcal{H}$-polytope.
- $\mathcal{V}$-polyhedron, polytope.
- Primitive cone.

8.5 Problems

Problem 8.1. Prove Proposition 8.1.

Problem 8.2. Describe an icosahedron both as an $\mathcal{H}$-polytope and as a $\mathcal{V}$-polytope. Do the same thing for a dodecahedron. What do you observe?

Chapter 9

Linear Programs

In this chapter we introduce linear programs and the basic notions relating to this concept. We define the $\mathcal{H}$-polyhedron $\mathcal{P}(A, b)$ of feasible solutions. Then we define bounded and unbounded linear programs and the notion of optimal solution. We define slack variables and the important notion of *linear program in standard form*.

We show that if a linear program in standard form has a feasible solution and is bounded above, then it has an optimal solution. This is not an obvious result and the proof relies on the fact that a polyhedral cone is closed (this result was shown in the previous chapter).

Next we show that in order to find optimal solutions it suffices to consider solutions of a special form called *basic feasible solutions*. We prove that if a linear program in standard form has a feasible solution and is bounded above, then some basic feasible solution is an optimal solution (Theorem 9.4).

Geometrically, a basic feasible solution corresponds to a *vertex*. In Theorem 9.6 we prove that a basic feasible solution of a linear program in standard form is a vertex of the polyhedron $\mathcal{P}(A, b)$. Finally, we prove that if a linear program in standard form has some feasible solution, then it has a basic feasible solution (see Theorem 9.7). This fact allows the simplex algorithm described in the next chapter to get started.

9.1 Linear Programs, Feasible Solutions, Optimal Solutions

The purpose of linear programming is to solve the following type of optimization problem.

Definition 9.1. *A Linear Program* (P) is the following kind of optimization problem:

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \\ &\quad a_1 x \leq b_1 \\ &\quad \cdots \\ &\quad a_m x \leq b_m \\ &\quad x \geq 0, \end{aligned}$$

where $x \in \mathbb{R}^n$, $c, a_1, \ldots, a_m \in (\mathbb{R}^n)^*$, $b_1, \ldots, b_m \in \mathbb{R}$.

The linear form c defines the *objective function* $x \mapsto cx$ of the Linear Program (P) (from $\mathbb{R}^n$ to $\mathbb{R}$), and the inequalities $a_i x \leq b_i$ and $x_j \geq 0$ are called the *constraints* of the Linear Program (P).

If we define the $m \times n$ matrix

$$A = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}$$

whose rows are the row vectors $a_1, \ldots, a_m$ and b as the column vector

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix},$$

the m inequality constraints $a_i x \leq b_i$ can be written in matrix form as

$$Ax \leq b.$$

Thus the Linear Program (P) can also be stated as the Linear Program (P):

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0. \end{aligned}$$

We should note that in many applications, the natural primal optimization problem is actually the *minimization* of some objective function $cx = c_1 x_1 + \cdots + c_n x_n$, rather its maximization. For example, many of the optimization problems considered in Papadimitriou and Steiglitz [60] are minimization problems.

Of course, minimizing cx is equivalent to maximizing $-cx$, so our presentation covers minimization too.

Here is an explicit example of a linear program of Type (P):

Example 9.1.

$$\begin{aligned}
&\text{maximize} && x_1 + x_2 \\
&\text{subject to} && \\
& && x_2 - x_1 \leq 1 \\
& && x_1 + 6x_2 \leq 15 \\
& && 4x_1 - x_2 \leq 10 \\
& && x_1 \geq 0,\ x_2 \geq 0,
\end{aligned}$$

and in matrix form

$$\begin{aligned}
&\text{maximize} && \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
&\text{subject to} && \\
& && \begin{pmatrix} -1 & 1 \\ 1 & 6 \\ 4 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \leq \begin{pmatrix} 1 \\ 15 \\ 10 \end{pmatrix} \\
& && x_1 \geq 0,\ x_2 \geq 0.
\end{aligned}$$

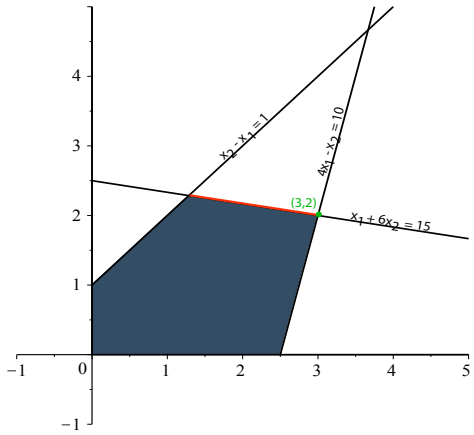


Figure 9.1: The $\mathcal{H}$-polyhedron associated with Example 9.1. The green point $(3, 2)$ is the unique optimal solution.

It turns out that $x_1 = 3, x_2 = 2$ yields the maximum of the objective function $x_1 + x_2$, which is 5. This is illustrated in Figure 9.1. Observe that the set of points that satisfy the above constraints is a convex region cut out by half planes determined by the lines of

equations

$$\begin{aligned} x_2 - x_1 &= 1 \\ x_1 + 6x_2 &= 15 \\ 4x_1 - x_2 &= 10 \\ x_1 &= 0 \\ x_2 &= 0. \end{aligned}$$

In general, each constraint $a_i x \leq b_i$ corresponds to the affine form φ_i given by $\varphi_i(x) = a_i x - b_i$ and defines the half-space $H_-(\varphi_i)$, and each inequality $x_j \geq 0$ defines the half-space $H_+(x_j)$. The intersection of these half-spaces is the set of solutions of all these constraints. It is a (possibly empty) $\mathcal{H}$-polyhedron denoted $\mathcal{P}(A, b)$.

Definition 9.2. If $\mathcal{P}(A, b) = \emptyset$, we say that the Linear Program (P) has *no feasible solution*, and otherwise any $x \in \mathcal{P}(A, b)$ is called a *feasible solution* of (P).

The linear program shown in Example 9.2 obtained by reversing the direction of the inequalities $x_2 - x_1 \leq 1$ and $4x_1 - x_2 \leq 10$ in the linear program of Example 9.1 has no feasible solution; see Figure 9.2.

Example 9.2.

$$\begin{aligned} &\text{maximize} && x_1 + x_2 \\ &\text{subject to} \\ & && x_1 - x_2 \leq -1 \\ & && x_1 + 6x_2 \leq 15 \\ & && x_2 - 4x_1 \leq -10 \\ & && x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

Assume $\mathcal{P}(A, b) \neq \emptyset$, so that the Linear Program (P) has a feasible solution. In this case, consider the image $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ of $\mathcal{P}(A, b)$ under the objective function $x \mapsto cx$.

Definition 9.3. If the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ is unbounded above, then we say that the Linear Program (P) is *unbounded*.

The linear program shown in Example 9.3 obtained from the linear program of Example 9.1 by deleting the constraints $4x_1 - x_2 \leq 10$ and $x_1 + 6x_2 \leq 15$ is unbounded.

Example 9.3.

$$\begin{aligned} &\text{maximize} && x_1 + x_2 \\ &\text{subject to} \\ & && x_2 - x_1 \leq 1 \\ & && x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

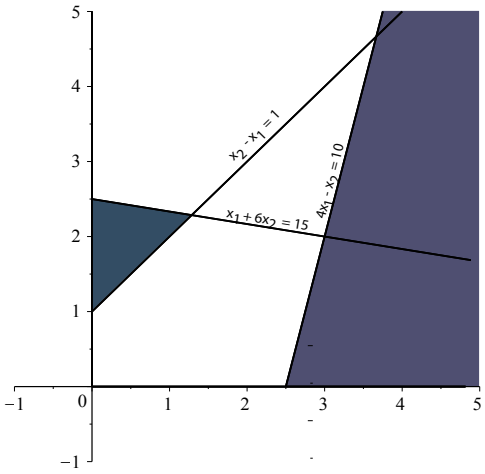


Figure 9.2: There is no $\mathcal{H}$-polyhedron associated with Example 9.2 since the blue and purple regions do not overlap.

Otherwise, we will prove shortly that if μ is the least upper bound of the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$, then there is some $p \in \mathcal{P}(A, b)$ such that

$$cp = \mu,$$

that is, the objective function $x \mapsto cx$ has a maximum value μ on $\mathcal{P}(A, b)$ which is achieved by some $p \in \mathcal{P}(A, b)$.

Definition 9.4. If the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ is nonempty and bounded above, any point $p \in \mathcal{P}(A, b)$ such that $cp = \max\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ is called an *optimal solution* (or *optimum*) of (P). Optimal solutions are often denoted by an upper $*$; for example, p^*.

The linear program of Example 9.1 has a unique optimal solution $(3, 2)$, but observe that the linear program of Example 9.4 in which the objective function is $(1/6)x_1 + x_2$ has infinitely many optimal solutions; the maximum of the objective function is $15/6$ which occurs along the points of orange boundary line in Figure 9.1.

Example 9.4.

$$\begin{aligned} &\text{maximize} && \frac{1}{6}x_1 + x_2 \\ &\text{subject to} \\ & && x_2 - x_1 \leq 1 \\ & && x_1 + 6x_2 \leq 15 \\ & && 4x_1 - x_2 \leq 10 \\ & && x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

The proof that if the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$ is nonempty and bounded above, then there is an optimal solution $p \in \mathcal{P}(A, b)$, is not as trivial as it might seem. It relies on the fact that a polyhedral cone is closed, a fact that was shown in Section 8.3.

We also use a trick that makes the proof simpler, which is that a Linear Program (P) with inequality constraints $Ax \leq b$

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

is equivalent to the Linear Program (P_2) with equality constraints

$$\begin{aligned} &\text{maximize} && \widehat{c}\,\widehat{x} \\ &\text{subject to} && \widehat{A}\widehat{x} = b \text{ and } \widehat{x} \geq 0, \end{aligned}$$

where $\widehat{A}$ is an $m \times (n+m)$ matrix, $\widehat{c}$ is a linear form in $(\mathbb{R}^{n+m})^*$, and $\widehat{x} \in \mathbb{R}^{n+m}$, given by

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad \widehat{c} = \begin{pmatrix} c & 0_m^\top \end{pmatrix}, \quad \text{and} \quad \widehat{x} = \begin{pmatrix} x \\ z \end{pmatrix},$$

with $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$.

Indeed, $\widehat{A}\widehat{x} = b$ and $\widehat{x} \geq 0$ iff

$$Ax + z = b, \quad x \geq 0,\ z \geq 0,$$

iff

$$Ax \leq b, \quad x \geq 0,$$

and $\widehat{c}\,\widehat{x} = cx$.

Definition 9.5. The variables z are called *slack variables*, and a linear program of the form (P_2) is called a linear program in *standard form*.

The result of converting the linear program of Example 9.4 to standard form is the program shown in Example 9.5.

Example 9.5.

$$\begin{aligned} &\text{maximize} && \frac{1}{6}x_1 + x_2 \\ &\text{subject to} && \\ & && x_2 - x_1 + z_1 = 1 \\ & && x_1 + 6x_2 + z_2 = 15 \\ & && 4x_1 - x_2 + z_3 = 10 \\ & && x_1 \geq 0,\ x_2 \geq 0,\ z_1 \geq 0,\ z_2 \geq 0,\ z_3 \geq 0. \end{aligned}$$

We can now prove that if a linear program has a feasible solution and is bounded, then it has an optimal solution.

Proposition 9.1. *Let (P_2) be a linear program in standard form, with equality constraint $Ax = b$. If $\mathcal{P}(A, b)$ is nonempty and bounded above, and if μ is the least upper bound of the set $\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$, then there is some $p \in \mathcal{P}(A, b)$ such that*

$$cp = \mu,$$

that is, the objective function $x \mapsto cx$ has a maximum value μ on $\mathcal{P}(A, b)$ which is achieved by some optimum solution $p \in \mathcal{P}(A, b)$.

Proof. Since $\mu = \sup\{cx \in \mathbb{R} \mid x \in \mathcal{P}(A, b)\}$, there is a sequence $(x^{(k)})_{k\geq 0}$ of vectors $x^{(k)} \in \mathcal{P}(A, b)$ such that $\lim_{k\mapsto\infty} cx^{(k)} = \mu$. In particular, if we write $x^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})$ we have $x_j^{(k)} \geq 0$ for $j = 1, \ldots, n$ and for all $k \geq 0$. Let $\widetilde{A}$ be the $(m + 1) \times n$ matrix

$$\widetilde{A} = \begin{pmatrix} c \\ A \end{pmatrix},$$

and consider the sequence $(\widetilde{A}x^{(k)})_{k\geq 0}$ of vectors $\widetilde{A}x^{(k)} \in \mathbb{R}^{m+1}$. We have

$$\widetilde{A}x^{(k)} = \begin{pmatrix} c \\ A \end{pmatrix} x^{(k)} = \begin{pmatrix} cx^{(k)} \\ Ax^{(k)} \end{pmatrix} = \begin{pmatrix} cx^{(k)} \\ b \end{pmatrix},$$

since by hypothesis $x^{(k)} \in \mathcal{P}(A, b)$, and the constraints are $Ax = b$ and $x \geq 0$. Since by hypothesis $\lim_{k\mapsto\infty} cx^{(k)} = \mu$, the sequence $(\widetilde{A}x^{(k)})_{k\geq 0}$ converges to the vector $\begin{pmatrix} \mu \\ b \end{pmatrix}$. Now, observe that each vector $\widetilde{A}x^{(k)}$ can be written as the convex combination

$$\widetilde{A}x^{(k)} = \sum_{j=1}^{n} x_j^{(k)} \widetilde{A}^j,$$

with $x_j^{(k)} \geq 0$ and where $\widetilde{A}^j \in \mathbb{R}^{m+1}$ is the jth column of $\widetilde{A}$. Therefore, $\widetilde{A}x^{(k)}$ belongs to the polyheral cone

$$C = \text{cone}(\widetilde{A}^1, \ldots, \widetilde{A}^n) = \{\widetilde{A}x \mid x \in \mathbb{R}^n,\ x \geq 0\},$$

and since by Proposition 8.2 this cone is closed, $\lim_{k\geq\infty} \widetilde{A}x^{(k)} \in C$, which means that there is some $u \in \mathbb{R}^n$ with $u \geq 0$ such that

$$\begin{pmatrix} \mu \\ b \end{pmatrix} = \lim_{k\geq\infty} \widetilde{A}x^{(k)} = \widetilde{A}u = \begin{pmatrix} cu \\ Au \end{pmatrix},$$

that is, $cu = \mu$ and $Au = b$. Hence, u is an optimal solution of (P_2). □

The next question is, how do we find such an optimal solution? It turns out that for linear programs in standard form where the constraints are of the form $Ax = b$ and $x \geq 0$, there are always optimal solutions of a special type called *basic feasible solutions*.

9.2 Basic Feasible Solutions and Vertices

If the system $Ax = b$ has a solution and if some row of A is a linear combination of other rows, then the corresponding equation is redundant, so we may assume that the rows of A are linearly independent; that is, we may assume that A has rank m, so $m \leq n$.

Definition 9.6. If A is an $m \times n$ matrix, for any nonempty subset K of $\{1, \ldots, n\}$, let A_K be the submatrix of A consisting of the columns of A whose indices belong to K. We denote the jth column of the matrix A by A^j.

Definition 9.7. Given a Linear Program (P_2)

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

where A has rank m, a vector $x \in \mathbb{R}^n$ is a *basic feasible solution* of (P) if $x \in \mathcal{P}(A, b) \neq \emptyset$, and if there is some subset K of $\{1, \ldots, n\}$ of size m such that

(1) The matrix A_K is invertible (that is, the columns of A_K are linearly independent).

(2) $x_j = 0$ for all $j \notin K$.

The subset K is called a *basis* of x. Every index $k \in K$ is called *basic*, and every index $j \notin K$ is called *nonbasic*. Similarly, the columns A^k corresponding to indices $k \in K$ are called *basic*, and the columns A^j corresponding to indices $j \notin K$ are called *nonbasic*. The variables corresponding to basic indices $k \in K$ are called *basic variables*, and the variables corresponding to indices $j \notin K$ are called *nonbasic*.

For example, the linear program

$$\begin{aligned} &\text{maximize} \quad x_1 + x_2 \\ &\text{subject to} \quad x_1 + x_2 + x_3 = 1 \text{ and } x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0, \qquad (*) \end{aligned}$$

has three basic feasible solutions; the basic feasible solution $K = \{1\}$ corresponds to the point $(1, 0, 0)$; the basic feasible solution $K = \{2\}$ corresponds to the point $(0, 1, 0)$; the basic feasible solution $K = \{3\}$ corresponds to the point $(0, 0, 1)$. Each of these points corresponds to the vertices of the slanted purple triangle illustrated in Figure 9.3. The vertices $(1, 0, 0)$ and $(0, 1, 0)$ optimize the objective function with a value of 1.

We now show that if the Standard Linear Program (P_2) as in Definition 9.7 has some feasible solution and is bounded above, then some basic feasible solution is an optimal solution. We follow Matousek and Gardner [54] (Chapter 4, Section 2, Theorem 4.2.3).

First we obtain a more convenient characterization of a basic feasible solution.

Proposition 9.2. *Given any Standard Linear Program (P_2) where $Ax = b$ and A is an $m \times n$ matrix of rank m, for any feasible solution x, if $J_> = \{j \in \{1, \ldots, n\} \mid x_j > 0\}$, then x is a basic feasible solution iff the columns of the matrix $A_{J_>}$ are linearly independent.*

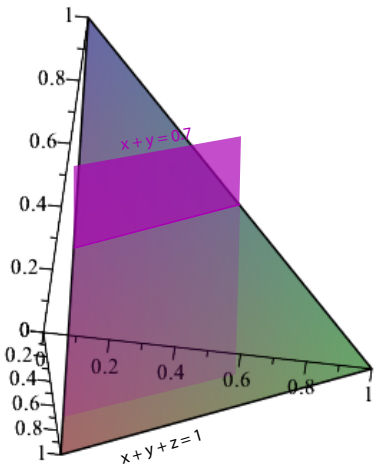


Figure 9.3: The $\mathcal{H}$-polytope associated with Linear Program $(*)$. The objective function (with $x_1 \to x$ and $x_2 \to y$) is represented by vertical planes parallel to the purple plane $x + y = 0.7$, and reaches it maximal value when $x + y = 1$.

Proof. If x is a basic feasible solution, then there is some subset $K \subseteq \{1, \ldots, n\}$ of size m such that the columns of A_K are linearly independent and $x_j = 0$ for all $j \notin K$, so by definition, $J_> \subseteq K$, which implies that the columns of the matrix $A_{J_>}$ are linearly independent.

Conversely, assume that x is a feasible solution such that the columns of the matrix $A_{J_>}$ are linearly independent. If $|J_>| = m$, we are done since we can pick $K = J_>$ and then x is a basic feasible solution. If $|J_>| < m$, we can extend $J_>$ to an m-element subset K by adding $m - |J_>|$ column indices so that the columns of A_K are linearly independent, which is possible since A has rank m. □

Next we prove that if a linear program in standard form has any feasible solution x_0 and is bounded above, then is has some basic feasible solution $\widetilde{x}$ which is as good as x_0, in the sense that $c\widetilde{x} \geq cx_0$.

Proposition 9.3. *Let (P_2) be any standard linear program with objective function cx, where $Ax = b$ and A is an $m \times n$ matrix of rank m. If (P_2) is bounded above and if x_0 is some feasible solution of (P_2), then there is some basic feasible solution $\widetilde{x}$ such that $c\widetilde{x} \geq cx_0$.*

Proof. Among the feasible solutions x such that $cx \geq cx_0$ (x_0 is one of them) pick one with the *maximum* number of coordinates x_j equal to 0, say $\widetilde{x}$. Let

$$K = J_> = \{j \in \{1, \ldots, n\} \mid \widetilde{x}_j > 0\}$$

and let $s = |K|$. We claim that $\widetilde{x}$ is a basic feasible solution, and by construction $c\widetilde{x} \geq cx_0$.

If the columns of A_K are linearly independent, then by Proposition 9.2 we know that $\widetilde{x}$ is a basic feasible solution and we are done.

Otherwise, the columns of A_K are linearly dependent, so there is some nonzero vector $v = (v_1, \ldots, v_s)$ such that $A_K v = 0$. Let $w \in \mathbb{R}^n$ be the vector obtained by extending v by setting $w_j = 0$ for all $j \notin K$. By construction,

$$Aw = A_K v = 0.$$

We will derive a contradiction by exhibiting a feasible solution $x(t_0)$ such that $cx(t_0) \geq cx_0$ with more zero coordinates than $\widetilde{x}$.

For this we claim that we may assume that w satisfies the following two conditions:

(1) $cw \geq 0$.

(2) There is some $j \in K$ such that $w_j < 0$.

If $cw = 0$ and if Condition (2) fails, since $w \neq 0$, we have $w_j > 0$ for some $j \in K$, in which case we can use $-w$, for which $w_j < 0$.

If $cw < 0$, then $c(-w) > 0$, so we may assume that $cw > 0$. If $w_j > 0$ for all $j \in K$, since $\widetilde{x}$ is feasible, $\widetilde{x} \geq 0$, and so $x(t) = \widetilde{x} + tw \geq 0$ for all $t \geq 0$. Furthermore, since $Aw = 0$ and $\widetilde{x}$ is feasible, we have

$$Ax(t) = A\widetilde{x} + tAw = b,$$

and thus $x(t)$ is feasible for all $t \geq 0$. We also have

$$cx(t) = c\widetilde{x} + tcw.$$

Since $cw > 0$, as $t > 0$ goes to infinity the objective function $cx(t)$ also tends to infinity, contradicting the fact that is is bounded above. Therefore, some w satisfying Conditions (1) and (2) above must exist.

We show that there is some $t_0 > 0$ such that $cx(t_0) \geq cx_0$ and $x(t_0) = \widetilde{x} + t_0 w$ is feasible, yet $x(t_0)$ has more zero coordinates than $\widetilde{x}$, a contradiction.

Since $x(t) = \widetilde{x} + tw$, we have

$$x(t)_i = \widetilde{x}_i + tw_i,$$

so if we let $I = \{i \in \{1, \ldots, n\} \mid w_i < 0\} \subseteq K$, which is nonempty since w satisfies Condition (2) above, if we pick

$$t_0 = \min_{i \in I} \left\{ \frac{-\widetilde{x}_i}{w_i} \right\},$$

then $t_0 > 0$, because $w_i < 0$ for all $i \in I$, and by definition of K we have $\widetilde{x}_i > 0$ for all $i \in K$. By the definition of $t_0 > 0$ and since $\widetilde{x} \geq 0$, we have

$$x(t_0)_j = \widetilde{x}_j + t_0 w_j \geq 0 \quad \text{for all } j \in K,$$

so $x(t_0) \geq 0$, and $x(t_0)_i = 0$ for some $i \in I$. Since $Ax(t_0) = b$ (for any t), $x(t_0)$ is a feasible solution,

$$cx(t_0) = c\widetilde{x} + t_0 cw \geq cx_0 + t_0 cw \geq cx_0,$$

and $x(t_0)_i = 0$ for some $i \in I$, we see that $x(t_0)$ has more zero coordinates than $\widetilde{x}$, a contradiction. □

Proposition 9.3 implies the following important result.

Theorem 9.4. *Let (P_2) be any standard linear program with objective function cx, where $Ax = b$ and A is an $m \times n$ matrix of rank m. If (P_2) has some feasible solution and if it is bounded above, then some basic feasible solution $\widetilde{x}$ is an optimal solution of (P_2).*

Proof. By Proposition 9.3, for any feasible solution x there is some basic feasible solution $\widetilde{x}$ such that $cx \leq c\widetilde{x}$. But there are only finitely many basic feasible solutions, so one of them has to yield the maximum of the objective function. □

Geometrically, basic solutions are exactly the vertices of the polyhedron $\mathcal{P}(A, b)$, a notion that we now define.

Definition 9.8. Given an $\mathcal{H}$-polyhedron $\mathcal{P} \subseteq \mathbb{R}^n$, a *vertex* of $\mathcal{P}$ is a point $v \in \mathcal{P}$ with property that there is some nonzero linear form $c \in (\mathbb{R}^n)^*$ and some $\mu \in \mathbb{R}$, such that v is the unique point of $\mathcal{P}$ for which the map $x \mapsto cx$ has the maximum value μ ; that is, $cy < cv = \mu$ for all $y \in \mathcal{P} - \{v\}$. Geometrically, this means that the hyperplane of equation $cy = \mu$ touches $\mathcal{P}$ exactly at v. More generally, a convex subset F of $\mathcal{P}$ is a *k-dimensional face* of $\mathcal{P}$ if F has dimension k and if there is some affine form $\varphi(x) = cx - \mu$ such that $cy = \mu$ for all $y \in F$, and $cy < \mu$ for all $y \in \mathcal{P} - F$. A 1-dimensional face is called an *edge*.

The concept of a vertex is illustrated in Figure 9.4, while the concept of an edge is illustrated in Figure 9.5.

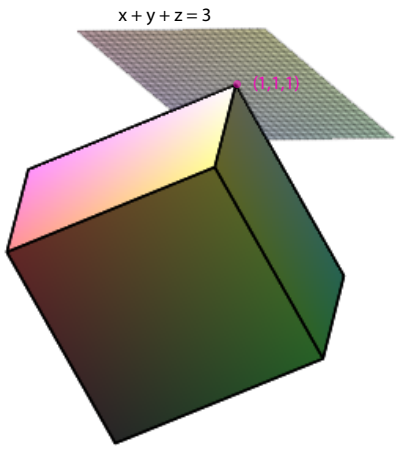


Figure 9.4: The cube centered at the origin with diagonal through $(-1, -1, -1)$ and $(1, 1, 1)$ has eight vertices. The vertex $(1, 1, 1)$ is associated with the linear form $x + y + z = 3$.

Since a k-dimensional face F of $\mathcal{P}$ is equal to the intersection of the hyperplane $H(\varphi)$ of equation $cx = \mu$ with $\mathcal{P}$, it is indeed convex and the notion of dimension makes sense.

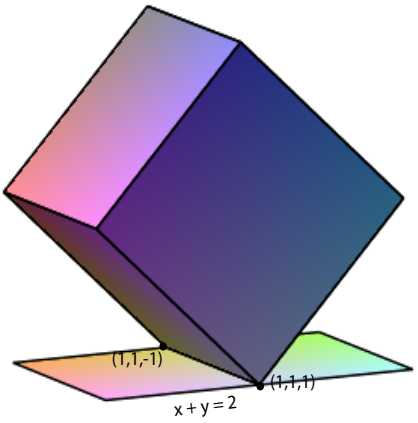


Figure 9.5: The cube centered at the origin with diagonal through $(-1,-1,-1)$ and $(1,1,1)$ has twelve edges. The edge from $(1,1,-1)$ to $(1,1,1)$ is associated with the linear form $x+y=2$.

Observe that a 0-dimensional face of $\mathcal{P}$ is a vertex. If $\mathcal{P}$ has dimension d, then the $(d-1)$-dimensional faces of $\mathcal{P}$ are called its *facets*.

If (P) is a linear program in standard form, then its basic feasible solutions are exactly the vertices of the polyhedron $\mathcal{P}(A,b)$. To prove this fact we need the following simple proposition

Proposition 9.5. *Let $Ax=b$ be a linear system where A is an $m\times n$ matrix of rank m. For any subset $K\subseteq\{1,\ldots,n\}$ of size m, if A_K is invertible, then there is at most one basic feasible solution $x\in\mathbb{R}^n$ with $x_j=0$ for all $j\notin K$ (of course, $x\geq 0$)*

Proof. In order for x to be feasible we must have $Ax=b$. Write $N=\{1,\ldots,n\}-K$, x_K for the vector consisting of the coordinates of x with indices in K, and x_N for the vector consisting of the coordinates of x with indices in N. Then

$$Ax=A_Kx_K+A_Nx_N=b.$$

In order for x to be a basic feasible solution we must have $x_N=0$, so

$$A_Kx_K=b.$$

Since by hypothesis A_K is invertible, $x_K=A_K^{-1}b$ is uniquely determined. If $x_K\geq 0$ then x is a basic feasible solution, otherwise it is not. This proves that there is at most one basic feasible solution $x\in\mathbb{R}^n$ with $x_j=0$ for all $j\notin K$. □

Theorem 9.6. *Let (P) be a linear program in standard form, where $Ax=b$ and A is an $m\times n$ matrix of rank m. For every $v\in\mathcal{P}(A,b)$, the following conditions are equivalent:*

(1) v is a vertex of the Polyhedron $\mathcal{P}(A, b)$.

(2) v is a basic feasible solution of the Linear Program (P).

Proof. First, assume that v is a vertex of $\mathcal{P}(A, b)$, and let $\varphi(x) = cx - \mu$ be a linear form such that $cy < \mu$ for all $y \in \mathcal{P}(A, b)$ and $cv = \mu$. This means that v is the unique point of $\mathcal{P}(A, b)$ for which the objective function $x \mapsto cx$ has the maximum value μ on $\mathcal{P}(A, b)$, so by Theorem 9.4, since this maximum is achieved by some basic feasible solution, by uniqueness v must be a basic feasible solution.

Conversely, suppose v is a basic feasible solution of (P) corresponding to a subset $K \subseteq \{1, \ldots, n\}$ of size m. Let $\widehat{c} \in (\mathbb{R}^n)^*$ be the linear form defined by

$$\widehat{c}_j = \begin{cases} 0 & \text{if } j \in K \\ -1 & \text{if } j \notin K. \end{cases}$$

By construction $\widehat{c}\, v = 0$ and $\widehat{c}\, x \leq 0$ for any $x \geq 0$, hence the function $x \mapsto \widehat{c}\, x$ on $\mathcal{P}(A, b)$ has a maximum at v. Furthermore, $\widehat{c}\, x < 0$ for any $x \geq 0$ such that $x_j > 0$ for some $j \notin K$. However, by Proposition 9.5, the vector v is the only basic feasible solution such that $v_j = 0$ for all $j \notin K$, and therefore v is the only point of $\mathcal{P}(A, b)$ maximizing the function $x \mapsto \widehat{c}\, x$, so it is a vertex. □

In theory, to find an optimal solution we try all $\binom{n}{m}$ possible m-elements subsets K of $\{1, \ldots, n\}$ and solve for the corresponding unique solution x_K of $A_K x = b$. Then we check whether such a solution satisfies $x_K \geq 0$, compute cx_K, and return some feasible x_K for which the objective function is maximum. This is a totally impractical algorithm.

A practical algorithm is the *simplex algorithm*. Basically, the simplex algorithm tries to "climb" in the polyhderon $\mathcal{P}(A, b)$ from vertex to vertex along edges (using basic feasible solutions), trying to maximize the objective function. We present the simplex algorithm in the next chapter. The reader may also consult texts on linear programming. In particular, we recommend Matousek and Gardner [54], Chvatal [24], Papadimitriou and Steiglitz [60], Bertsimas and Tsitsiklis [14], Ciarlet [25], Schrijver [67], and Vanderbei [80].

Observe that Theorem 9.4 asserts that if a Linear Program (P) in standard form (where $Ax = b$ and A is an $m \times n$ matrix of rank m) has some feasible solution and is bounded above, then some basic feasible solution is an optimal solution. By Theorem 9.6, the polyhedron $\mathcal{P}(A, b)$ must have some vertex.

But suppose we only know that $\mathcal{P}(A, b)$ is nonempty; that is, we don't know that the objective function cx is bounded above. Does $\mathcal{P}(A, b)$ have some vertex?

The answer to the above question is *yes*, and this is important because the simplex algorithm needs an initial basic feasible solution to get started. Here we prove that if $\mathcal{P}(A, b)$ is nonempty, then it must contain a vertex. This proof still doesn't constructively yield a vertex, but we will see in the next chapter that the simplex algorithm always finds a vertex if there is one (provided that we use a pivot rule that prevents cycling).

Theorem 9.7. *Let (P) be a linear program in standard form, where $Ax = b$ and A is an $m \times n$ matrix of rank m. If $\mathcal{P}(A, b)$ is nonempty (there is a feasible solution), then $\mathcal{P}(A, b)$ has some vertex; equivalently, (P) has some basic feasible solution.*

Proof. The proof relies on a trick, which is to add slack variables $x_{n+1}, \ldots, x_{n+m}$ and use the new objective function $-(x_{n+1} + \cdots + x_{n+m})$.

If we let $\widehat{A}$ be the $m \times (m+n)$-matrix, and x, $\overline{x}$, and $\widehat{x}$ be the vectors given by

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n, \quad \overline{x} = \begin{pmatrix} x_{n+1} \\ \vdots \\ x_{n+m} \end{pmatrix} \in \mathbb{R}^m, \quad \widehat{x} = \begin{pmatrix} x \\ \overline{x} \end{pmatrix} \in \mathbb{R}^{n+m},$$

then consider the Linear Program $(\widehat{P})$ in standard form

$$\begin{aligned} &\text{maximize} && -(x_{n+1} + \cdots + x_{n+m}) \\ &\text{subject to} && \widehat{A}\,\widehat{x} = b \text{ and } \widehat{x} \geq 0. \end{aligned}$$

Since $x_i \geq 0$ for all i, the objective function $-(x_{n+1} + \cdots + x_{n+m})$ is bounded above by 0. The system $\widehat{A}\,\widehat{x} = b$ is equivalent to the system

$$Ax + \overline{x} = b,$$

so for every feasible solution $u \in \mathcal{P}(A, b)$, since $Au = b$, the vector $(u, 0_m)$ is also a feasible solution of $(\widehat{P})$, in fact an optimal solution since the value of the objective function $-(x_{n+1} + \cdots + x_{n+m})$ for $\overline{x} = 0$ is 0. By Proposition 9.3, the linear program $(\widehat{P})$ has some basic feasible solution (u^*, w^*) for which the value of the objective function is greater than or equal to the value of the objective function for $(u, 0_m)$, and since $(u, 0_m)$ is an optimal solution, (u^*, w^*) is also an optimal solution of $(\widehat{P})$. This implies that $w^* = 0$, since otherwise the objective function $-(x_{n+1} + \cdots + x_{n+m})$ would have a strictly negative value.

Therefore, $(u^*, 0_m)$ is a basic feasible solution of $(\widehat{P})$, and thus the columns corresponding to nonzero components of u^* are linearly independent. Some of the coordinates of u^* could be equal to 0, but since A has rank m we can add columns of A to obtain a basis K associated with u^*, and u^* is indeed a basic feasible solution of (P). □

The definition of a basic feasible solution can be adapted to linear programs where the constraints are of the form $Ax \leq b$, $x \geq 0$; see Matousek and Gardner [54] (Chapter 4, Section 4, Definition 4.4.2).

The most general type of linear program allows constraints of the form $a_i x \geq b_i$ or $a_i x = b_i$ besides constraints of the form $a_i x \leq b_i$. The variables x_i may also take negative values. It is always possible to convert such programs to the type considered in Definition 9.1. We proceed as follows.

Every constraint $a_i x \geq b_i$ is replaced by the constraint $-a_i x \leq -b_i$. Every equality constraint $a_i x = b_i$ is replaced by the two constraints $a_i x \leq b_i$ and $-a_i x \leq -b_i$.

If there are n variables x_i, we create n new variables y_i and n new variables z_i and replace every variable x_i by $y_i - z_i$. We also add the $2n$ constraints $y_i \geq 0$ and $z_i \geq 0$. If the constraints are given by the inequalities $Ax \leq b$, we now have constraints given by

$$\begin{pmatrix} A & -A \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} \leq b, \quad y \geq 0,\ z \geq 0.$$

We replace the objective function cx by $cy - cz$.

Remark: We also showed that we can replace the inequality constraints $Ax \leq b$ by equality constraints $Ax = b$, by adding slack variables constrained to be nonnegative.

9.3 Summary

The main concepts and results of this chapter are listed below:

- Linear program.
- Objective function, constraints.
- Feasible solution.
- Bounded and unbounded linear programs.
- Optimal solution, optimum.
- Slack variables, linear program in standard form.
- Basic feasible solution.
- Basis of a variable.
- Basic, nonbasic index, basic, nonbasic variable.
- Vertex, face, edge, facet.

9.4 Problems

Problem 9.1. Convert the following program to standard form:

$$\begin{aligned} &\text{maximize} && x_1 + x_2 \\ &\text{subject to} && \\ & && x_2 - x_1 \leq 1 \\ & && x_1 + 6x_2 \leq 15 \\ & && -4x_1 + x_2 \geq 10. \end{aligned}$$

Problem 9.2. Convert the following program to standard form:

$$\begin{aligned} &\text{maximize} && 3x_1 - 2x_2 \\ &\text{subject to} && \\ & && 2x_1 - x_2 \leq 4 \\ & && x_1 + 3x_2 \geq 5 \\ & && x_2 \geq 0. \end{aligned}$$

Problem 9.3. The notion of basic feasible solution for linear programs where the constraints are of the form $Ax \leq b$, $x \geq 0$ is defined as follows. A basic feasible solution of a (general) linear program with n variables is a feasible solution for which some n linearly independent constraints hold with equality.

Prove that the definition of a basic feasible solution for linear programs in standard form is a special case of the above definition.

Problem 9.4. Consider the linear program

$$\begin{aligned} &\text{maximize} && x_1 + x_2 \\ &\text{subject to} && \\ & && x_1 + x_2 \leq 1. \end{aligned}$$

Show that none of the optimal solutions are basic.

Problem 9.5. The *standard n-simplex* is the subset Δ^n of $\mathbb{R}^{n+1}$ given by

$$\Delta^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1 + \cdots + x_{n+1} = 1,\ x_i \geq 0,\ 1 \leq i \leq n+1\}.$$

(1) Prove that Δ^n is convex and that it is the convex hull of the $n+1$ vectors $e_1, \ldots e_{n+1}$, where e_i is the ith canonical unit basis vector, $i = 1, \ldots, n+1$.

(2) Prove that Δ^n is the intersection of $n+1$ half spaces and determine the hyperplanes defining these half-spaces.

Remark: The volume under the standard simplex Δ^n is $1/(n+1)!$.

Problem 9.6. The n-dimensional *cross-polytope* is the subset XP_n of $\mathbb{R}^n$ given by

$$XP_n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid |x_1| + \cdots + |x_n| \leq 1\}.$$

(1) Prove that XP_n is convex and that it is the convex hull of the $2n$ vectors $\pm e_i$, where e_i is the ith canonical unit basis vector, $i = 1, \ldots, n$.

(2) Prove that XP_n is the intersection of 2^n half spaces and determine the hyperplanes defining these half-spaces.

Remark: The volume of XP_n is $2^n/n!$.

Problem 9.7. The n-dimensional *hypercube* is the subset C_n of $\mathbb{R}^n$ given by

$$C_n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid |x_i| \leq 1, 1 \leq i \leq n\}.$$

(1) Prove that C_n is convex and that it is the convex hull of the 2^n vectors $(\pm 1, \ldots, \pm 1)$, $i = 1, \ldots, n$.

(2) Prove that C_n is the intersection of $2n$ half spaces and determine the hyperplanes defining these half-spaces.

Remark: The volume of C_n is 2^n.

Chapter 10

The Simplex Algorithm

10.1 The Idea Behind the Simplex Algorithm

The simplex algorithm, due to Dantzig, applies to a linear program (P) in standard form, where the constraints are given by $Ax = b$ and $x \geq 0$, with A a $m \times n$ matrix of rank m, and with an objective function $x \mapsto cx$. This algorithm either reports that (P) has no feasible solution, or that (P) is unbounded, or yields an optimal solution. Geometrically, the algorithm climbs from vertex to vertex in the polyhedron $\mathcal{P}(A, b)$, trying to improve the value of the objective function. Since vertices correspond to basic feasible solutions, the simplex algorithm actually works with basic feasible solutions.

Recall that a basic feasible solution x is a feasible solution for which there is a subset $K \subseteq \{1, \ldots, n\}$ of size m such that the matrix A_K consisting of the columns of A whose indices belong to K are linearly independent, and that $x_j = 0$ for all $j \notin K$. We also let $J_>(x)$ be the set of indices

$$J_>(x) = \{j \in \{1, \ldots, n\} \mid x_j > 0\},$$

so for a basic feasible solution x associated with K, we have $J_>(x) \subseteq K$. In fact, by Proposition 9.2, a feasible solution x is a basic feasible solution iff the columns of $A_{J_>(x)}$ are linearly independent.

If $J_>(x)$ had cardinality m for all basic feasible solutions x, then the simplex algorithm would make progress at every step, in the sense that it would strictly increase the value of the objective function. Unfortunately, it is possible that $|J_>(x)| < m$ for certain basic feasible solutions, and in this case a step of the simplex algorithm may not increase the value of the objective function. Worse, in rare cases, it is possible that the algorithm enters an infinite loop. This phenomenon called *cycling* can be detected, but in this case the algorithm fails to give a conclusive answer.

Fortunately, there are ways of preventing the simplex algorithm from cycling (for example, Bland's rule discussed later), although proving that these rules work correctly is quite involved.

The potential "bad" behavior of a basic feasible solution is recorded in the following definition.

Definition 10.1. Given a Linear Program (P) in standard form where the constraints are given by $Ax = b$ and $x \geq 0$, with A an $m \times n$ matrix of rank m, a basic feasible solution x is *degenerate* if $|J_>(x)| < m$, otherwise it is *nondegenerate*.

The origin 0_n, if it is a basic feasible solution, is degenerate. For a less trivial example, $x = (0, 0, 0, 2)$ is a degenerate basic feasible solution of the following linear program in which $m = 2$ and $n = 4$.

Example 10.1.

$$\begin{aligned} &\text{maximize} && x_2 \\ &\text{subject to} && \\ & && -x_1 + x_2 + x_3 = 0 \\ & && x_1 + x_4 = 2 \\ & && x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0. \end{aligned}$$

The matrix A and the vector b are given by

$$A = \begin{pmatrix} -1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 2 \end{pmatrix},$$

and if $x = (0, 0, 0, 2)$, then $J_>(x) = \{4\}$. There are two ways of forming a set of two linearly independent columns of A containing the fourth column.

Given a basic feasible solution x associated with a subset K of size m, since the columns of the matrix A_K are linearly independent, by abuse of language we call the columns of A_K a *basis* of x.

If u is a vertex of (P), that is, a basic feasible solution of (P) associated with a basis K (of size m), in "normal mode," the simplex algorithm tries to move along an edge from the vertex u to an adjacent vertex v (with $u, v \in \mathcal{P}(A, b) \subseteq \mathbb{R}^n$) corresponding to a basic feasible solution whose basis is obtained by replacing one of the basic vectors A^k with $k \in K$ by another nonbasic vector A^j for some $j \notin K$, in such a way that the value of the objective function is increased.

Let us demonstrate this process on an example.

Example 10.2. Let (P) be the following linear program in standard form.

$$\begin{aligned} &\text{maximize} && x_1 + x_2 \\ &\text{subject to} && \\ & && -x_1 + x_2 + x_3 = 1 \\ & && x_1 + x_4 = 3 \\ & && x_2 + x_5 = 2 \\ & && x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0,\ x_5 \geq 0. \end{aligned}$$

The matrix A and the vector b are given by

$$A = \begin{pmatrix} -1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}.$$

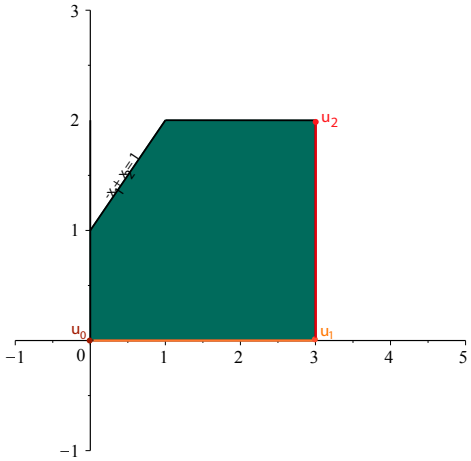


Figure 10.1: The planar $\mathcal{H}$-polyhedron associated with Example 10.2. The initial basic feasible solution is the origin. The simplex algorithm first moves along the horizontal orange line to feasible solution at vertex u_1. It then moves along the vertical red line to obtain the optimal feasible solution u_2.

The vector $u_0 = (0, 0, 1, 3, 2)$ corresponding to the basis $K = \{3, 4, 5\}$ is a basic feasible solution, and the corresponding value of the objective function is $0 + 0 = 0$. Since the columns (A^3, A^4, A^5) corresponding to $K = \{3, 4, 5\}$ are linearly independent we can express A^1 and A^2 as

$$\begin{aligned} A^1 &= -A^3 + A^4 \\ A^2 &= A^3 + A^5. \end{aligned}$$

Since

$$1A^3 + 3A^4 + 2A^5 = Au_0 = b,$$

for any $\theta \in \mathbb{R}$, we have

$$\begin{aligned} b &= 1A^3 + 3A^4 + 2A^5 - \theta A^1 + \theta A^1 \\ &= 1A^3 + 3A^4 + 2A^5 - \theta(-A^3 + A^4) + \theta A^1 \\ &= \theta A^1 + (1 + \theta)A^3 + (3 - \theta)A^4 + 2A^5, \end{aligned}$$

and

$$\begin{aligned} b &= 1A^3 + 3A^4 + 2A^5 - \theta A^2 + \theta A^2 \\ &= 1A^3 + 3A^4 + 2A^5 - \theta(A^3 + A^5) + \theta A^2 \\ &= \theta A^2 + (1-\theta)A^3 + 3A^4 + (2-\theta)A^5. \end{aligned}$$

In the first case, the vector $(\theta, 0, 1+\theta, 3-\theta, 2)$ is a feasible solution iff $0 \leq \theta \leq 3$, and the new value of the objective function is θ.

In the second case, the vector $(0, \theta, 1-\theta, 3, 2-\theta, 1)$ is a feasible solution iff $0 \leq \theta \leq 1$, and the new value of the objective function is also θ.

Consider the first case. It is natural to ask whether we can get another vertex and increase the objective function by setting to zero one of the coordinates of $(\theta, 0, 1+\theta, 3-\theta, 2)$, in this case the fouth one, by picking $\theta = 3$. This yields the feasible solution $(3, 0, 4, 0, 2)$, which corresponds to the basis (A^1, A^3, A^5), and so is indeed a basic feasible solution, with an improved value of the objective function equal to 3. Note that A^4 *left* the basis (A^3, A^4, A^5) and A^1 *entered* the new basis (A^1, A^3, A^5).

We can now express A^2 and A^4 in terms of the basis (A^1, A^3, A^5), which is easy to do since we already have A^1 and A^2 in term of (A^3, A^4, A^5), and A^1 and A^4 are swapped. Such a step is called a *pivoting step*. We obtain

$$\begin{aligned} A^2 &= A^3 + A^5 \\ A^4 &= A^1 + A^3. \end{aligned}$$

Then we repeat the process with $u_1 = (3, 0, 4, 0, 2)$ and the basis (A^1, A^3, A^5). We have

$$\begin{aligned} b &= 3A^1 + 4A^3 + 2A^5 - \theta A^2 + \theta A^2 \\ &= 3A^1 + 4A^3 + 2A^5 - \theta(A^3 + A^5) + \theta A^2 \\ &= 3A^1 + \theta A^2 + (4-\theta)A^3 + (2-\theta)A^5, \end{aligned}$$

and

$$\begin{aligned} b &= 3A^1 + 4A^3 + 2A^5 - \theta A^4 + \theta A^4 \\ &= 3A^1 + 4A^3 + 2A^5 - \theta(A^1 + A^3) + \theta A^4 \\ &= (3-\theta)A^1 + (4-\theta)A^3 + \theta A^4 + 2A^5. \end{aligned}$$

In the first case, the point $(3, \theta, 4-\theta, 0, 2-\theta)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the new value of the objective function is $3+\theta$. In the second case, the point $(3-\theta, 0, 4-\theta, \theta, 2)$ is a feasible solution iff $0 \leq \theta \leq 3$, and the new value of the objective function is $3-\theta$. To increase the objective function, we must choose the first case and we pick $\theta = 2$. Then we get the feasible solution $u_2 = (3, 2, 2, 0, 0)$, which corresponds to the basis (A^1, A^2, A^3), and thus is a basic feasible solution. The new value of the objective function is 5.

Next we express A^4 and A^5 in terms of the basis (A^1, A^2, A^3). Again this is easy to do since we just swapped A^5 and A^2 (a pivoting step), and we get

$$A^5 = A^2 - A^3$$
$$A^4 = A^1 + A^3.$$

We repeat the process with $u_2 = (3, 2, 2, 0, 0)$ and the basis (A^1, A^2, A^3). We have

$$\begin{aligned} b &= 3A^1 + 2A^2 + 2A^3 - \theta A^4 + \theta A^4 \\ &= 3A^1 + 2A^2 + 2A^3 - \theta(A^1 + A^3) + \theta A^4 \\ &= (3 - \theta)A^1 + 2A^2 + (2 - \theta)A^3 + \theta A^4, \end{aligned}$$

and

$$\begin{aligned} b &= 3A^1 + 2A^2 + 2A^3 - \theta A^5 + \theta A^5 \\ &= 3A^1 + 2A^2 + 2A^3 - \theta(A^2 - A^3) + \theta A^5 \\ &= 3A^1 + (2 - \theta)A^2 + (2 + \theta)A^3 + \theta A^5. \end{aligned}$$

In the first case, the point $(3 - \theta, 2, 2 - \theta, \theta, 0)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the value of the objective function is $5 - \theta$. In the second case, the point $(3, 2 - \theta, 2 + \theta, 0, \theta)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the value of the objective function is also $5 - \theta$. Since we must have $\theta \geq 0$ to have a feasible solution, there is no way to increase the objective function. In this situation, it turns out that we have reached an optimal solution, in our case $u_2 = (3, 2, 2, 0, 0)$, with the maximum of the objective function equal to 5.

We could also have applied the simplex algorithm to the vertex $u_0 = (0, 0, 1, 3, 2)$ and to the vector $(0, \theta, 1 - \theta, 3, 2 - \theta, 1)$, which is a feasible solution iff $0 \leq \theta \leq 1$, with new value of the objective function θ. By picking $\theta = 1$, we obtain the feasible solution $(0, 1, 0, 3, 1)$, corresponding to the basis (A^2, A^4, A^5), which is indeed a vertex. The new value of the objective function is 1. Then we express A^1 and A^3 in terms the basis (A^2, A^4, A^5) obtaining

$$A^1 = A^4 - A^3$$
$$A^3 = A^2 - A^5,$$

and repeat the process with $(0, 1, 0, 3, 1)$ and the basis (A^2, A^4, A^5). After three more steps we will reach the optimal solution $u_2 = (3, 2, 2, 0, 0)$.

Let us go back to the linear program of Example 10.1 with objective function x_2 and where the matrix A and the vector b are given by

$$A = \begin{pmatrix} -1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

Recall that $u_0 = (0, 0, 0, 2)$ is a degenerate basic feasible solution, and the objective function has the value 0. See Figure 10.2 for a planar picture of the $\mathcal{H}$-polyhedron associated with Example 10.1.

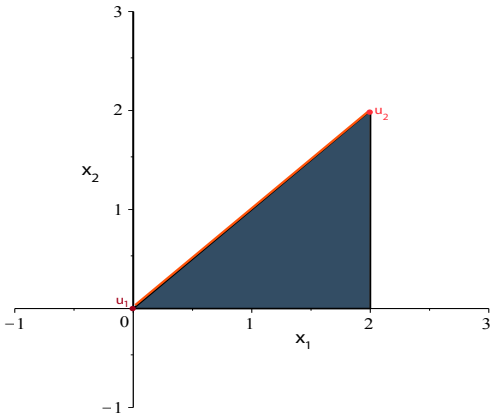


Figure 10.2: The planar $\mathcal{H}$-polyhedron associated with Example 10.1. The initial basic feasible solution is the origin. The simplex algorithm moves along the slanted orange line to the apex of the triangle.

Pick the basis (A^3, A^4). Then we have

$$\begin{aligned} A^1 &= -A^3 + A^4 \\ A^2 &= A^3, \end{aligned}$$

and we get

$$\begin{aligned} b &= 2A^4 - \theta A^1 + \theta A^1 \\ &= 2A^4 - \theta(-A^3 + A^4) + \theta A^1 \\ &= \theta A^1 + \theta A^3 + (2 - \theta)A^4, \end{aligned}$$

and

$$\begin{aligned} b &= 2A^4 - \theta A^2 + \theta A^2 \\ &= 2A^4 - \theta A^3 + \theta A^2 \\ &= \theta A^2 - \theta A^3 + 2A^4. \end{aligned}$$

In the first case, the point $(\theta, 0, \theta, 2 - \theta)$ is a feasible solution iff $0 \leq \theta \leq 2$, and the value of the objective function is 0, and in the second case the point $(0, \theta, -\theta, 2)$ is a feasible solution iff $\theta = 0$, and the value of the objective function is θ. However, since we must have $\theta = 0$ in the second case, there is no way to increase the objective function either.

It turns out that in order to make the cases considered by the simplex algorithm as mutually exclusive as possible, since in the second case the coefficient of θ in the value of the objective function is nonzero, namely 1, we should choose the second case. We must

pick $\theta = 0$, but we can swap the vectors A^3 and A^2 (because A^2 is coming in and A^3 has the coefficient $-\theta$, which is the reason why θ must be zero), and we obtain the basic feasible solution $u_1 = (0, 0, 0, 2)$ with the new basis (A^2, A^4). Note that this basic feasible solution corresponds to the same vertex $(0, 0, 0, 2)$ as before, but the basis has changed. The vectors A^1 and A^3 can be expressed in terms of the basis (A^2, A^4) as

$$\begin{aligned} A^1 &= -A^2 + A^4 \\ A^3 &= A^2. \end{aligned}$$

We now repeat the procedure with $u_1 = (0, 0, 0, 2)$ and the basis (A^2, A^4), and we get

$$\begin{aligned} b &= 2A^4 - \theta A^1 + \theta A^1 \\ &= 2A^4 - \theta(-A^2 + A^4) + \theta A^1 \\ &= \theta A^1 + \theta A^2 + (2 - \theta)A^4, \end{aligned}$$

and

$$\begin{aligned} b &= 2A^4 - \theta A^3 + \theta A^3 \\ &= 2A^4 - \theta A^2 + \theta A^3 \\ &= -\theta A^2 + \theta A^3 + 2A^4. \end{aligned}$$

In the first case, the point $(\theta, \theta, 0, 2-\theta)$ is a feasible solution iff $0 \leq \theta \leq 2$ and the value of the objective function is θ, and in the second case the point $(0, -\theta, \theta, 2)$ is a feasible solution iff $\theta = 0$ and the value of the objective function is θ. In order to increase the objective function we must choose the first case and pick $\theta = 2$. We obtain the feasible solution $u_2 = (2, 2, 0, 0)$ whose corresponding basis is (A^1, A^2) and the value of the objective function is 2.

The vectors A^3 and A^4 are expressed in terms of the basis (A^1, A^2) as

$$\begin{aligned} A^3 &= A^2 \\ A^4 &= A^1 + A^3, \end{aligned}$$

and we repeat the procedure with $u_2 = (2, 2, 0, 0)$ and the basis (A^1, A^2). We get

$$\begin{aligned} b &= 2A^1 + 2A^2 - \theta A^3 + \theta A^3 \\ &= 2A^1 + 2A^2 - \theta A^2 + \theta A^3 \\ &= 2A^1 + (2 - \theta)A^2 + \theta A^3, \end{aligned}$$

and

$$\begin{aligned} b &= 2A^1 + 2A^2 - \theta A^4 + \theta A^4 \\ &= 2A^1 + 2A^2 - \theta(A^1 + A^3) + \theta A^4 \\ &= (2 - \theta)A^1 + 2A^2 - \theta A^3 + \theta A^4. \end{aligned}$$

In the first case, the point $(2, 2-\theta, 0, \theta)$ is a feasible solution iff $0 \leq \theta \leq 2$ and the value of the objective function is $2-\theta$, and in the second case, the point $(2-\theta, 2, -\theta, \theta)$ is a feasible solution iff $\theta = 0$ and the value of the objective function is 2. This time there is no way to improve the objective function and we have reached an optimal solution $u_2 = (2, 2, 0, 0)$ with the maximum of the objective function equal to 2.

Let us now consider an example of an unbounded linear program.

Example 10.3. Let (P) be the following linear program in standard form.

$$\begin{aligned} &\text{maximize} && x_1 \\ &\text{subject to} && \\ & && x_1 - x_2 + x_3 = 1 \\ & && -x_1 + x_2 + x_4 = 2 \\ & && x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0. \end{aligned}$$

The matrix A and the vector b are given by

$$A = \begin{pmatrix} 1 & -1 & 1 & 0 \\ -1 & 1 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

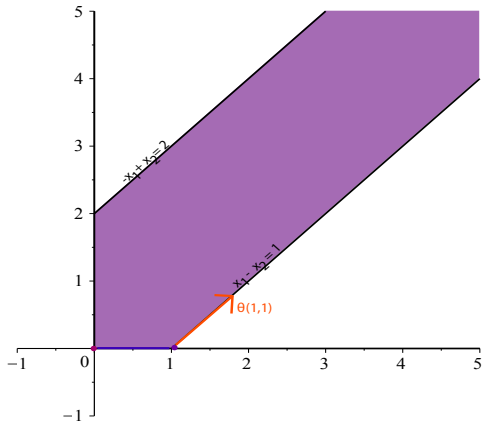


Figure 10.3: The planar $\mathcal{H}$-polyhedron associated with Example 10.3. The initial basic feasible solution is the origin. The simplex algorithm first moves along the horizontal indigo line to basic feasible solution at vertex $(1, 0)$. Any optimal feasible solution occurs by moving along the boundary line parameterized by the orange arrow $\theta(1, 1)$.

The vector $u_0 = (0, 0, 1, 2)$ corresponding to the basis $K = \{3, 4\}$ is a basic feasible solution, and the corresponding value of the objective function is 0. The vectors A^1 and A^2

are expressed in terms of the basis (A^3, A^4) by

$$\begin{aligned} A^1 &= A^3 - A^4 \\ A^2 &= -A^3 + A^4. \end{aligned}$$

Starting with $u_0 = (0, 0, 1, 2)$, we get

$$\begin{aligned} b &= A^3 + 2A^4 - \theta A^1 + \theta A^1 \\ &= A^3 + 2A^4 - \theta(A^3 - A^4) + \theta A^1 \\ &= \theta A^1 + (1 - \theta)A^3 + (2 + \theta)A^4, \end{aligned}$$

and

$$\begin{aligned} b &= A^3 + 2A^4 - \theta A^2 + \theta A^2 \\ &= A^3 + 2A^4 - \theta(-A^3 + A^4) + \theta A^2 \\ &= \theta A^2 + (1 + \theta)A^3 + (2 - \theta)A^4. \end{aligned}$$

In the first case, the point $(\theta, 0, 1 - \theta, 2 + \theta)$ is a feasible solution iff $0 \leq \theta \leq 1$ and the value of the objective function is θ, and in the second case, the point $(0, \theta, 1 + \theta, 2 - \theta)$ is a feasible solution iff $0 \leq \theta \leq 2$ and the value of the objective function is 0. In order to increase the objective function we must choose the first case, and we pick $\theta = 1$. We get the feasible solution $u_1 = (1, 0, 0, 3)$ corresponding to the basis (A^1, A^4), so it is a basic feasible solution, and the value of the objective function is 1.

The vectors A^2 and A^3 are given in terms of the basis (A^1, A^4) by

$$\begin{aligned} A^2 &= -A^1 \\ A^3 &= A^1 + A^4. \end{aligned}$$

Repeating the process with $u_1 = (1, 0, 0, 3)$, we get

$$\begin{aligned} b &= A^1 + 3A^4 - \theta A^2 + \theta A^2 \\ &= A^1 + 3A^4 - \theta(-A^1) + \theta A^2 \\ &= (1 + \theta)A^1 + \theta A^2 + 3A^4, \end{aligned}$$

and

$$\begin{aligned} b &= A^1 + 3A^4 - \theta A^3 + \theta A^3 \\ &= A^1 + 3A^4 - \theta(A^1 + A^4) + \theta A^3 \\ &= (1 - \theta)A^1 + \theta A^3 + (3 - \theta)A^4. \end{aligned}$$

In the first case, the point $(1 + \theta, \theta, 0, 3)$ is a feasible solution for all $\theta \geq 0$ and the value of the objective function if $1 + \theta$, and in the second case, the point $(1 - \theta, 0, \theta, 3 - \theta)$ is a

feasible solution iff $0 \leq \theta \leq 1$ and the value of the objective function is $1 - \theta$. This time, we are in the situation where the points

$$(1 + \theta, \theta, 0, 3) = (1, 0, 0, 3) + \theta(1, 1, 0, 0), \quad \theta \geq 0$$

form an infinite ray in the set of feasible solutions, and the objective function $1 + \theta$ is unbounded from above on this ray. This indicates that our linear program, although feasible, is unbounded.

Let us now describe a step of the simplex algorithm in general.

10.2 The Simplex Algorithm in General

We assume that we already have an initial vertex u_0 to start from. This vertex corresponds to a basic feasible solution with basis K_0. We will show later that it is always possible to find a basic feasible solution of a Linear Program (P) is standard form, or to detect that (P) has no feasible solution.

The idea behind the simplex algorithm is this: Given a pair (u, K) consisting of a basic feasible solution u and a basis K for u, find another pair (u^+, K^+) consisting of another basic feasible solution u^+ and a basis K^+ for u^+, such that K^+ is obtained from K by deleting some basic index $k^- \in K$ and adding some nonbasic index $j^+ \notin K$, in such a way that the value of the objective function increases (preferably strictly). The step which consists in swapping the vectors A^{k^-} and A^{j^+} is called a *pivoting step*.

Let u be a given vertex corresponds to a basic feasible solution with basis K. Since the m vectors A^k corresponding to indices $k \in K$ are linearly independent, they form a basis, so for every nonbasic $j \notin K$, we write

$$A^j = \sum_{k \in K} \gamma_k^j A^k. \tag{$*$}$$

We let $\gamma_K^j \in \mathbb{R}^m$ be the vector given by $\gamma_K^j = (\gamma_k^j)_{k \in K}$. Actually, since the vector γ_K^j depends on K, to be very precise we should denote its components by $(\gamma_K^j)_k$, but to simplify notation we usually write γ_k^j instead of $(\gamma_K^j)_k$ (unless confusion arises). We will explain later how the coefficients γ_k^j can be computed efficiently.

Since u is a feasible solution we have $u \geq 0$ and $Au = b$, that is,

$$\sum_{k \in K} u_k A^k = b. \tag{$**$}$$

For every nonbasic $j \notin K$, a candidate for entering the basis K, we try to find a new vertex $u(\theta)$ that improves the objective function, and for this we add $-\theta A^j + \theta A^j = 0$ to b in

Equation $(**)$ and then replace the occurrence of A^j in $-\theta A^j$ by the right hand side of Equation $(*)$ to obtain

$$\begin{aligned} b &= \sum_{k\in K} u_k A^k - \theta A^j + \theta A^j \\ &= \sum_{k\in K} u_k A^k - \theta\Big(\sum_{k\in K} \gamma_k^j A^k\Big) + \theta A^j \\ &= \sum_{k\in K} \Big(u_k - \theta\gamma_k^j\Big) A^k + \theta A^j. \end{aligned}$$

Consequently, the vector $u(\theta)$ appearing on the right-hand side of the above equation given by

$$u(\theta)_i = \begin{cases} u_i - \theta\gamma_i^j & \text{if } i \in K \\ \theta & \text{if } i = j \\ 0 & \text{if } i \notin K \cup \{j\} \end{cases}$$

automatically satisfies the constraints $Au(\theta) = b$, and this vector is a feasible solution iff

$$\theta \geq 0 \quad \text{and} \quad u_k \geq \theta\gamma_k^j \quad \text{for all } k \in K.$$

Obviously $\theta = 0$ is a solution, and if

$$\theta^j = \min\left\{ \frac{u_k}{\gamma_k^j} \;\middle|\; \gamma_k^j > 0,\ k \in K \right\} > 0,$$

then we have a range of feasible solutions for $0 \leq \theta \leq \theta^j$. The value of the objective function for $u(\theta)$ is

$$cu(\theta) = \sum_{k\in K} c_k(u_k - \theta\gamma_k^j) + \theta c_j = cu + \theta\Big(c_j - \sum_{k\in K} \gamma_k^j c_k\Big).$$

Since the potential change in the objective function is

$$\theta\Big(c_j - \sum_{k\in K} \gamma_k^j c_k\Big)$$

and $\theta \geq 0$, if $c_j - \sum_{k\in K} \gamma_k^j c_k \leq 0$, then the objective function can't be increased.

However, if $c_{j^+} - \sum_{k\in K} \gamma_k^{j^+} c_k > 0$ for some $j^+ \notin K$, and if $\theta^{j^+} > 0$, then the objective function can be strictly increased by choosing any $\theta > 0$ such that $\theta \leq \theta^{j^+}$, so it is natural to zero at least one coefficient of $u(\theta)$ by picking $\theta = \theta^{j^+}$, which also maximizes the increase of the objective function. In this case (Case below (B2)), we obtain a new feasible solution $u^+ = u(\theta^{j^+})$.

Now, if $\theta^{j^+} > 0$, then there is some index $k \in K$ such $u_k > 0$, $\gamma_k^{j^+} > 0$, and $\theta^{j^+} = u_k/\gamma_k^{j^+}$, so we can pick such an index k^- for the vector A^{k^-} leaving the basis K. We claim that

$K^+ = (K - \{k^-\}) \cup \{j^+\}$ is a basis. This is because the coefficient $\gamma_{k^-}^{j^+}$ associated with the column A^{k^-} is nonzero (in fact, $\gamma_{k^-}^{j^+} > 0$), so Equation $(*)$, namely

$$A^{j^+} = \gamma_{k^-}^{j^+} A^{k^-} + \sum_{k \in K - \{k^-\}} \gamma_k^{j^+} A^k,$$

yields the equation

$$A^{k^-} = (\gamma_{k^-}^{j^+})^{-1} A^{j^+} - \sum_{k \in K - \{k^-\}} (\gamma_{k^-}^{j^+})^{-1} \gamma_k^{j^+} A^k,$$

and these equations imply that the subspaces spanned by the vectors $(A^k)_{k \in K}$ and the vectors $(A^k)_{k \in K^+}$ are identical. However, K is a basis of dimension m so this subspace has dimension m, and since K^+ also has m elements, it must be a basis. Therefore, $u^+ = u(\theta^{j^+})$ is a basic feasible solution.

The above case is the most common one, but other situations may arise. In what follows, we discuss all eventualities.

Case (A).

We have $c_j - \sum_{k \in K} \gamma_k^j c_k \leq 0$ for all $j \notin K$. Then it turns out that u is an *optimal solution*. Otherwise, we are in Case (B).

Case (B).

We have $c_j - \sum_{k \in K} \gamma_k^j c_k > 0$ for some $j \notin K$ (not necessarily unique). There are three subcases.

Case (B1).

If for some $j \notin K$ as above we also have $\gamma_k^j \leq 0$ for all $k \in K$, since $u_k \geq 0$ for all $k \in K$, this places no restriction on θ, and the objective function is *unbounded above*. This is demonstrated by Example 10.3 with $K = \{3, 4\}$ and $j = 2$ since $\gamma_{\{3,4\}}^2 = (-1, 0)$.

Case (B2).

There is some index $j^+ \notin K$ such that simultaneously

(1) $c_{j^+} - \sum_{k \in K} \gamma_k^{j^+} c_k > 0$, which means that the objective function can potentially be increased;

(2) There is some $k \in K$ such that $\gamma_k^{j^+} > 0$, and for every $k \in K$, if $\gamma_k^{j^+} > 0$ then $u_k > 0$, which implies that $\theta^{j^+} > 0$.

If we pick $\theta = \theta^{j^+}$ where

$$\theta^{j^+} = \min\left\{ \frac{u_k}{\gamma_k^{j^+}} \;\middle|\; \gamma_k^{j^+} > 0,\ k \in K \right\} > 0,$$

then the feasible solution u^+ given by

$$u_i^+ = \begin{cases} u_i - \theta^{j^+}\gamma_i^{j^+} & \text{if } i \in K \\ \theta^{j^+} & \text{if } i = j^+ \\ 0 & \text{if } i \notin K \cup \{j^+\} \end{cases}$$

is a vertex of $\mathcal{P}(A, b)$. If we pick any index $k^- \in K$ such that $\theta^{j^+} = u_{k^-}/\gamma_{k^-}^{j+}$, then $K^+ = (K - \{k^-\}) \cup \{j^+\}$ is a basis for u^+. The vector A^{j^+} enters the new basis K^+, and the vector A^{k^-} leaves the old basis K. This is a *pivoting step*. The objective function increases strictly. This is demonstrated by Example 10.2 with $K = \{3, 4, 5\}$, $j = 1$, and $k = 4$, Then $\gamma^1_{\{3,4,5\}} = (-1, 1, 0)$, with $\gamma_4^1 = 1$. Since $u = (0, 0, 1, 3, 2)$, $\theta^1 = \frac{u_4}{\gamma_4^1} = 3$, and the new optimal solutions becomes $u^+ = (3, 0, 1 - 3(-1), 3 - 3(1), 2 - 3(0)) = (3, 0, 4, 0, 2)$.

Case (B3).

There is some index $j \notin K$ such that $c_j - \sum_{k\in K} \gamma_k^j c_k > 0$, and for each of the indices $j \notin K$ satisfying the above property we have simultaneously

(1) $c_j - \sum_{k\in K} \gamma_k^j c_k > 0$, which means that the objective function can potentially be increased;

(2) There is some $k \in K$ such that $\gamma_k^j > 0$, and $u_k = 0$, which implies that $\theta^j = 0$.

Consequently, the objective function *does not change*. In this case, u is a degenerate basic feasible solution.

We can associate to $u^+ = u$ a new basis K^+ as follows: Pick any index $j^+ \notin K$ such that

$$c_{j^+} - \sum_{k\in K} \gamma_k^{j^+} c_k > 0,$$

and any index $k^- \in K$ such that

$$\gamma_{k^-}^{j^+} > 0,$$

and let $K^+ = (K - \{k^-\}) \cup \{j^+\}$. As in Case (B2), The vector A^{j^+} enters the new basis K^+, and the vector A^{k^-} leaves the old basis K. This is a *pivoting step*. However, the objective function *does not change* since $\theta^{j+} = 0$. This is demonstrated by Example 10.1 with $K = \{3, 4\}$, $j = 2$, and $k = 3$.

It is easy to prove that in Case (A) the basic feasible solution u is an optimal solution, and that in Case (B1) the linear program is unbounded. We already proved that in Case (B2) the vector u^+ and its basis K^+ constitutes a basic feasible solution, and the proof in Case (B3) is similar. For details, see Ciarlet [25] (Chapter 10).

It is convenient to reinterpret the various cases considered by introducing the following sets:

$$B_1 = \Big\{ j \notin K \mid c_j - \sum_{k \in K} \gamma_k^j c_k > 0,\ \max_{k \in K} \gamma_k^j \leq 0 \Big\}$$

$$B_2 = \Big\{ j \notin K \mid c_j - \sum_{k \in K} \gamma_k^j c_k > 0,\ \max_{k \in K} \gamma_k^j > 0,\ \min\Big\{ \frac{u_k}{\gamma_k^j} \;\Big|\; k \in K, \gamma_k^j > 0 \Big\} > 0 \Big\}$$

$$B_3 = \Big\{ j \notin K \mid c_j - \sum_{k \in K} \gamma_k^j c_k > 0,\ \max_{k \in K} \gamma_k^j > 0,\ \min\Big\{ \frac{u_k}{\gamma_k^j} \;\Big|\; k \in K, \gamma_k^j > 0 \Big\} = 0 \Big\},$$

and

$$B = B_1 \cup B_2 \cup B_3 = \Big\{ j \notin K \mid c_j - \sum_{k \in K} \gamma_k^j c_k > 0 \Big\}.$$

Then it is easy to see that the following equivalences hold:

$$\begin{aligned} &\text{Case (A)} \Longleftrightarrow B = \emptyset, \qquad \text{Case (B)} \Longleftrightarrow B \neq \emptyset \\ &\text{Case (B1)} \Longleftrightarrow B_1 \neq \emptyset \\ &\text{Case (B2)} \Longleftrightarrow B_2 \neq \emptyset \\ &\text{Case (B3)} \Longleftrightarrow B_3 \neq \emptyset. \end{aligned}$$

Furthermore, Cases (A) and (B), Cases (B1) and (B3), and Cases (B2) and (B3) are mutually exclusive, while Cases (B1) and (B2) are not.

If Case (B1) and Case (B2) arise simultaneously, we opt for Case (B1) which says that the Linear Program (P) is unbounded and terminate the algorithm.

Here are a few remarks about the method.

In Case (B2), which is the path followed by the algorithm most frequently, various choices have to be made for the index $j^+ \notin K$ for which $\theta^{j^+} > 0$ (the new index in K^+). Similarly, various choices have to be made for the index $k^- \in K$ leaving K, but such choices are typically less important.

Similarly in Case (B3), various choices have to be made for the new index $j^+ \notin K$ going into K^+. In Cases (B2) and (B3), criteria for making such choices are called *pivot rules*.

Case (B3) only arises when u is a degenerate vertex. But even if u is degenerate, Case (B2) may arise if $u_k > 0$ whenever $\gamma_k^j > 0$. It may also happen that u is nondegenerate but as a result of Case (B2), the new vertex u^+ is degenerate because at least two components $u_{k_1} - \theta^{j^+}\gamma_{k_1}^{j^+}$ and $u_{k_2} - \theta^{j^+}\gamma_{k_2}^{j^+}$ vanish for some distinct $k_1, k_2 \in K$.

Cases (A) and (B1) correspond to situations where the algorithm terminates, and Case (B2) can only arise a finite number of times during execution of the simplex algorithm, since the objective function is strictly increased from vertex to vertex and there are only finitely many vertices. Therefore, if the simplex algorithm is started on any initial basic feasible solution u_0, then one of three mutually exclusive situations may arise:

(1) There is a finite sequence of occurrences of Case (B2) and/or Case (B3) ending with an occurrence of Case (A). Then the last vertex produced by the algorithm is an optimal solution. This is what occurred in Examples 10.1 and 10.2.

(2) There is a finite sequence of occurrences of Case (B2) and/or Case (B3) ending with an occurrence of Case (B1). We conclude that the problem is unbounded, and thus has no solution. This is what occurred in Example 10.3.

(3) There is a finite sequence of occurrences of Case (B2) and/or Case (B3), followed by an infinite sequence of Case (B3). If this occurs, the algorithm visits the some basis twice. This a phenomenon known as *cycling*. In this eventually the algorithm fails to come to a conclusion.

There are examples for which cycling occur, although this is rare in practice. Such an example is given in Chvatal [24]; see Chapter 3, pages 31-32, for an example with seven variables and three equations that cycles after six iterations under a certain pivot rule.

The third possibility can be avoided by the choice of a suitable pivot rule. Two of these rules are *Bland's rule* and the *lexicographic rule*; see Chvatal [24] (Chapter 3, pages 34-38).

Bland's rule says: choose the smallest of the eligible incoming indices $j^+ \notin K$, and similarly choose the smallest of the eligible outgoing indices $k^- \in K$.

It can be proven that cycling cannot occur if Bland's rule is chosen as the pivot rule. The proof is very technical; see Chvatal [24] (Chapter 3, pages 37-38), Matousek and Gardner [54] (Chapter 5, Theorem 5.8.1), and Papadimitriou and Steiglitz [60] (Section 2.7). Therefore, assuming that some initial basic feasible solution is provided, and using a suitable pivot rule (such as Bland's rule), the simplex algorithm always terminates and either yields an optimal solution or reports that the linear program is unbounded. Unfortunately, Bland's rules is one of the slowest pivot rules.

The choice of a pivot rule affects greatly the number of pivoting steps that the simplex algorithms goes through. It is not our intention here to explain the various pivot rules. We simply mention the following rules, referring the reader to Matousek and Gardner [54] (Chapter 5, Section 5.7) or to the texts cited in Section 8.1.

1. Largest coefficient, or Dantzig's rule.

2. Largest increase.

3. Steepest edge.

4. Bland's Rule.

5. Random edge.

The steepest edge rule is one of the most popular. The idea is to maximize the ratio

$$\frac{c(u^+ - u)}{\|u^+ - u\|}.$$

The random edge rule picks the index $j^+ \notin K$ of the entering basis vector uniformly at random among all eligible indices.

Let us now return to the issue of the initialization of the simplex algorithm. We use the Linear Program $(\widehat{P})$ introduced during the proof of Theorem 9.7.

Consider a Linear Program $(P2)$

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax = b \text{ and } x \geq 0, \end{aligned}$$

in standard form where A is an $m \times n$ matrix of rank m.

First, observe that since the constraints are equations, we can ensure that $b \geq 0$, because every equation $a_i x = b_i$ where $b_i < 0$ can be replaced by $-a_i x = -b_i$. The next step is to introduce the Linear Program $(\widehat{P})$ in standard form

$$\begin{aligned} &\text{maximize} && -(x_{n+1} + \cdots + x_{n+m}) \\ &\text{subject to} && \widehat{A}\,\widehat{x} = b \text{ and } \widehat{x} \geq 0, \end{aligned}$$

where $\widehat{A}$ and $\widehat{x}$ are given by

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad \widehat{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_{n+m} \end{pmatrix}.$$

Since we assumed that $b \geq 0$, the vector $\widehat{x} = (0_n, b)$ is a feasible solution of $(\widehat{P})$, in fact a basic feasible solutions since the matrix associated with the indices $n+1, \ldots, n+m$ is the identity matrix I_m. Furthermore, since $x_i \geq 0$ for all i, the objective function $-(x_{n+1} + \cdots + x_{n+m})$ is bounded above by 0.

If we execute the simplex algorithm with a pivot rule that prevents cycling, starting with the basic feasible solution $(0_n, d)$, since the objective function is bounded by 0, the simplex algorithm terminates with an optimal solution given by some basic feasible solution, say (u^*, w^*), with $u^* \in \mathbb{R}^n$ and $w^* \in \mathbb{R}^m$.

As in the proof of Theorem 9.7, for every feasible solution $u \in \mathcal{P}(A, b)$, the vector $(u, 0_m)$ is an optimal solution of $(\widehat{P})$. Therefore, if $w^* \neq 0$, then $\mathcal{P}(A, b) = \emptyset$, since otherwise for every feasible solution $u \in \mathcal{P}(A, b)$ the vector $(u, 0_m)$ would yield a value of the objective function $-(x_{n+1} + \cdots + x_{n+m})$ equal to 0, but (u^*, w^*) yields a strictly negative value since $w^* \neq 0$.

Otherwise, $w^* = 0$, and u^* is a feasible solution of $(P2)$. Since $(u^*, 0_m)$ is a basic feasible solution of $(\widehat{P})$ the columns corresponding to nonzero components of u^* are linearly independent. Some of the coordinates of u^* could be equal to 0, but since A has rank m we can add columns of A to obtain a basis K^* associated with u^*, and u^* is indeed a basic feasible solution of $(P2)$.

Running the simplex algorithm on the Linear Program $\widehat{P}$ to obtain an initial feasible solution (u_0, K_0) of the linear program $(P2)$ is called *Phase I* of the simplex algorithm. Running the simplex algorithm on the Linear Program $(P2)$ with some initial feasible solution (u_0, K_0) is called *Phase II* of the simplex algorithm. If a feasible solution of the Linear Program $(P2)$ is readily available then Phase I is skipped. Sometimes, at the end of Phase I, an optimal solution of $(P2)$ is already obtained.

In summary, we proved the following fact worth recording.

Proposition 10.1. *For any Linear Program* $(P2)$

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax = b \text{ and } x \geq 0, \end{aligned}$$

in standard form, where A is an $m \times n$ matrix of rank m and $b \geq 0$, consider the Linear Program $(\widehat{P})$ in standard form

$$\begin{aligned} &\text{maximize} && -(x_{n+1} + \cdots + x_{n+m}) \\ &\text{subject to} && \widehat{A}\widehat{x} = b \text{ and } \widehat{x} \geq 0. \end{aligned}$$

The simplex algorithm with a pivot rule that prevents cycling started on the basic feasible solution $\widehat{x} = (0_n, b)$ of $(\widehat{P})$ terminates with an optimal solution (u^*, w^*).

(1) If $w^* \neq 0$, then $\mathcal{P}(A, b) = \emptyset$, that is, the Linear Program $(P2)$ has no feasible solution.

(2) If $w^* = 0$, then $\mathcal{P}(A, b) \neq \emptyset$, and u^* is a basic feasible solution of $(P2)$ associated with some basis K.

Proposition 10.1 shows that determining whether the polyhedron $\mathcal{P}(A, b)$ defined by a system of equations $Ax = b$ and inequalities $x \geq 0$ is nonempty is decidable. This decision procedure uses a fail-safe version of the simplex algorithm (that prevents cycling), and the proof that it always terminates and returns an answer is nontrivial.

10.3 How to Perform a Pivoting Step Efficiently

We now discuss briefly how to perform the computation of (u^+, K^+) from a basic feasible solution (u, K).

In order to avoid applying permutation matrices it is preferable to allow a basis K to be a sequence of indices, possibly out of order. Thus, for any $m \times n$ matrix A (with $m \leq n$) and any sequence $K = (k_1, k_2, \cdots, k_m)$ of m elements with $k_i \in \{1, \ldots, n\}$, the matrix A_K denotes the $m \times m$ matrix whose ith column is the k_ith column of A, and similarly for any vector $u \in \mathbb{R}^n$ (resp. any linear form $c \in (\mathbb{R}^n)^*$), the vector $u_K \in \mathbb{R}^m$ (the linear form $c_K \in (\mathbb{R}^m)^*$) is the vector whose ith entry is the k_ith entry in u (resp. the linear whose ith entry is the k_ith entry in c).

For each nonbasic $j \notin K$, we have

$$A^j = \gamma^j_{k_1} A^{k_1} + \cdots + \gamma^j_{k_m} A^{k_m} = A_K \gamma^j_K,$$

so the vector γ^j_K is given by $\gamma^j_K = A_K^{-1} A^j$, that is, by solving the system

$$A_K \gamma^j_K = A^j. \tag{$*_\gamma$}$$

To be very precise, since the vector γ^j_K depends on K its components should be denoted by $(\gamma^j_K)_{k_i}$, but as we said before, to simplify notation we write $\gamma^j_{k_i}$ instead of $(\gamma^j_K)_{k_i}$.

In order to decide which case applies ((A), (B1), (B2), (B3)), we need to compute the numbers $c_j - \sum_{k \in K} \gamma^j_k c_k$ for all $j \notin K$. For this, observe that

$$c_j - \sum_{k \in K} \gamma^j_k c_k = c_j - c_K \gamma^j_K = c_j - c_K A_K^{-1} A^j.$$

If we write $\beta_K = c_K A_K^{-1}$, then

$$c_j - \sum_{k \in K} \gamma^j_k c_k = c_j - \beta_K A^j,$$

and we see that $\beta_K^\top \in \mathbb{R}^m$ is the solution of the system $\beta_K^\top = (A_K^{-1})^\top c_k^\top$, which means that $\beta_K^\top$ is the solution of the system

$$A_K^\top \beta_K^\top = c_K^\top. \tag{$*_\beta$}$$

Remark: Observe that since u is a basis feasible solution of (P), we have $u_j = 0$ for all $j \notin K$, so u is the solution of the equation $A_K u_K = b$. As a consequence, the value of the objective function for u is $cu = c_K u_K = c_K A_K^{-1} b$. This fact will play a crucial role in Section 11.2 to show that when the simplex algorithm terminates with an optimal solution of the Linear Program (P), then it also produces an optimal solution of the Dual Linear Program (D).

Assume that we have a basic feasible solution u, a basis K for u, and that we also have the matrix A_K as well its inverse A_K^{-1} (perhaps implicitly) and also the inverse $(A_K^\top)^{-1}$ of $A_K^\top$ (perhaps implicitly). Here is a description of an iteration step of the simplex algorithm, following almost exactly Chvatal (Chvatal [24], Chapter 7, Box 7.1).

An Iteration Step of the (Revised) Simplex Method

Step 1. Compute the numbers $c_j - \sum_{k\in K} \gamma_k^j c_k = c_j - \beta_K A^j$ for all $j \notin K$, and for this, compute $\beta_K^\top$ as the solution of the system

$$A_K^\top \beta_K^\top = c_K^\top.$$

If $c_j - \beta_K A^j \leq 0$ for all $j \notin K$, stop and return the optimal solution u (Case (A)).

Step 2. If Case (B) arises, use a pivot rule to determine which index $j^+ \notin K$ should enter the new basis K^+ (the condition $c_{j^+} - \beta_K A^{j^+} > 0$ should hold).

Step 3. Compute $\max_{k\in K} \gamma_k^{j^+}$. For this, solve the linear system

$$A_K \gamma_K^{j^+} = A^{j^+}.$$

Step 4. If $\max_{k\in K} \gamma_k^{j^+} \leq 0$, then stop and report that Linear Program (P) is unbounded (Case (B1)).

Step 5. If $\max_{k\in K} \gamma_k^{j^+} > 0$, use the ratios $u_k/\gamma_k^{j^+}$ for all $k \in K$ such that $\gamma_k^{j^+} > 0$ to compute θ^{j^+}, and use a pivot rule to determine which index $k^- \in K$ such that $\theta^{j^+} = u_{k^-}/\gamma_{k^-}^{j+}$ should leave K (Case (B2)).

If $\max_{k\in K} \gamma_k^{j^+} = 0$, then use a pivot rule to determine which index k^- for which $\gamma_{k^-}^{j^+} > 0$ should leave the basis K (Case (B3)).

Step 6. Update u, K, and A_K, to u^+ and K^+, and A_{K^+}. During this step, given the basis K specified by the sequence $K = (k_1, \ldots, k_\ell, \ldots, k_m)$, with $k^- = k_\ell$, then K^+ is the sequence obtained by replacing k_ℓ by the incoming index j^+, so $K^+ = (k_1, \ldots, j^+, \ldots, k_m)$ with j^+ in the ℓth slot.

The vector u is easily updated. To compute A_{K^+} from A_K we take advantage of the fact that A_K and A_{K^+} only differ by a *single column*, namely the ℓth column A^{j^+}, which is given by the linear combination

$$A^{j^+} = A_K \gamma_K^{j^+}.$$

To simplify notation, denote $\gamma_K^{j^+}$ by γ, and recall that $k^- = k_\ell$. If $K = (k_1, \ldots, k_m)$, then $A_K = [A^{k_1} \cdots A^{k^-} \cdots A^{i_m}]$, and since A_{K^+} is the result of replacing the ℓth column A^{k^-} of A_K by the column A^{j^+}, we have

$$A_{K^+} = [A^{k_1} \cdots A^{j^+} \cdots A^{i_m}] = [A^{k_1} \cdots A_K \gamma \cdots A^{i_m}] = A_K E(\gamma),$$

where $E(\gamma)$ is the following invertible matrix obtained from the identity matrix I_m by re-

placing its ℓth column by γ:

$$E(\gamma) = \begin{pmatrix} 1 & & \gamma_1 & & \\ & \ddots & \vdots & & \\ & 1 & \gamma_{\ell-1} & & \\ & & \gamma_\ell & & \\ & & \gamma_{\ell+1} & 1 & \\ & & \vdots & & \ddots \\ & & \gamma_m & & 1 \end{pmatrix}.$$

Since $\gamma_\ell = \gamma_{k^-}^{j^+} > 0$, the matrix $E(\gamma)$ is invertible, and it is easy to check that its inverse is given by

$$E(\gamma)^{-1} = \begin{pmatrix} 1 & & -\gamma_\ell^{-1}\gamma_1 & & \\ & \ddots & \vdots & & \\ & 1 & -\gamma_\ell^{-1}\gamma_{\ell-1} & & \\ & & \gamma_\ell^{-1} & & \\ & & -\gamma_\ell^{-1}\gamma_{\ell+1} & 1 & \\ & & \vdots & & \ddots \\ & & -\gamma_\ell^{-1}\gamma_m & & 1 \end{pmatrix},$$

which is very cheap to compute. We also have

$$A_{K^+}^{-1} = E(\gamma)^{-1} A_K^{-1}.$$

Consequently, if A_K and A_K^{-1} are available, then A_{K^+} and $A_{K^+}^{-1}$ can be computed cheaply in terms of A_K and A_K^{-1} and matrices of the form $E(\gamma)$. Then the systems $(*_\gamma)$ to find the vectors γ_K^j can be solved cheaply.

Since

$$A_{K^+}^\top = E(\gamma)^\top A_K^\top$$

and

$$(A_{K^+}^\top)^{-1} = (A_K^\top)^{-1}(E(\gamma)^\top)^{-1},$$

the matrices $A_{K^+}^\top$ and $(A_{K^+}^\top)^{-1}$ can also be computed cheaply from $A_K^\top$, $(A_K^\top)^{-1}$, and matrices of the form $E(\gamma)^\top$. Thus the systems $(*_\beta)$ to find the linear forms β_K can also be solved cheaply.

A matrix of the form $E(\gamma)$ is called an *eta matrix*; see Chvatal [24] (Chapter 7). We showed that the matrix A_{K^s} obtained after s steps of the simplex algorithm can be written as

$$A_{K^s} = A_{K^{s-1}} E_s$$

for some eta matrix E_s, so A_{k^s} can be written as the product

$$A_{K^s} = E_1 E_2 \cdots E_s$$

of s eta matrices. Such a factorization is called an *eta factorization*. The eta factorization can be used to either invert A_{K^s} or to solve a system of the form $A_{K_s}\gamma = A^{j^+}$ iteratively. Which method is more efficient depends on the sparsity of the E_i.

In summary, there are cheap methods for finding the next basic feasible solution (u^+, K^+) from (u, K). We simply wanted to give the reader a flavor of these techniques. We refer the reader to texts on linear programming for detailed presentations of methods for implementing efficiently the simplex method. In particular, the *revised simplex method* is presented in Chvatal [24], Papadimitriou and Steiglitz [60], Bertsimas and Tsitsiklis [14], and Vanderbei [80].

10.4 The Simplex Algorithm Using Tableaux

We now describe a formalism for presenting the simplex algorithm, namely *(full) tableaux*. This is the traditional formalism used in all books, modulo minor variations. A particularly nice feature of the tableau formalism is that the update of a tableau can be performed using elementary row operations *identical* to the operations used during the reduction of a matrix to row reduced echelon form (rref). What differs is the criterion for the choice of the pivot.

Since the quantities $c_j - c_K\gamma_K^j$ play a crucial role in determining which column A^j should come into the basis, the notation $\overline{c}_j$ is used to denote $c_j - c_K\gamma_K^j$, which is called the *reduced cost* of the variable x_j. The reduced costs actually depend on K so to be very precise we should denote them by $(\overline{c}_K)_j$, but to simplify notation we write $\overline{c}_j$ instead of $(\overline{c}_K)_j$. We will see shortly how $(\overline{c}_{K^+})_i$ is computed in terms of $(\overline{c}_K)_i$.

Observe that the data needed to execute the next step of the simplex algorithm are

(1) The current basic solution u_K and its basis $K = (k_1, \ldots, k_m)$.

(2) The reduced costs $\overline{c}_j = c_j - c_K A_K^{-1} A^j = c_j - c_K\gamma_K^j$, for all $j \notin K$.

(3) The vectors $\gamma_K^j = (\gamma_{k_i}^j)_{i=1}^m$ for all $j \notin K$, that allow us to express each A^j as $A_K\gamma_K^j$.

All this information can be packed into a $(m+1) \times (n+1)$ matrix called a *(full) tableau* organized as follows:

$c_K u_K$	$\overline{c}_1$	$\cdots$	$\overline{c}_j$	$\cdots$	$\overline{c}_n$
u_{k_1}	γ_1^1	$\cdots$	γ_1^j	$\cdots$	γ_1^n
$\vdots$	$\vdots$		$\vdots$		$\vdots$
u_{k_m}	γ_m^1	$\cdots$	γ_m^j	$\cdots$	γ_m^n

It is convenient to think as the first row as Row 0, and of the first column as Column 0. Row 0 contains the current value of the objective function and the reduced costs. Column 0, except for its top entry, contains the components of the current basic solution u_K, and

the remaining columns, except for their top entry, contain the vectors γ_K^j. Observe that the γ_K^j corresponding to indices j in K constitute a permutation of the identity matrix I_m. The entry $\gamma_{k^-}^{j^+}$ is called the *pivot* element. A tableau together with the new basis $K^+ = (K - \{k^-\}) \cup \{j^+\}$ contains all the data needed to compute the new u_{K^+}, the new $\gamma_{K^+}^j$, and the new reduced costs $(\overline{c}_{K^+})_j$.

If we define the $m \times n$ matrix Γ as the matrix $\Gamma = [\gamma_K^1 \ \cdots \ \gamma_K^n]$ whose jth column is γ_K^j, and $\overline{c}$ as the row vector $\overline{c} = (\overline{c}_1 \ \cdots \ \overline{c}_n)$, then the above tableau is denoted concisely by

$c_K u_K$	$\overline{c}$
u_K	Γ

We now show that the update of a tableau can be performed using elementary row operations identical to the operations used during the reduction of a matrix to row reduced echelon form (rref).

If $K = (k_1, \ldots, k_m)$, j^+ is the index of the incoming basis vector, $k^- = k_\ell$ is the index of the column leaving the basis, and if $K^+ = (k_1, \ldots, k_{\ell-1}, j^+, k_{\ell+1}, \ldots, k_m)$, since $A_{K^+} = A_K E(\gamma_K^{j^+})$, the new columns $\gamma_{K^+}^j$ are computed in terms of the old columns γ_K^j using $(*_\gamma)$ and the equations

$$\gamma_{K^+}^j = A_{K^+}^{-1} A^j = E(\gamma_K^{j^+})^{-1} A_K^{-1} A^j = E(\gamma_K^{j^+})^{-1} \gamma_K^j.$$

Consequently, the matrix Γ^+ is given in terms of Γ by

$$\Gamma^+ = E(\gamma_K^{j^+})^{-1} \Gamma.$$

But the matrix $E(\gamma_K^{j^+})^{-1}$ is of the form

$$E(\gamma_K^{j^+})^{-1} = \begin{pmatrix} 1 & & -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_1}^{j^+} & & \\ & \ddots & \vdots & & \\ & 1 & -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_{\ell-1}}^{j^+} & & \\ & & (\gamma_{k^-}^{j^+})^{-1} & & \\ & & -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_{\ell+1}}^{j^+} & 1 & \\ & & \vdots & & \ddots \\ & & -(\gamma_{k^-}^{j^+})^{-1}\gamma_{k_m}^{j^+} & & 1 \end{pmatrix},$$

with the column involving the γs in the ℓth column, and Γ^+ is obtained by applying the following elementary row operations to Γ:

1. Multiply Row ℓ by $1/\gamma_{k^-}^{j^+}$ (the inverse of the pivot) to make the entry on Row ℓ and Column j^+ equal to 1.

2. Subtract $\gamma_{k_i}^{j^+} \times$ (the normalized) Row ℓ from Row i, for $i = 1, \ldots, \ell - 1, \ell + 1, \ldots, m$.

These are *exactly* the elementary row operations that reduce the ℓth column $\gamma_K^{j^+}$ of Γ to the ℓth column of the identity matrix I_m. Thus, this step is identical to the sequence of steps that the procedure to convert a matrix to row reduced echelon from executes on the ℓth column of the matrix. The only difference is the criterion for the choice of the pivot.

Since the new basic solution u_{K^+} is given by $u_{K^+} = A_{K^+}^{-1} b$, we have

$$u_{K^+} = E(\gamma_K^{j^+})^{-1} A_K^{-1} b = E(\gamma_K^{j^+})^{-1} u_K.$$

This means that u_+ is obtained from u_K by applying exactly the *same* elementary row operations that were applied to Γ. Consequently, just as in the procedure for reducing a matrix to rref, we can apply elementary row operations to the matrix $[u_k \ \Gamma]$, which consists of rows $1, \ldots, m$ of the tableau.

Once the new matrix Γ^+ is obtained, the new reduced costs are given by the following proposition.

Proposition 10.2. *Given any Linear Program $(P2)$ in standard form*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

where A is an $m \times n$ matrix of rank m, if (u, K) is a basic (not necessarily feasible) solution of $(P2)$ and if $K^+ = (K - \{k^-\}) \cup \{j^+\}$, with $K = (k_1, \ldots, k_m)$ and $k^- = k_\ell$, then for $i = 1, \ldots, n$ we have

$$c_i - c_{K^+}\gamma_{K^+}^i = c_i - c_K \gamma_K^i - \frac{\gamma_{k^-}^i}{\gamma_{k^-}^{j^+}}(c_{j^+} - c_K \gamma_K^{j^+}).$$

Using the reduced cost notation, the above equation is

$$(\overline{c}_{K^+})_i = (\overline{c}_K)_i - \frac{\gamma_{k^-}^i}{\gamma_{k^-}^{j^+}}(\overline{c}_K)_{j^+}.$$

Proof. Without any loss of generality and to simplify notation assume that $K = (1, \ldots, m)$ and write j for j^+ and ℓ for k_m. Since $\gamma_K^i = A_K^{-1} A^i$, $\gamma_{K^+}^i = A_{K^+}^{-1} A^i$, and $A_{K^+} = A_K E(\gamma_K^j)$, we have

$$c_i - c_{K^+}\gamma_{K^+}^i = c_i - c_{K^+} A_{K^+}^{-1} A^i = c_i - c_{K^+} E(\gamma_K^j)^{-1} A_K^{-1} A^i = c_i - c_{K^+} E(\gamma_K^j)^{-1} \gamma_K^i,$$

where

$$E(\gamma_K^j)^{-1} = \begin{pmatrix} 1 & & & -(\gamma_\ell^j)^{-1}\gamma_1^j & & \\ & \ddots & & \vdots & & \\ & & 1 & -(\gamma_\ell^j)^{-1}\gamma_{\ell-1}^j & & \\ & & & (\gamma_\ell^j)^{-1} & & \\ & & & -(\gamma_\ell^j)^{-1}\gamma_{\ell+1}^j & 1 & \\ & & & \vdots & & \ddots \\ & & & -(\gamma_\ell^j)^{-1}\gamma_m^j & & 1 \end{pmatrix}$$

where the ℓth column contains the γs. Since $c_{K^+} = (c_1, \ldots, c_{\ell-1}, c_j, c_{\ell+1}, \ldots, c_m)$, we have

$$c_{K^+} E(\gamma_K^j)^{-1} = \left(c_1, \ldots, c_{\ell-1}, \frac{c_j}{\gamma_\ell^j} - \sum_{k=1, k\neq\ell}^{m} c_k \frac{\gamma_k^j}{\gamma_\ell^j}, c_{\ell+1}, \ldots, c_m\right),$$

and

$$\begin{aligned}
c_{K^+} E(\gamma_K^j)^{-1}\gamma_K^i &= \begin{pmatrix} c_1 & \ldots & c_{\ell-1} & \dfrac{c_j}{\gamma_\ell^j} - \displaystyle\sum_{k=1,k\neq\ell}^{m} c_k \frac{\gamma_k^j}{\gamma_\ell^j} & c_{\ell+1} & \ldots & c_m \end{pmatrix} \begin{pmatrix} \gamma_1^i \\ \vdots \\ \gamma_{\ell-1}^i \\ \gamma_\ell^i \\ \gamma_{\ell+1}^i \\ \vdots \\ \gamma_m^i \end{pmatrix} \\
&= \sum_{k=1,k\neq\ell}^{m} c_k\gamma_k^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}\left(c_j - \sum_{k=1,k\neq\ell}^{m} c_k\gamma_k^j\right) \\
&= \sum_{k=1,k\neq\ell}^{m} c_k\gamma_k^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}\left(c_j + c_\ell\gamma_\ell^j - \sum_{k=1}^{m} c_k\gamma_k^j\right) \\
&= \sum_{k=1}^{m} c_k\gamma_k^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}\left(c_j - \sum_{k=1}^{m} c_k\gamma_k^j\right) \\
&= c_K\gamma_K^i + \frac{\gamma_\ell^i}{\gamma_\ell^j}(c_j - c_K\gamma_K^j),
\end{aligned}$$

and thus

$$c_i - c_{K^+}\gamma_{K^+}^i = c_i - c_{K^+}E(\gamma_K^j)^{-1}\gamma_K^i = c_i - c_K\gamma_K^i - \frac{\gamma_\ell^i}{\gamma_\ell^j}(c_j - c_K\gamma_K^j),$$

as claimed. □

Since $(\gamma_{k^-}^1, \ldots, \gamma_{k^-}^n)$ is the ℓth row of Γ, we see that Proposition 10.2 shows that

$$\overline{c}_{K^+} = \overline{c}_K - \frac{(\overline{c}_K)_{j^+}}{\gamma_{k^-}^{j^+}}\,\Gamma_\ell, \tag{$\dagger$}$$

where Γ_ℓ denotes the ℓ-th row of Γ and $\gamma_{k^-}^{j^+}$ is the pivot. This means that $\overline{c}_{K^+}$ is obtained by the elementary row operations which consist of first normalizing the ℓth row by dividing it by the pivot $\gamma_{k^-}^{j^+}$, and then subtracting $(\overline{c}_K)_{j^+}\times$ the normalized Row ℓ from $\overline{c}_K$. *These are exactly the row operations that make the reduced cost $(\overline{c}_K)_{j^+}$ zero.*

Remark: It easy easy to show that we also have

$$\overline{c}_{K^+} = c - c_{K^+}\Gamma^+.$$

We saw in Section 10.2 that the change in the objective function after a pivoting step during which column j^+ comes in and column k^- leaves is given by

$$\theta^{j^+}\left(c_{j^+} - \sum_{k\in K}\gamma_k^{j^+}c_k\right) = \theta^{j^+}(\overline{c}_K)_{j^+},$$

where

$$\theta^{j^+} = \frac{u_{k^-}}{\gamma_{k^-}^{j^+}}.$$

If we denote the value of the objective function $c_K u_K$ by z_K, then we see that

$$z_{K^+} = z_K + \frac{(\overline{c}_K)_{j^+}}{\gamma_{k^-}^{j^+}}\, u_{k^-}.$$

This means that the new value z_{K^+} of the objective function is obtained by first normalizing the ℓth row by dividing it by the pivot $\gamma_{k^-}^{j^+}$, and then adding $(\overline{c}_K)_{j^+}\times$ the zeroth entry of the normalized ℓth line by $(\overline{c}_K)_{j^+}$ to the zeroth entry of line 0.

In updating the reduced costs, we subtract rather than add $(\overline{c}_K)_{j^+}\times$ the normalized row ℓ from row 0. This suggests storing $-z_K$ as the zeroth entry on line 0 rather than z_K, because then all the entries row 0 are updated by the *same* elementary row operations. Therefore, from now on, we use tableau of the form

$-c_K u_K$	$\overline{c}_1$	$\cdots$	$\overline{c}_j$	$\cdots$	$\overline{c}_n$
u_{k_1}	γ_1^1	$\cdots$	γ_1^j	$\cdots$	γ_1^n
$\vdots$	$\vdots$		$\vdots$		$\vdots$
u_{k_m}	γ_m^1	$\cdots$	γ_m^j	$\cdots$	γ_m^n

The simplex algorithm first chooses the incoming column j^+ by picking some column for which $\overline{c}_j > 0$, and then chooses the outgoing column k^- by considering the ratios $u_k/\gamma_k^{j^+}$ for which $\gamma_k^{j^+} > 0$ (along column j^+), and picking k^- to achieve the minimum of these ratios.

Here is an illustration of the simplex algorithm using elementary row operations on an example from Papadimitriou and Steiglitz [60] (Section 2.9).

Example 10.4. Consider the linear program

$$\begin{aligned}
&\text{maximize} && -2x_2 - x_4 - 5x_7 \\
&\text{subject to} && \\
& && x_1 + x_2 + x_3 + x_4 = 4 \\
& && x_1 + x_5 = 2 \\
& && x_3 + x_6 = 3 \\
& && 3x_2 + x_3 + x_7 = 6 \\
& && x_1, x_2, x_3, x_4, x_5, x_6, x_7 \geq 0.
\end{aligned}$$

We have the basic feasible solution $u = (0, 0, 0, 4, 2, 3, 6)$, with $K = (4, 5, 6, 7)$. Since $c_K = (-1, 0, 0, -5)$ and $c = (0, -2, 0, -1, 0, 0 - 5)$ the first tableau is

34	1	14	6	0	0	0	0
$u_4 = 4$	1	1	1	1	0	0	0
$u_5 = 2$	(1)	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 6$	0	3	1	0	0	0	1

Since $\bar{c}_j = c_j - c_K \gamma_K^j$, Row 0 is obtained by subtracting $-1\times$ Row 1 and $-5\times$ Row 4 from $c = (0, -2, 0, -1, 0, 0, -5)$. Let us pick Column $j^+ = 1$ as the incoming column. We have the ratios (for positive entries on Column 1)

$$4/1,\ 2/1,$$

and since the minimum is 2, we pick the outgoing column to be Column $k^- = 5$. The pivot 1 is indicated in red. The new basis is $K = (4, 1, 6, 7)$. Next we apply row operations to reduce Column 1 to the second vector of the identity matrix I_4. For this, we subtract Row 2 from Row 1. We get the tableau

34	1	14	6	0	0	0	0
$u_4 = 2$	0	1	1	1	-1	0	0
$u_1 = 2$	(1)	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 6$	0	3	1	0	0	0	1

To compute the new reduced costs, we want to set $\bar{c}_1$ to 0, so we apply the identical row operations and subtract Row 2 from Row 0 to obtain the tableau

32	0	14	6	0	-1	0	0
$u_4 = 2$	0	1	(1)	1	-1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 6$	0	3	1	0	0	0	1

Next, pick Column $j^+ = 3$ as the incoming column. We have the ratios (for positive entries on Column 3)

$$2/1,\ 3/1,\ 6/1,$$

and since the minimum is 2, we pick the outgoing column to be Column $k^- = 4$. The pivot 1 is indicated in red and the new basis is $K = (3, 1, 6, 7)$. Next we apply row operations to reduce Column 3 to the first vector of the identity matrix I_4. For this, we subtract Row 1 from Row 3 and from Row 4 and obtain the tableau:

32	0	14	6	0	-1	0	0
$u_3 = 2$	0	1	(1)	1	-1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 1$	0	-1	0	-1	1	1	0
$u_7 = 4$	0	2	0	-1	1	0	1

To compute the new reduced costs, we want to set $\overline{c}_3$ to 0, so we subtract $6\times$ Row 1 from Row 0 to get the tableau

20	0	8	0	-6	5	0	0
$u_3 = 2$	0	(1)	1	1	-1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 1$	0	-1	0	-1	1	1	0
$u_7 = 4$	0	2	0	-1	1	0	1

Next we pick $j^+ = 2$ as the incoming column. We have the ratios (for positive entries on Column 2)

$$2/1,\ 4/2,$$

and since the minimum is 2, we pick the outgoing column to be Column $k^- = 3$. The pivot 1 is indicated in red and the new basis is $K = (2, 1, 6, 7)$. Next we apply row operations to reduce Column 2 to the first vector of the identity matrix I_4. For this, we add Row 1 to Row 3 and subtract $2\times$ Row 1 from Row 4 to obtain the tableau:

20	0	8	0	-6	5	0	0
$u_2 = 2$	0	(1)	1	1	-1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 0$	0	0	-2	-3	3	0	1

To compute the new reduced costs, we want to set $\overline{c}_2$ to 0, so we subtract $8\times$ Row 1 from Row 0 to get the tableau

4	0	0	−8	−14	13	0	0
$u_2 = 2$	0	1	1	1	−1	0	0
$u_1 = 2$	1	0	0	0	1	0	0
$u_6 = 3$	0	0	1	0	0	1	0
$u_7 = 0$	0	0	−2	−3	3	0	1

The only possible incoming column corresponds to $j^+ = 5$. We have the ratios (for positive entries on Column 5)

$$2/1,\ 0/3,$$

and since the minimum is 0, we pick the outgoing column to be Column $k^- = 7$. The pivot 3 is indicated in red and the new basis is $K = (2, 1, 6, 5)$. Since the minimum is 0, the basis $K = (2, 1, 6, 5)$ is degenerate (indeed, the component corresponding to the index 5 is 0). Next we apply row operations to reduce Column 5 to the fourth vector of the identity matrix I_4. For this, we multiply Row 4 by 1/3, and then add the normalized Row 4 to Row 1 and subtract the normalized Row 4 from Row 2 to obtain the tableau:

4	0	0	−8	−14	13	0	0
$u_2 = 2$	0	1	1/3	0	0	0	1/3
$u_1 = 2$	1	0	2/3	1	0	0	−1/3
$u_6 = 3$	0	0	1	0	0	1	0
$u_5 = 0$	0	0	−2/3	−1	1	0	1/3

To compute the new reduced costs, we want to set $\overline{c}_5$ to 0, so we subtract 13× Row 4 from Row 0 to get the tableau

4	0	0	2/3	−1	0	0	−13/3
$u_2 = 2$	0	1	1/3	0	0	0	1/3
$u_1 = 2$	1	0	2/3	1	0	0	−1/3
$u_6 = 3$	0	0	1	0	0	1	0
$u_5 = 0$	0	0	−2/3	−1	1	0	1/3

The only possible incoming column corresponds to $j^+ = 3$. We have the ratios (for positive entries on Column 3)

$$2/(1/3) = 6,\ 2/(2/3) = 3,\ 3/1 = 3,$$

and since the minimum is 3, we pick the outgoing column to be Column $k^- = 1$. The pivot 2/3 is indicated in red and the new basis is $K = (2, 3, 6, 5)$. Next we apply row operations to reduce Column 3 to the second vector of the identity matrix I_4. For this, we multiply Row 2 by 3/2, subtract (1/3)× (normalized Row 2) from Row 1, and subtract normalized Row 2 from Row 3, and add (2/3)× (normalized Row 2) to Row 4 to obtain the tableau:

4	0	0	2/3	−1	0	0	−13/3
$u_2 = 1$	−1/2	1	0	−1/2	0	0	1/2
$u_3 = 3$	3/2	0	(1)	3/2	0	0	−1/2
$u_6 = 0$	−3/2	0	0	−3/2	0	1	1/2
$u_5 = 2$	1	0	0	0	1	0	0

To compute the new reduced costs, we want to set $\bar{c}_3$ to 0, so we subtract $(2/3)\times$ Row 2 from Row 0 to get the tableau

2	−1	0	0	−2	0	0	−4
$u_2 = 1$	−1/2	1	0	−1/2	0	0	1/2
$u_3 = 3$	3/2	0	1	3/2	0	0	−1/2
$u_6 = 0$	−3/2	0	0	−3/2	0	1	1/2
$u_5 = 2$	1	0	0	0	1	0	0

Since all the reduced cost are ≤ 0, we have reached an optimal solution, namely $(0, 1, 3, 0, 2, 0, 0, 0)$, with optimal value -2.

The progression of the simplex algorithm from one basic feasible solution to another corresponds to the visit of vertices of the polyhedron $\mathcal{P}$ associated with the constraints of the linear program illustrated in Figure 10.4.

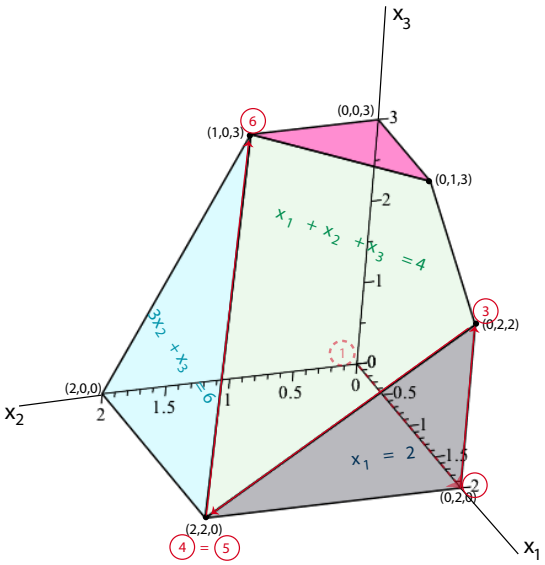


Figure 10.4: The polytope $\mathcal{P}$ associated with the linear program optimized by the tableau method. The red arrowed path traces the progression of the simplex method from the origin to the vertex $(0, 1, 3)$.

As a final comment, if it is necessary to run Phase I of the simplex algorithm, in the event that the simplex algorithm terminates with an optimal solution $(u^*, 0_m)$ and a basis K^* such that some $u_i = 0$, then the basis K^* contains indices of basic columns A^j corresponding to slack variables that need to be *driven out* of the basis. This is easy to achieve by performing a pivoting step involving some other column j^+ corresponding to one of the original variables (not a slack variable) for which $(\gamma_{K^*})_i^{j^+} \neq 0$. In such a step, it doesn't matter whether $(\gamma_{K^*})_i^{j^+} < 0$ or $(\overline{c}_{K^*})_{j^+} \leq 0$. If the original matrix A has no redundant equations, such a step is always possible. Otherwise, $(\gamma_{K^*})_i^j = 0$ for all non-slack variables, so we detected that the ith equation is redundant and we can delete it.

Other presentations of the tableau method can be found in Bertsimas and Tsitsiklis [14] and Papadimitriou and Steiglitz [60].

10.5 Computational Efficiency of the Simplex Method

Let us conclude with a few comments about the efficiency of the simplex algorithm. In *practice*, it was observed by Dantzig that for linear programs with $m < 50$ and $m + n < 200$, the simplex algorithms typically requires less than $3m/2$ iterations, but at most $3m$ iterations. This fact agrees with more recent empirical experiments with much larger programs that show that the number iterations is bounded by $3m$. Thus, it was somewhat of a shock in 1972 when Klee and Minty found a linear program with n variables and n equations for which the simplex algorithm with Dantzig's pivot rule requires requires $2^n - 1$ iterations. This program (taken from Chvatal [24], page 47) is reproduced below:

$$\begin{aligned}
&\text{maximize} \quad \sum_{j=1}^{n} 10^{n-j} x_j \\
&\text{subject to} \\
&\left(2 \sum_{j=1}^{i-1} 10^{i-j} x_j\right) + x_i \leq 100^{i-1} \\
&x_j \geq 0,
\end{aligned}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, n$.

If $p = \max(m, n)$, then, in terms of worse case behavior, for all currently known pivot rules, the simplex algorithm has exponential complexity in p. However, as we said earlier, in practice, nasty examples such as the Klee–Minty example seem to be rare, and the number of iterations appears to be linear in m.

Whether or not a pivot rule (a clairvoyant rule) for which the simplex algorithms runs in polynomial time in terms of m is still an *open problem*.

The *Hirsch conjecture* claims that there is some pivot rule such that the simplex algorithm finds an optimal solution in $O(p)$ steps. The best bound known so far due to Kalai and Kleitman is $m^{1+\ln n} = (2n)^{\ln m}$. For more on this topic, see Matousek and Gardner [54] (Section 5.9) and Bertsimas and Tsitsiklis [14] (Section 3.7).

Researchers have investigated the problem of finding upper bounds on the expected number of pivoting steps if a randomized pivot rule is used. Bounds better than 2^m (but of course, not polynomial) have been found.

Understanding the complexity of linear programing, in particular of the simplex algorithm, is still ongoing. The interested reader is referred to Matousek and Gardner [54] (Chapter 5, Section 5.9) for some pointers.

In the next section we consider important theoretical criteria for determining whether a set of constraints $Ax \leq b$ and $x \geq 0$ has a solution or not.

10.6 Summary

The main concepts and results of this chapter are listed below:

- Degenerate and nondegenerate basic feasible solution.
- Pivoting step.
- Pivot rule.
- Cycling.
- Bland's rule, Dantzig's rule, steepest edge rule, random edge rule, largest increase rule, lexicographic rule.
- Phase I and Phase II of the simplex algorithm.
- eta matrix, eta factorization.
- Revised simplex method.
- Reduced cost.
- Full tableaux.
- The Hirsch conjecture.

10.7 Problems

Problem 10.1. In Section 10.2 prove that if Case (A) arises, then the basic feasible solution u is an optimal solution. Prove that if Case (B1) arises, then the linear program is unbounded. Prove that if Case (B3) arises, then (u^+, K^+) is a basic feasible solution.

Problem 10.2. In Section 10.2 prove that the following equivalences hold:

$$\begin{aligned}
&\text{Case (A)} \Longleftrightarrow B = \emptyset, \qquad \text{Case (B)} \Longleftrightarrow B \neq \emptyset\\
&\text{Case (B1)} \Longleftrightarrow B_1 \neq \emptyset\\
&\text{Case (B2)} \Longleftrightarrow B_2 \neq \emptyset\\
&\text{Case (B3)} \Longleftrightarrow B_3 \neq \emptyset.
\end{aligned}$$

Furthermore, prove that Cases (A) and (B), Cases (B1) and (B3), and Cases (B2) and (B3) are mutually exclusive, while Cases (B1) and (B2) are not.

Problem 10.3. Consider the linear program (due to E.M.L. Beale):

$$\begin{aligned}
\text{maximize} \quad & (3/4)x_1 - 150x_2 + (1/50)x_3 - 6x_4\\
\text{subject to} \quad & \\
& (1/4)x_1 - 60x_2 - (1/25)x_3 + 9x_4 \leq 0\\
& (1/4)x_1 - 90x_2 - (1/50)x_3 + 3x_4 \leq 0\\
& \qquad\qquad\qquad\qquad\qquad\quad x_3 \leq 1\\
& x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0.
\end{aligned}$$

(1) Convert the above program to standard form.

(2) Show that if we apply the simplex algorithm with the pivot rule which selects the column entering the basis as the column of smallest index, then the method cycles.

Problem 10.4. Read carefully the proof given by Chvatal that the lexicographic pivot rule and Bland's pivot rule prevent cycling; see Chvatal [24] (Chapter 3, pages 34-38).

Problem 10.5. Solve the following linear program (from Chvatal [24], Chapter 3, page 44) using the two-phase simplex algorithm:

$$\begin{aligned}
\text{maximize} \quad & 3x_1 + x_2\\
\text{subject to} \quad & \\
& x_1 - x_2 \leq -1\\
& -x_1 - x_2 \leq -3\\
& 2x_1 + x_2 \leq 4\\
& x_1 \geq 0,\ x_2 \geq 0.
\end{aligned}$$

Problem 10.6. Solve the following linear program (from Chvatal [24], Chapter 3, page 44) using the two-phase simplex algorithm:

$$\begin{aligned} &\text{maximize} && 3x_1 + x_2 \\ &\text{subject to} && \\ & && x_1 - x_2 \leq -1 \\ & && -x_1 - x_2 \leq -3 \\ & && 2x_1 + x_2 \leq 2 \\ & && x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

Problem 10.7. Solve the following linear program (from Chvatal [24], Chapter 3, page 44) using the two-phase simplex algorithm:

$$\begin{aligned} &\text{maximize} && 3x_1 + x_2 \\ &\text{subject to} && \\ & && x_1 - x_2 \leq -1 \\ & && -x_1 - x_2 \leq -3 \\ & && 2x_1 - x_2 \leq 2 \\ & && x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

Problem 10.8. Show that the following linear program (from Chvatal [24], Chapter 3, page 43) is unbounded.

$$\begin{aligned} &\text{maximize} && x_1 + 3x_2 - x_3 \\ &\text{subject to} && \\ & && 2x_1 + 2x_2 - x_3 \leq 10 \\ & && 3x_1 - 2x_2 + x_3 \leq 10 \\ & && x_1 - 3x_2 + x_3 \leq 10 \\ & && x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0. \end{aligned}$$

Hint. Try $x_1 = 0, x_3 = t$, and a suitable value for x_2.

Chapter 11

Linear Programming and Duality

11.1 Variants of the Farkas Lemma

If A is an $m \times n$ matrix and if $b \in \mathbb{R}^m$ is a vector, it is known from linear algebra that the linear system $Ax = b$ has no solution iff there is some linear form $y \in (\mathbb{R}^m)^*$ such that $yA = 0$ and $yb \neq 0$. This means that the linear from y vanishes on the columns $A^1, \ldots, A^n$ of A but does not vanish on b. Since the linear form y defines the linear hyperplane H of equation $yz = 0$ (with $z \in \mathbb{R}^m$), geometrically the equation $Ax = b$ has no solution iff there is a linear hyperplane H containing $A^1, \ldots, A^n$ and not containing b. This is a kind of separation theorem that says that the vectors $A^1, \ldots, A^n$ and b can be separated by some linear hyperplane H.

What we would like to do is to generalize this kind of criterion, first to a system $Ax = b$ subject to the constraints $x \geq 0$, and next to sets of inequality constraints $Ax \leq b$ and $x \geq 0$. There are indeed such criteria going under the name of *Farkas lemma*.

The key is a separation result involving polyhedral cones known as the Farkas–Minkowski proposition. We have the following fundamental separation lemma.

Proposition 11.1. *Let $C \subseteq \mathbb{R}^n$ be a closed nonempty (convex) cone. For any point $a \in \mathbb{R}^n$, if $a \notin C$, then there is a linear hyperplane H (through 0) such that*

1. *C lies in one of the two half-spaces determined by H.*
2. *$a \notin H$*
3. *a lies in the other half-space determined by H.*

We say that H strictly separates C and a.

Proposition 11.1, which is illustrated in Figure 11.1, is an easy consequence of another separation theorem that asserts that given any two nonempty closed convex sets A and B of $\mathbb{R}^n$ with A compact, there is a hyperplane H strictly separating A and B (which means

that $A \cap H = \emptyset$, $B \cap H = \emptyset$, that A lies in one of the two half-spaces determined by H, and B lies in the other half-space determined by H); see Gallier [34] (Chapter 7, Corollary 7.4 and Proposition 7.3). This proof is nontrivial and involves a geometric version of the Hahn–Banach theorem.

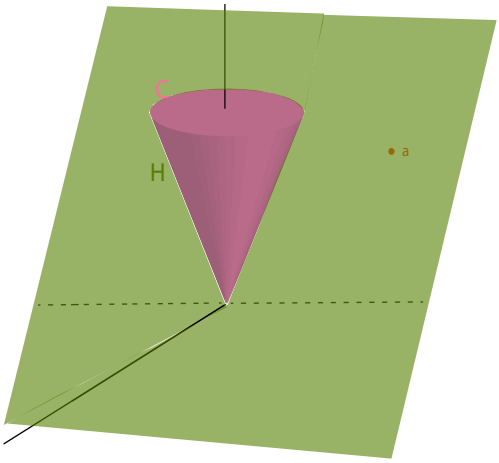


Figure 11.1: In $\mathbb{R}^3$, the olive green hyperplane H separates the cone C from the orange point a.

The Farkas–Minkowski proposition is Proposition 11.1 applied to a polyhedral cone

$$C = \{\lambda_1 a_1 + \cdots + \lambda_n a_n \mid \lambda_i \geq 0,\ i = 1, \ldots, n\}$$

where $\{a_1, \ldots, a_n\}$ is a *finite* number of vectors $a_i \in \mathbb{R}^n$. By Proposition 8.2, any polyhedral cone is closed, so Proposition 11.1 applies and we obtain the following separation lemma.

Proposition 11.2. *(Farkas–Minkowski) Let $C \subseteq \mathbb{R}^n$ be a nonempty polyhedral cone $C = \text{cone}(\{a_1, \ldots, a_n\})$. For any point $b \in \mathbb{R}^n$, if $b \notin C$, then there is a linear hyperplane H (through 0) such that*

1. *C lies in one of the two half-spaces determined by H.*
2. *$b \notin H$*
3. *b lies in the other half-space determined by H.*

Equivalently, there is a nonzero linear form $y \in (\mathbb{R}^n)^*$ such that

1. *$ya_i \geq 0$ for $i = 1, \ldots, n$.*
2. *$yb < 0$.*

A direct proof of the Farkas–Minkowski proposition not involving Proposition 11.1 is given at the end of this section.

Remark: There is a generalization of the Farkas–Minkowski proposition applying to infinite dimensional real Hilbert spaces; see Theorem 12.12 (or Ciarlet [25], Chapter 9).

Proposition 11.2 implies our first version of Farkas' lemma.

Proposition 11.3. *(Farkas Lemma, Version I) Let A be an $m \times n$ matrix and let $b \in \mathbb{R}^m$ be any vector. The linear system $Ax = b$ has no solution $x \geq 0$ iff there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $yA \geq 0_n^\top$ and $yb < 0$.*

Proof. First assume that there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $yA \geq 0$ and $yb < 0$. If $x \geq 0$ is a solution of $Ax = b$, then we get

$$yAx = yb,$$

but if $yA \geq 0$ and $x \geq 0$, then $yAx \geq 0$, and yet by hypothesis $yb < 0$, a contradiction.

Next assume that $Ax = b$ has no solution $x \geq 0$. This means that b does not belong to the polyhedral cone $C = \text{cone}(\{A^1, \ldots, A^n\})$ spanned by the columns of A. By Proposition 11.2, there is a nonzero linear form $y \in (\mathbb{R}^m)^*$ such that

1. $yA^j \geq 0$ for $j = 1, \ldots, n$.
2. $yb < 0$,

which says that $yA \geq 0_n^\top$ and $yb < 0$. □

Next consider the solvability of a system of inequalities of the form $Ax \leq b$ and $x \geq 0$.

Proposition 11.4. *(Farkas Lemma, Version II) Let A be an $m \times n$ matrix and let $b \in \mathbb{R}^m$ be any vector. The system of inequalities $Ax \leq b$ has no solution $x \geq 0$ iff there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $y \geq 0_m^\top$, $yA \geq 0_n^\top$ and $yb < 0$.*

Proof. We use the trick of linear programming which consists of adding "slack variables" z_i to convert inequalities $a_i x \leq b_i$ into equations $a_i x + z_i = b_i$ with $z_i \geq 0$ already discussed just before Definition 8.9. If we let $z = (z_1, \ldots, z_m)$, it is obvious that the system $Ax \leq b$ has a solution $x \geq 0$ iff the equation

$$\begin{pmatrix} A & I_m \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix} = b$$

has a solution $\begin{pmatrix} x \\ z \end{pmatrix}$ with $x \geq 0$ and $z \geq 0$. Now by Farkas I, the above system has no solution with with $x \geq 0$ and $z \geq 0$ iff there is some nonzero linear form $y \in (\mathbb{R}^m)^*$ such that

$$y \begin{pmatrix} A & I_m \end{pmatrix} \geq 0_{n+m}^\top$$

and $yb < 0$, that is, $yA \geq 0_n^\top$, $y \geq 0_m^\top$ and $yb < 0$. □

In the next section we use Farkas II to prove the duality theorem in linear programming. Observe that by taking the negation of the equivalence in Farkas II we obtain a criterion of solvability, namely:

The system of inequalities $Ax \leq b$ has a solution $x \geq 0$ iff for every nonzero linear form $y \in (\mathbb{R}^m)^*$ such that $y \geq 0_m^\top$, if $yA \geq 0_n^\top$, then $yb \geq 0$.

We now prove the Farkas–Minkowski proposition without using Proposition 11.1. This approach uses a basic property of the distance function from a point to a closed set.

Definition 11.1. Let $X \subseteq \mathbb{R}^n$ be any nonempty set and let $a \in \mathbb{R}^n$ be any point. The *distance* $d(a, X)$ from a to X is defined as

$$d(a, X) = \inf_{x \in X} \|a - x\|.$$

Here, $\| \|$ denotes the Euclidean norm.

Proposition 11.5. *Let $X \subseteq \mathbb{R}^n$ be any nonempty set and let $a \in \mathbb{R}^n$ be any point. If X is closed, then there is some $z \in X$ such that $\|a - z\| = d(a, X)$.*

Proof. Since X is nonempty, pick any $x_0 \in X$, and let $r = \|a - x_0\|$. If $B_r(a)$ is the closed ball $B_r(a) = \{x \in \mathbb{R}^n \mid \|x - a\| \leq r\}$, then clearly

$$d(a, X) = \inf_{x \in X} \|a - x\| = \inf_{x \in X \cap B_r(a)} \|a - x\|.$$

Since $B_r(a)$ is compact and X is closed, $K = X \cap B_r(a)$ is also compact. But the function $x \mapsto \|a - x\|$ defined on the compact set K is continuous, and the image of a compact set by a continuous function is compact, so by Heine–Borel it has a minimum that is achieved by some $z \in K \subseteq X$. □

Remark: If U is a nonempty, closed and convex subset of a Hilbert space V, a standard result of Hilbert space theory (the projection lemma, see Proposition 12.5) asserts that for any $v \in V$ there is a *unique* $p \in U$ such that

$$\|v - p\| = \inf_{u \in U} \|v - u\| = d(v, U),$$

and

$$\langle p - v, u - p \rangle \geq 0 \quad \text{for all } u \in U.$$

Here $\|w\| = \sqrt{\langle w, w \rangle}$, where $\langle -, - \rangle$ is the inner product of the Hilbert space V.

We can now give a proof of the Farkas–Minkowski proposition (Proposition 11.2) that does not use Proposition 11.1. This proof is adapted from Matousek and Gardner [54] (Chapter 6, Sections 6.5).

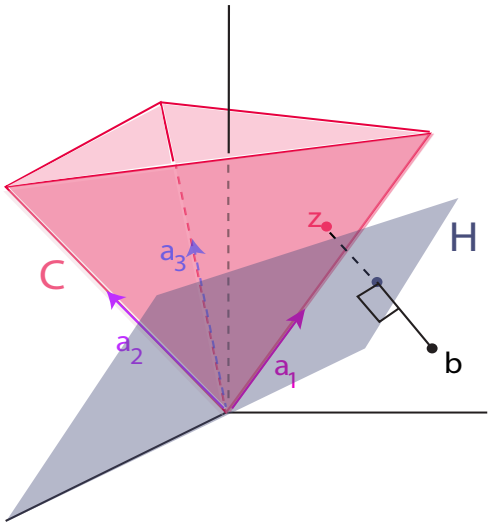


Figure 11.2: The hyperplane H, perpendicular to $z - b$, separates the point b from $C = \text{cone}(\{a_1, a_2, a_3\})$.

Proof of the Farkas–Minkowski proposition. Let $C = \text{cone}(\{a_1, \ldots, a_m\})$ be a polyhedral cone (nonempty) and assume that $b \notin C$. By Proposition 8.2, the polyhedral cone is closed, and by Proposition 11.5 there is some $z \in C$ such that $d(b, C) = \|b - z\|$; that is, z is a point of C closest to b. Since $b \notin C$ and $z \in C$ we have $u = z - b \neq 0$, and we claim that the linear hyperplane H orthogonal to u does the job, as illustrated in Figure 11.2.

First let us show that

$$\langle u, z\rangle = \langle z - b, z\rangle = 0. \tag{$*_1$}$$

This is trivial if $z = 0$, so assume $z \neq 0$. If $\langle u, z\rangle \neq 0$, then either $\langle u, z\rangle > 0$ or $\langle u, z\rangle < 0$. In either case we show that we can find some point $z' \in C$ closer to b than z is, a contradiction.

Case 1: $\langle u, z\rangle > 0$.

Let $z' = (1 - \alpha)z$ for any α such that $0 < \alpha < 1$. Then $z' \in C$ and since $u = z - b$,

$$z' - b = (1 - \alpha)z - (z - u) = u - \alpha z,$$

so

$$\|z' - b\|^2 = \|u - \alpha z\|^2 = \|u\|^2 - 2\alpha\langle u, z\rangle + \alpha^2 \|z\|^2 .$$

If we pick $\alpha > 0$ such that $\alpha < 2\langle u, z\rangle / \|z\|^2$, then $-2\alpha\langle u, z\rangle + \alpha^2 \|z\|^2 < 0$, so $\|z' - b\|^2 < \|u\|^2 = \|z - b\|^2$, contradicting the fact that z is a point of C closest to b.

Case 2: $\langle u, z\rangle < 0$.

Let $z' = (1 + \alpha)z$ for any α such that $\alpha \geq -1$. Then $z' \in C$ and since $u = z - b$, we have $z' - b = (1 + \alpha)z - (z - u) = u + \alpha z$ so

$$\|z' - b\|^2 = \|u + \alpha z\|^2 = \|u\|^2 + 2\alpha\langle u, z\rangle + \alpha^2 \|z\|^2 ,$$

and if

$$0 < \alpha < -2\langle u, z\rangle / \|z\|^2,$$

then $2\alpha\langle u, z\rangle + \alpha^2 \|z\|^2 < 0$, so $\|z' - b\|^2 < \|u\|^2 = \|z - b\|^2$, a contradiction as above.

Therefore $\langle u, z\rangle = 0$. We have

$$\langle u, u\rangle = \langle u, z - b\rangle = \langle u, z\rangle - \langle u, b\rangle = -\langle u, b\rangle,$$

and since $u \neq 0$, we have $\langle u, u\rangle > 0$, so $\langle u, u\rangle = -\langle u, b\rangle$ implies that

$$\langle u, b\rangle < 0. \tag{$*_2$}$$

It remains to prove that $\langle u, a_i\rangle \geq 0$ for $i = 1, \ldots, m$. Pick any $x \in C$ such that $x \neq z$. We claim that

$$\langle b - z, x - z\rangle \leq 0. \tag{$*_3$}$$

Otherwise $\langle b - z, x - z\rangle > 0$, that is, $\langle z - b, x - z\rangle < 0$, and we show that we can find some point $z' \in C$ on the line segment $[z, x]$ closer to b than z is.

For any α such that $0 \leq \alpha \leq 1$, we have $z' = (1 - \alpha)z + \alpha x = z + \alpha(x - z) \in C$, and since $z' - b = z - b + \alpha(x - z)$ we have

$$\|z' - b\|^2 = \|z - b + \alpha(x - z)\|^2 = \|z - b\|^2 + 2\alpha\langle z - b, x - z\rangle + \alpha^2 \|x - z\|^2,$$

so for any $\alpha > 0$ such that

$$\alpha < -2\langle z - b, x - z\rangle / \|x - z\|^2,$$

we have $2\alpha\langle z - b, x - z\rangle + \alpha^2 \|x - z\|^2 < 0$, which implies that $\|z' - b\|^2 < \|z - b\|^2$, contradicting that z is a point of C closest to b.

Since $\langle b - z, x - z\rangle \leq 0$, $u = z - b$, and by $(*_1)$, $\langle u, z\rangle = 0$, we have

$$0 \geq \langle b - z, x - z\rangle = \langle -u, x - z\rangle = -\langle u, x\rangle + \langle u, z\rangle = -\langle u, x\rangle,$$

which means that

$$\langle u, x\rangle \geq 0 \quad \text{for all } x \in C, \tag{$*_3$}$$

as claimed. In particular,

$$\langle u, a_i\rangle \geq 0 \quad \text{for } i = 1, \ldots, m. \tag{$*_4$}$$

Then by $(*_2)$ and $(*_4)$, the linear form defined by $y = u^\top$ satisfies the properties $yb < 0$ and $ya_i \geq 0$ for $i = 1, \ldots, m$, which proves the Farkas–Minkowski proposition. □

There are other ways of proving the Farkas–Minkowski proposition, for instance using minimally infeasible systems or Fourier–Motzkin elimination; see Matousek and Gardner [54] (Chapter 6, Sections 6.6 and 6.7).

11.2 The Duality Theorem in Linear Programming

Let (P) be the linear program

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

with A a $m \times n$ matrix, and assume that (P) has a feasible solution and is bounded above. Since by hypothesis the objective function $x \mapsto cx$ is bounded on $\mathcal{P}(A, b)$, it might be useful to deduce an *upper bound* for cx from the inequalities $Ax \leq b$, for any $x \in \mathcal{P}(A, b)$. We can do this as follows: for every inequality

$$a_i x \leq b_i \quad 1 \leq i \leq m,$$

pick a nonnegative scalar y_i, multiply both sides of the above inequality by y_i obtaining

$$y_i a_i x \leq y_i b_i \quad 1 \leq i \leq m,$$

(the direction of the inequality is preserved since $y_i \geq 0$), and then add up these m equations, which yields

$$(y_1 a_1 + \cdots + y_m a_m)x \leq y_1 b_1 + \cdots + y_m b_m.$$

If we can pick the $y_i \geq 0$ such that

$$c \leq y_1 a_1 + \cdots + y_m a_m,$$

then since $x_j \geq 0$, we have

$$cx \leq (y_1 a_1 + \cdots + y_m a_m)x \leq y_1 b_1 + \cdots + y_m b_m,$$

namely we found an upper bound of the value cx of the objective function of (P) for any feasible solution $x \in \mathcal{P}(A, b)$. If we let y be the linear form $y = (y_1, \ldots, y_m)$, then since

$$A = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}$$

$y_1 a_1 + \cdots + y_m a_m = yA$, and $y_1 b_1 + \cdots + y_m b_m = yb$, what we did was to look for some $y \in (\mathbb{R}^m)^*$ such that

$$c \leq yA, \quad y \geq 0,$$

so that we have

$$cx \leq yb \quad \text{for all } x \in \mathcal{P}(A, b). \tag{$*$}$$

Then it is natural to look for a "best" value of yb, namely a minimum value, which leads to the definition of the *dual* of the linear program (P), a notion due to John von Neumann.

Definition 11.2. Given any Linear Program (P)

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

with A an $m \times n$ matrix, the *dual* (D) of (P) is the following optimization problem:

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c \text{ and } y \geq 0, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$.

The variables $y_1, \ldots, y_m$ are called the *dual variables*. The original Linear Program (P) is called the *primal* linear program and the original variables $x_1, \ldots, x_n$ are the *primal variables*.

Here is an explicit example of a linear program and its dual.

Example 11.1. Consider the linear program illustrated by Figure 11.3

$$\begin{aligned} &\text{maximize} \quad 2x_1 + 3x_2 \\ &\text{subject to} \\ &\qquad 4x_1 + 8x_2 \leq 12 \\ &\qquad 2x_1 + x_2 \leq 3 \\ &\qquad 3x_1 + 2x_2 \leq 4 \\ &\qquad x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

Its dual linear program is illustrated in Figure 11.4

$$\begin{aligned} &\text{minimize} \quad 12y_1 + 3y_2 + 4y_3 \\ &\text{subject to} \\ &\qquad 4y_1 + 2y_2 + 3y_3 \geq 2 \\ &\qquad 8y_1 + y_2 + 2y_3 \geq 3 \\ &\qquad y_1 \geq 0,\ y_2 \geq 0,\ y_3 \geq 0. \end{aligned}$$

It can be checked that $(x_1, x_2) = (1/2, 5/4)$ is an optimal solution of the primal linear program, with the maximum value of the objective function $2x_1 + 3x_2$ equal to $19/4$, and that $(y_1, y_2, y_3) = (5/16, 0, 1/4)$ is an optimal solution of the dual linear program, with the minimum value of the objective function $12y_1 + 3y_2 + 4y_3$ also equal to $19/4$.

Observe that in the Primal Linear Program (P), we are looking for a *vector* $x \in \mathbb{R}^n$ maximizing the form cx, and that the constraints are determined by the action of the *rows* of the matrix A on x. On the other hand, in the Dual Linear Program (D), we are looking

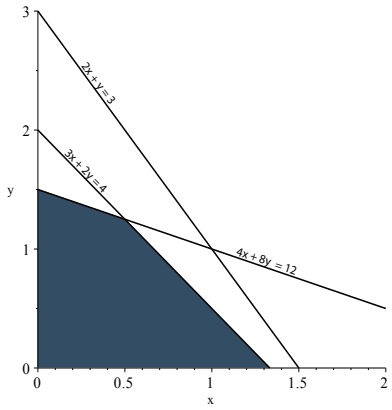


Figure 11.3: The $\mathcal{H}$-polytope for the linear program of Example 11.1. Note $x_1 \to x$ and $x_2 \to y$.

for a *linear form* $y \in (\mathbb{R}^*)^m$ minimizing the form yb, and the constraints are determined by the action of y on the *columns* of A. This is the sense in which (D) is the *dual* (P). In most presentations, the fact that (P) and (D) perform a search for a solution in spaces that are dual to each other is obscured by excessive use of transposition.

To convert the Dual Program (D) to a standard maximization problem we change the objective function yb to $-b^\top y^\top$ and the inequality $yA \geq c$ to $-A^\top y^\top \leq -c^\top$. The Dual Linear Program (D) is now stated as (D')

$$\begin{aligned} &\text{maximize} && -b^\top y^\top \\ &\text{subject to} && -A^\top y^\top \leq -c^\top \text{ and } y^\top \geq 0, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$. Observe that the dual in maximization form (D'') of the Dual Program (D') gives back the Primal Program (P).

The above discussion established the following inequality known as *weak duality*.

Proposition 11.6. *(Weak Duality) Given any Linear Program* (P)

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

with A *an* $m \times n$ *matrix, for any feasible solution* $x \in \mathbb{R}^n$ *of the Primal Problem* (P) *and every feasible solution* $y \in (\mathbb{R}^m)^*$ *of the Dual Problem* (D)*, we have*

$$cx \leq yb.$$

Definition 11.3. We say that the Dual Linear Program (D) is *bounded below* if $\{yb \mid y^\top \in \mathcal{P}(-A^\top, -c^\top)\}$ is bounded below.

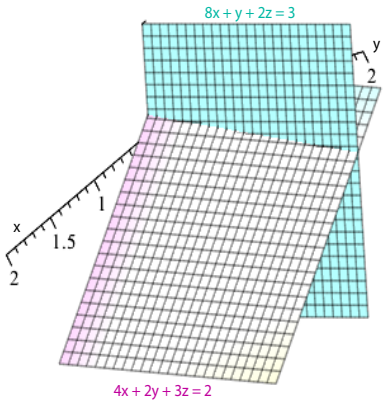


Figure 11.4: The $\mathcal{H}$-polyhedron for the dual linear program of Example 11.1 is the spacial region "above" the pink plane and in "front" of the blue plane. Note $y_1 \to x$, $y_2 \to y$, and $y_3 \to z$.

What happens if x^* is an optimal solution of (P) and if y^* is an optimal solution of (D)? We have $cx^* \leq y^*b$, but is there a "duality gap," that is, is it possible that $cx^* < y^*b$?

The answer is **no**, this is the *strong duality theorem*. Actually, the strong duality theorem asserts more than this.

Theorem 11.7. *(Strong Duality for Linear Programming) Let (P) be any linear program*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

with A an $m \times n$ matrix. The Primal Problem (P) has a feasible solution and is bounded above iff the Dual Problem (D) has a feasible solution and is bounded below. Furthermore, if (P) has a feasible solution and is bounded above, then for every optimal solution x^* of (P) and every optimal solution y^* of (D), we have

$$cx^* = y^*b.$$

Proof. If (P) has a feasible solution and is bounded above, then we know from Proposition 9.1 that (P) has some optimal solution. Let x^* be any optimal solution of (P). First we will show that (D) has a feasible solution v.

Let $\mu = cx^*$ be the maximum of the objective function $x \mapsto cx$. Then for any $\epsilon > 0$, the system of inequalities

$$Ax \leq b, \quad x \geq 0, \quad cx \geq \mu + \epsilon$$

has no solution, since otherwise μ would not be the maximum value of the objective function cx. We would like to apply Farkas II, so first we transform the above system of inequalities

into the system

$$\begin{pmatrix} A \\ -c \end{pmatrix} x \leq \begin{pmatrix} b \\ -(\mu + \epsilon) \end{pmatrix}.$$

By Proposition 11.4 (Farkas II), there is some linear form $(\lambda, z) \in (\mathbb{R}^{m+1})^*$ such that $\lambda \geq 0$, $z \geq 0$,

$$\begin{pmatrix} \lambda & z \end{pmatrix} \begin{pmatrix} A \\ -c \end{pmatrix} \geq 0_m^\top,$$

and

$$\begin{pmatrix} \lambda & z \end{pmatrix} \begin{pmatrix} b \\ -(\mu + \epsilon) \end{pmatrix} < 0,$$

which means that

$$\lambda A - zc \geq 0_m^\top, \quad \lambda b - z(\mu + \epsilon) < 0,$$

that is,

$$\begin{aligned} \lambda A &\geq zc \\ \lambda b &< z(\mu + \epsilon) \\ \lambda \geq 0, &\ \ z \geq 0. \end{aligned}$$

On the other hand, since $x^* \geq 0$ is an optimal solution of the system $Ax \leq b$, by Farkas II again (by taking the negation of the equivalence), since $\lambda A \geq 0$ (for the same λ as before), we must have

$$\lambda b \geq 0. \tag{$*_1$}$$

We claim that $z > 0$. Otherwise, since $z \geq 0$, we must have $z = 0$, but then

$$\lambda b < z(\mu + \epsilon)$$

implies

$$\lambda b < 0, \tag{$*_2$}$$

and since $\lambda b \geq 0$ by $(*_1)$, we have a contradiction. Consequently, we can divide by $z > 0$ without changing the direction of inequalities, and we obtain

$$\begin{aligned} \frac{\lambda}{z} A &\geq c \\ \frac{\lambda}{z} b &< \mu + \epsilon \\ \frac{\lambda}{z} &\geq 0, \end{aligned}$$

which shows that $v = \lambda/z$ is a feasible solution of the Dual Problem (D). However, weak duality (Proposition 11.6) implies that $cx^* = \mu \leq yb$ for any feasible solution $y \geq 0$ of the Dual Program (D), so (D) is bounded below and by Proposition 9.1 applied to the version of (D) written as a maximization problem, we conclude that (D) has some optimal solution.

For any optimal solution y^* of (D), since v is a feasible solution of (D) such that $vb < \mu + \epsilon$, we must have

$$\mu \leq y^* b < \mu + \epsilon,$$

and since our reasoning is valid for *any* $\epsilon > 0$, we conclude that $cx^* = \mu = y^* b$.

If we assume that the dual program (D) has a feasible solution and is bounded below, since the dual of (D) is (P), we conclude that (P) is also feasible and bounded above. □

The strong duality theorem can also be proven by the simplex method, because when it terminates with an optimal solution of (P), the final tableau also produces an optimal solution y of (D) that can be read off the reduced costs of columns $n+1, \ldots, n+m$ by flipping their signs. We follow the proof in Ciarlet [25] (Chapter 10).

Theorem 11.8. *Consider the Linear Program (P),*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

its equivalent version $(P2)$ *in standard form,*

$$\begin{aligned} &\text{maximize} \quad \widehat{c}\,\widehat{x} \\ &\text{subject to} \quad \widehat{A}\widehat{x} = b \text{ and } \widehat{x} \geq 0, \end{aligned}$$

where $\widehat{A}$ *is an* $m \times (n+m)$ *matrix,* $\widehat{c}$ *is a linear form in* $(\mathbb{R}^{n+m})^*$*, and* $\widehat{x} \in \mathbb{R}^{n+m}$*, given by*

$$\widehat{A} = \begin{pmatrix} A & I_m \end{pmatrix}, \quad \widehat{c} = \begin{pmatrix} c & 0_m^\top \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \overline{x} = \begin{pmatrix} x_{n+1} \\ \vdots \\ x_{n+m} \end{pmatrix}, \quad \widehat{x} = \begin{pmatrix} x \\ \overline{x} \end{pmatrix},$$

and the Dual (D) *of* (P) *given by*

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c \text{ and } y \geq 0, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$*. If the simplex algorithm applied to the Linear Program* $(P2)$ *terminates with an optimal solution* $(\widehat{u}^*, K^*)$*, where* $\widehat{u}^*$ *is a basic feasible solution and* K^* *is a basis for* $\widehat{u}^*$*, then* $y^* = \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}$ *is an optimal solution for* (D) *such that* $\widehat{c}\,\widehat{u}^* = y^* b$*. Furthermore,* y^* *is given in terms of the reduced costs by* $y^* = -((\overline{c}_{K^*})_{n+1} \ \ldots (\overline{c}_{K^*})_{n+m})$*.*

Proof. We know that K^* is a subset of $\{1, \ldots, n+m\}$ consisting of m indices such that the corresponding columns of $\widehat{A}$ are linearly independent. Let $N^* = \{1, \ldots, n+m\} - K^*$. The simplex method terminates with an optimal solution in Case (A), namely when

$$\widehat{c}_j - \sum_{k \in k} \gamma_k^j \widehat{c}_k \leq 0 \quad \text{for all } j \in N^*,$$

where $\widehat{A}^j = \sum_{k \in K^*} \gamma_k^j \widehat{A}^k$, or using the notations of Section 10.3,

$$\widehat{c}_j - \widehat{c}_{K^*} \widehat{A}_{K^*}^{-1} \widehat{A}^j \leq 0 \quad \text{for all } j \in N^*.$$

The above inequalities can be written as

$$\widehat{c}_{N^*} - \widehat{c}_{K^*} \widehat{A}_{K^*}^{-1} \widehat{A}_{N^*} \leq 0_n^\top,$$

or equivalently as

$$\widehat{c}_{K^*} \widehat{A}_{K^*}^{-1} \widehat{A}_{N^*} \geq \widehat{c}_{N^*}. \tag{$*_1$}$$

The value of the objective function for the optimal solution $\widehat{u}^*$ is $\widehat{c}\,\widehat{u}^* = \widehat{c}_{K^*} \widehat{u}_{K^*}^*$, and since $\widehat{u}_{K^*}^*$ satisfies the equation $\widehat{A}_{K^*} \widehat{u}_{K^*}^* = b$, the value of the objective function is

$$\widehat{c}_{K^*}\, \widehat{u}_{K^*}^* = \widehat{c}_{K^*} \widehat{A}_{K^*}^{-1} b. \tag{$*_2$}$$

Then if we let $y^* = \widehat{c}_{K^*} \widehat{A}_{K^*}^{-1}$, obviously we have $y^* b = \widehat{c}_{K^*} \widehat{u}_{K^*}$, so if we can prove that y^* is a feasible solution of the Dual Linear program (D), by weak duality, y^* is an optimal solution of (D). We have

$$y^* \widehat{A}_{K^*} = \widehat{c}_{K^*} \widehat{A}_{K^*}^{-1} \widehat{A}_{K^*} = \widehat{c}_{K^*}, \tag{$*_3$}$$

and by $(*_1)$ we get

$$y^* \widehat{A}_{N^*} = \widehat{c}_{K^*} \widehat{A}_{K^*}^{-1} \widehat{A}_{N^*} \geq \widehat{c}_{N^*}. \tag{$*_4$}$$

Let P be the $(n+m) \times (n+m)$ permutation matrix defined so that

$$\widehat{A}\, P = \begin{pmatrix} A & I_m \end{pmatrix} P = \begin{pmatrix} \widehat{A}_{K^*} & \widehat{A}_{N^*} \end{pmatrix}.$$

Then we also have

$$\widehat{c}\, P = \begin{pmatrix} c & 0_m^\top \end{pmatrix} P = \begin{pmatrix} \widehat{c}_{K^*} & \widehat{c}_{N^*} \end{pmatrix}.$$

Using Equations $(*_3)$ and $(*_4)$ we obtain

$$y^* \begin{pmatrix} \widehat{A}_{K^*} & \widehat{A}_{N^*} \end{pmatrix} \geq \begin{pmatrix} \widehat{c}_{K^*} & \widehat{c}_{N^*} \end{pmatrix},$$

that is,

$$y^* \begin{pmatrix} A & I_m \end{pmatrix} P \geq \begin{pmatrix} c & 0_m^\top \end{pmatrix} P,$$

which is equivalent to

$$y^* \begin{pmatrix} A & I_m \end{pmatrix} \geq \begin{pmatrix} c & 0_m^\top \end{pmatrix},$$

that is

$$y^* A \geq c, \quad y \geq 0,$$

and these are exactly the conditions that say that y^* is a feasible solution of the Dual Program (D).

The reduced costs are given by $(\widehat{c}_{K^*})_i = \widehat{c}_i - \widehat{c}_{K^*}\widehat{A}_{K^*}^{-1}\widehat{A}^i$, for $i = 1, \ldots, n+m$. But for $i = n + j$ with $j = 1, \ldots, m$ each column $\widehat{A}^{n+j}$ is the jth vector of the identity matrix I_m and by definition $\widehat{c}_{n+j} = 0$, so

$$(\widehat{c}_{K^*})_{n+j} = -(\widehat{c}_{K^*}\widehat{A}_{K^*}^{-1})_j = -y_j^* \quad j = 1, \ldots, m,$$

as claimed. □

The fact that the above proof is fairly short is deceptive because this proof relies on the fact that there are versions of the simplex algorithm using pivot rules that prevent cycling, but the proof that such pivot rules work correctly is quite lengthy. Other proofs are given in Matousek and Gardner [54] (Chapter 6, Sections 6.3), Chvatal [24] (Chapter 5), and Papadimitriou and Steiglitz [60] (Section 2.7).

Observe that since the last m rows of the final tableau are actually obtained by multipling $[u\ \widehat{A}]$ by $\widehat{A}_{K^*}^{-1}$, the $m \times m$ matrix consisting of the last m columns and last m rows of the final tableau is $\widehat{A}_{K^*}^{-1}$ (basically, the simplex algorithm has performed the steps of a Gauss–Jordan reduction). This fact allows saving some steps in the primal dual method.

By combining weak duality and strong duality, we obtain the following theorem which shows that exactly four cases arise.

Theorem 11.9. *(Duality Theorem of Linear Programming) Let (P) be any linear program*

$$\begin{aligned} &\text{maximize} && cx \\ &\text{subject to} && Ax \leq b \text{ and } x \geq 0, \end{aligned}$$

and let (D) be its dual program

$$\begin{aligned} &\text{minimize} && yb \\ &\text{subject to} && yA \geq c \text{ and } y \geq 0, \end{aligned}$$

with A an $m \times n$ matrix. Then exactly one of the following possibilities occur:

(1) Neither (P) nor (D) has a feasible solution.

(2) (P) is unbounded and (D) has no feasible solution.

(3) (P) has no feasible solution and (D) is unbounded.

(4) Both (P) and (D) have a feasible solution. Then both have an optimal solution, and for every optimal solution x^* of (P) and every optimal solution y^* of (D), we have

$$cx^* = y^*b.$$

An interesting corollary of Theorem 11.9 is that there is a test to determine whether a Linear Program (P) has an optimal solution.

Corollary 11.10. *The Primal Program (P) has an optimal solution iff the following set of constraints is satisfiable:*

$$\begin{aligned} Ax &\leq b \\ yA &\geq c \\ cx &\geq yb \\ x \geq 0,\ y &\geq 0_m^\top. \end{aligned}$$

In fact, for any feasible solution (x^*, y^*) of the above system, x^* is an optimal solution of (P) and y^* is an optimal solution of (D)

11.3 Complementary Slackness Conditions

Another useful corollary of the strong duality theorem is the following result known as the *equilibrium theorem*.

Theorem 11.11. *(Equilibrium Theorem) For any Linear Program (P) and its Dual Linear Program (D) (with set of inequalities $Ax \leq b$ where A is an $m \times n$ matrix, and objective function $x \mapsto cx$), for any feasible solution x of (P) and any feasible solution y of (D), x and y are optimal solutions iff*

$$y_i = 0 \qquad \text{for all } i \text{ for which } \textstyle\sum_{j=1}^n a_{ij}x_j < b_i \tag{$*_D$}$$

and

$$x_j = 0 \qquad \text{for all } j \text{ for which } \textstyle\sum_{i=1}^m y_i a_{ij} > c_j. \tag{$*_P$}$$

Proof. First assume that $(*_D)$ and $(*_P)$ hold. The equations in $(*_D)$ say that $y_i = 0$ unless $\sum_{j=1}^n a_{ij}x_j = b_i$, hence

$$yb = \sum_{i=1}^m y_i b_i = \sum_{i=1}^m y_i \sum_{j=1}^n a_{ij}x_j = \sum_{i=1}^m \sum_{j=1}^n y_i a_{ij} x_j.$$

Similarly, the equations in $(*_P)$ say that $x_j = 0$ unless $\sum_{i=1}^m y_i a_{ij} = c_j$, hence

$$cx = \sum_{j=1}^n c_j x_j = \sum_{j=1}^n \sum_{i=1}^m y_i a_{ij} x_j.$$

Consequently, we obtain

$$cx = yb.$$

By weak duality (Proposition 11.6), we have

$$cx \leq yb = cx$$

for all feasible solutions x of (P), so x is an optimal solution of (P). Similarly,

$$yb = cx \leq yb$$

for all feasible solutions y of (D), so y is an optimal solution of (D).

Let us now assume that x is an optimal solution of (P) and that y is an optimal solution of (D). Then as in the proof of Proposition 11.6,

$$\sum_{j=1}^{n} c_j x_j \leq \sum_{i=1}^{m}\sum_{j=1}^{n} y_i a_{ij} x_j \leq \sum_{i=1}^{m} y_i b_i.$$

By strong duality, since x and y are optimal solutions the above inequalities are actually equalities, so in particular we have

$$\sum_{j=1}^{n}\left(c_j - \sum_{i=1}^{m} y_i a_{ij}\right)x_j = 0.$$

Since x and y are feasible, $x_i \geq 0$ and $y_j \geq 0$, so if $\sum_{i=1}^{m} y_i a_{ij} > c_j$, we must have $x_j = 0$. Similarly, we have

$$\sum_{i=1}^{m} y_i\left(\sum_{j=1}^{m} a_{ij} x_j - b_i\right) = 0,$$

so if $\sum_{j=1}^{m} a_{ij} x_j < b_i$, then $y_i = 0$. □

The equations in $(*_D)$ and $(*_P)$ are often called *complementary slackness conditions*. These conditions can be exploited to solve for an optimal solution of the primal problem with the help of the dual problem, and conversely. Indeed, if we guess a solution to one problem, then we may solve for a solution of the dual using the complementary slackness conditions, and then check that our guess was correct. This is the essence of the *primal-dual* method. To present this method, first we need to take a closer look at the dual of a linear program already in standard form.

11.4 Duality for Linear Programs in Standard Form

Let (P) be a linear program in standard form, where $Ax = b$ for some $m \times n$ matrix of rank m and some objective function $x \mapsto cx$ (of course, $x \geq 0$). To obtain the dual of (P) we convert the equations $Ax = b$ to the following system of inequalities involving a $(2m) \times n$ matrix:

$$\begin{pmatrix} A \\ -A \end{pmatrix} x \leq \begin{pmatrix} b \\ -b \end{pmatrix}.$$

Then if we denote the $2m$ dual variables by (y', y''), with $y', y'' \in (\mathbb{R}^m)^*$, the dual of the above program is

$$\begin{aligned} &\text{minimize} \quad y'b - y''b \\ &\text{subject to} \quad \begin{pmatrix} y' & y'' \end{pmatrix} \begin{pmatrix} A \\ -A \end{pmatrix} \geq c \text{ and } y', y'' \geq 0, \end{aligned}$$

where $y', y'' \in (\mathbb{R}^m)^*$, which is equivalent to

$$\begin{aligned} &\text{minimize} \quad (y' - y'')b \\ &\text{subject to} \quad (y' - y'')A \geq c \text{ and } y', y'' \geq 0, \end{aligned}$$

where $y', y'' \in (\mathbb{R}^m)^*$. If we write $y = y' - y''$, we find that the above linear program is equivalent to the following Linear Program (D):

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$. Observe that y is *not required* to be nonnegative; it is arbitrary.

Next we would like to know what is the version of Theorem 11.8 for a linear program already in standard form. This is very simple.

Theorem 11.12. *Consider the Linear Program (P2) in standard form*

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

and its Dual (D) given by

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$. If the simplex algorithm applied to the Linear Program $(P2)$ terminates with an optimal solution (u^*, K^*), where u^* is a basic feasible solution and K^* is a basis for u^*, then $y^* = c_{K^*}A_{K^*}^{-1}$ is an optimal solution for (D) such that $cu^* = y^*b$. Furthermore, if we assume that the simplex algorithm is started with a basic feasible solution (u_0, K_0) where $K_0 = (n-m+1, \ldots, n)$ (the indices of the last m columns of A) and $A_{(n-m+1,\ldots,n)} = I_m$ (the last m columns of A constitute the identity matrix I_m), then the optimal solution $y^* = c_{K^*}A_{K^*}^{-1}$ for (D) is given in terms of the reduced costs by

$$y^* = c_{(n-m+1,\ldots,n)} - (\overline{c}_{K^*})_{(n-m+1,\ldots,n)},$$

and the $m \times m$ matrix consisting of last m columns and the last m rows of the final tableau is $A_{K^*}^{-1}$.

Proof. The proof of Theorem 11.8 applies with A instead of $\widehat{A}$, and we can show that

$$c_{K^*}A_{K^*}^{-1}A_{N^*} \geq c_{N^*},$$

and that $y^* = c_{K^*}A_{K^*}^{-1}$ satisfies, $cu^* = y^*b$, and

$$\begin{aligned} y^*A_{K^*} &= c_{K^*}A_{K^*}^{-1}A_{K^*} = c_{K^*}, \\ y^*A_{N^*} &= c_{K^*}A_{K^*}^{-1}A_{N^*} \geq c_{N^*}. \end{aligned}$$

Let P be the $n \times n$ permutation matrix defined so that

$$AP = \begin{pmatrix} A_{K^*} & A_{N^*} \end{pmatrix}.$$

Then we also have

$$cP = \begin{pmatrix} c_{K^*} & c_{N^*} \end{pmatrix},$$

and using the above equations and inequalities we obtain

$$y^* \begin{pmatrix} A_{K^*} & A_{N^*} \end{pmatrix} \geq \begin{pmatrix} c_{K^*} & c_{N^*} \end{pmatrix},$$

that is, $y^*AP \geq cP$, which is equivalent to

$$y^*A \geq c,$$

which shows that y^* is a feasible solution of (D) (remember, y^* is arbitrary so there is no need for the constraint $y^* \geq 0$).

The reduced costs are given by

$$(\overline{c}_{K^*})_i = c_i - c_{K^*}A_{K^*}^{-1}A^i,$$

and since for $j = n-m+1, \ldots, n$ the column A^j is the $(j+m-n)$th column of the identity matrix I_m, we have

$$(\overline{c}_{K^*})_j = c_j - (c_{K^*}A_{K^*}^{-1})_{j+m-n} \quad j = n-m+1, \ldots, n,$$

that is,

$$y^* = c_{(n-m+1,\ldots,n)} - (\overline{c}_{K^*})_{(n-m+1,\ldots,n)},$$

as claimed. Since the last m rows of the final tableau is obtained by multiplying $[u_0 \;\; A]$ by $A_{K^*}^{-1}$, and the last m columns of A constitute I_m, the last m rows and the last m columns of the final tableau constitute $A_{K^*}^{-1}$. □

Let us now take a look at the complementary slackness conditions of Theorem 11.11. If we go back to the version of (P) given by

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad \begin{pmatrix} A \\ -A \end{pmatrix} x \leq \begin{pmatrix} b \\ -b \end{pmatrix} \text{ and } x \geq 0, \end{aligned}$$

and to the version of (D) given by

$$\begin{aligned} &\text{minimize} \quad y'b - y''b \\ &\text{subject to} \quad \begin{pmatrix} y' & y'' \end{pmatrix} \begin{pmatrix} A \\ -A \end{pmatrix} \geq c \text{ and } y', y'' \geq 0, \end{aligned}$$

where $y', y'' \in (\mathbb{R}^m)^*$, since the inequalities $Ax \leq b$ and $-Ax \leq -b$ together imply that $Ax = b$, we have equality for all these inequality constraints, and so the Conditions $(*_D)$ place no constraints at all on y' and y'', while the Conditions $(*_P)$ assert that

$$x_j = 0 \qquad \text{for all } j \text{ for which } \textstyle\sum_{i=1}^m (y'_i - y''_i)a_{ij} > c_j.$$

If we write $y = y' - y''$, the above conditions are equivalent to

$$x_j = 0 \qquad \text{for all } j \text{ for which } \textstyle\sum_{i=1}^m y_i a_{ij} > c_j.$$

Thus we have the following version of Theorem 11.11.

Theorem 11.13. *(Equilibrium Theorem, Version 2) For any Linear Program $(P2)$ in standard form (with $Ax = b$ where A is an $m \times n$ matrix, $x \geq 0$, and objective function $x \mapsto cx$) and its Dual Linear Program (D), for any feasible solution x of (P) and any feasible solution y of (D), x and y are optimal solutions iff*

$$x_j = 0 \qquad \textit{for all } j \textit{ for which } \textstyle\sum_{i=1}^m y_i a_{ij} > c_j. \qquad (*_P)$$

Therefore, the slackness conditions applied to a Linear Program $(P2)$ in standard form and to its Dual (D) *only impose slackness conditions on the variables x_j of the primal problem*.

The above fact plays a crucial role in the primal-dual method.

11.5 The Dual Simplex Algorithm

Given a Linear Program $(P2)$ in standard form

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

where A is an $m \times n$ matrix of rank m, if no obvious feasible solution is available *but if* $c \leq 0$, rather than using the method for finding a feasible solution described in Section 10.2 we may use a method known as the dual simplex algorithm. This method uses basic solutions (u, K) where $Au = b$ and $u_j = 0$ for all $u_j \notin K$, but does not require $u \geq 0$, so u may not be feasible. However, $y = c_K A_K^{-1}$ is required to be feasible for the dual program

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c, \end{aligned}$$

where $y \in (\mathbb{R}^*)^m$. Since $c \leq 0$, observe that $y = 0_m^\top$ is a feasible solution of the dual.

If a basic solution u of $(P2)$ is found such that $u \geq 0$, then $cu = yb$ for $y = c_K A_K^{-1}$, and we have found an optimal solution u for $(P2)$ and y for (D). The dual simplex method makes progress by attempting to make negative components of u zero and by decreasing the objective function of the dual program.

The dual simplex method starts with a basic solution (u, K) of $Ax = b$ which is not feasible but for which $y = c_K A_K^{-1}$ is dual feasible. In many cases the original linear program is specified by a set of inequalities $Ax \leq b$ with some $b_i < 0$, so by adding slack variables it is easy to find such basic solution u, and if in addition $c \leq 0$, then because the cost associated with slack variables is 0, we see that $y = 0$ is a feasible solution of the dual.

Given a basic solution (u, K) of $Ax = b$ (feasible or not), $y = c_K A_K^{-1}$ is dual feasible iff $c_K A_K^{-1} A \geq c$, and since $c_K A_K^{-1} A_K = c_K$, the inequality $c_K A_K^{-1} A \geq c$ is equivalent to $c_K A_K^{-1} A_N \geq c_N$, that is,

$$c_N - c_K A_K^{-1} A_N \leq 0, \tag{$*_1$}$$

where $N = \{1, \ldots, n\} - K$. Equation $(*_1)$ is equivalent to

$$c_j - c_K \gamma_K^j \leq 0 \quad \text{for all } j \in N, \tag{$*_2$}$$

where $\gamma_K^j = A_K^{-1} A^j$. Recall that the notation $\overline{c}_j$ is used to denote $c_j - c_K \gamma_K^j$, which is called the *reduced cost* of the variable x_j.

As in the simplex algorithm we need to decide which column A^k leaves the basis K and which column A^j enters the new basis K^+, in such a way that $y^+ = c_{K^+} A_{K^+}^{-1}$ is a feasible solution of (D), that is, $c_{N^+} - c_{K^+} A_{K^+}^{-1} A_{N^+} \leq 0$, where $N^+ = \{1, \ldots, n\} - K^+$. We use Proposition 10.2 to decide wich column k^- should leave the basis.

Suppose (u, K) is a solution of $Ax = b$ for which $y = c_K A_K^{-1}$ is dual feasible.

Case (A). If $u \geq 0$, then u is an optimal solution of $(P2)$.

Case (B). There is some $k \in K$ such that $u_k < 0$. In this case pick some $k^- \in K$ such that $u_{k^-} < 0$ (according to some pivot rule).

Case (B1). Suppose that $\gamma_{k^-}^j \geq 0$ for all $j \notin K$ (in fact, for all j, since $\gamma_{k^-}^j \in \{0, 1\}$ for all $j \in K$). If so, we we claim that $(P2)$ is not feasible.

Indeed, let v be some basic feasible solution. We have $v \geq 0$ and $Av = b$, that is,

$$\sum_{j=1}^{n} v_j A^j = b,$$

so by multiplying both sides by A_K^{-1} and using the fact that by definition $\gamma_K^j = A_K^{-1} A^j$, we obtain

$$\sum_{j=1}^{n} v_j \gamma_K^j = A_K^{-1} b = u_K.$$

But recall that by hypothesis $u_{k^-} < 0$, yet $v_j \geq 0$ and $\gamma^j_{k^-} \geq 0$ for all j, so the component of index k^- is zero or positive on the left, and negative on the right, a contradiction. Therefore, $(P2)$ is indeed not feasible.

Case (B2). We have $\gamma^j_{k^-} < 0$ for some j.

We pick the column A^j entering the basis among those for which $\gamma^j_{k^-} < 0$. Since we assumed that $c_j - c_K\gamma^j_K \leq 0$ for all $j \in N$ by $(*_2)$, consider

$$\mu^+ = \max\left\{-\frac{c_j - c_K\gamma^j_K}{\gamma^j_{k^-}} \;\middle|\; \gamma^j_{k^-} < 0,\ j \in N\right\} = \max\left\{-\frac{\overline{c}_j}{\gamma^j_{k^-}} \;\middle|\; \gamma^j_{k^-} < 0,\ j \in N\right\} \leq 0,$$

and the set

$$N(\mu^+) = \left\{j \in N \;\middle|\; -\frac{\overline{c}_j}{\gamma^j_{k^-}} = \mu^+\right\}.$$

We pick some index $j^+ \in N(\mu^+)$ as the index of the column entering the basis (using some pivot rule).

Recall that by hypothesis $c_i - c_K\gamma^i_K \leq 0$ for all $j \notin K$ and $c_i - c_K\gamma^i_K = 0$ for all $i \in K$. Since $\gamma^{j^+}_{k^-} < 0$, for any index i such that $\gamma^i_{k^-} \geq 0$, we have $-\gamma^i_{k^-}/\gamma^{j^+}_{k^-} \geq 0$, and since by Proposition 10.2

$$c_i - c_{K^+}\gamma^i_{K^+} = c_i - c_K\gamma^i_K - \frac{\gamma^i_{k^-}}{\gamma^{j^+}_{k^-}}(c_{j^+} - c_K\gamma^{j^+}_K),$$

we have $c_i - c_{K^+}\gamma^i_{K^+} \leq 0$. For any index i such that $\gamma^i_{k^-} < 0$, by the choice of $j^+ \in K^*$,

$$-\frac{c_i - c_K\gamma^i_K}{\gamma^i_{k^-}} \leq -\frac{c_{j^+} - c_K\gamma^{j^+}_K}{\gamma^{j^+}_{k^-}},$$

so

$$c_i - c_K\gamma^i_K - \frac{\gamma^i_{k^-}}{\gamma^{j^+}_{k^-}}(c_{j^+} - c_K\gamma^{j^+}_K) \leq 0,$$

and again, $c_i - c_{K^+}\gamma^i_{K^+} \leq 0$. Therefore, if we let $K^+ = (K - \{k^-\}) \cup \{j^+\}$, then $y^+ = c_{K^+}A^{-1}_{K^+}$ is dual feasible. As in the simplex algorithm, θ^+ is given by

$$\theta^+ = u_{k^-}/\gamma^{j^+}_{k^-} \geq 0,$$

and u^+ is also computed as in the simplex algorithm by

$$u^+_i = \begin{cases} u_i - \theta^{j^+}\gamma^{j^+}_i & \text{if } i \in K \\ \theta^{j^+} & \text{if } i = j^+ \\ 0 & \text{if } i \notin K \cup \{j^+\} \end{cases}.$$

The change in the objective function of the primal and dual program (which is the same, since $u_K = A_K^{-1}b$ and $y = c_K A_K^{-1}$ is chosen such that $cu = c_K u_K = yb$) is the same as in the simplex algorithm, namely

$$\theta^+\left(c^{j^+} - c_K\gamma_K^{j^+}\right).$$

We have $\theta^+ > 0$ and $c^{j^+} - c_K\gamma_K^{j^+} \leq 0$, so if $c^{j^+} - c_K\gamma_K^{j^+} < 0$, then the objective function of the dual program decreases strictly.

Case (B3). $\mu^+ = 0$.

The possibity that $\mu^+ = 0$, that is, $c^{j^+} - c_K\gamma_K^{j^+} = 0$, may arise. In this case, the objective function doesn't change. This is a case of degeneracy similar to the degeneracy that arises in the simplex algorithm. We still pick $j^+ \in N(\mu^+)$, but we need a pivot rule that prevents cycling. Such rules exist; see Bertsimas and Tsitsiklis [14] (Section 4.5) and Papadimitriou and Steiglitz [60] (Section 3.6).

The reader surely noticed that the dual simplex algorithm is very similar to the simplex algorithm, except that the simplex algorithm preserves the property that (u, K) is (primal) feasible, whereas the dual simplex algorithm preserves the property that $y = c_K A_K^{-1}$ is dual feasible. One might then wonder whether the dual simplex algorithm is equivalent to the simplex algorithm applied to the dual problem. This is indeed the case, there is a one-to-one correspondence between the dual simplex algorithm and the simplex algorithm applied to the dual problem in maximization form. This correspondence is described in Papadimitriou and Steiglitz [60] (Section 3.7).

The comparison between the simplex algorithm and the dual simplex algorithm is best illustrated if we use a description of these methods in terms of *(full) tableaux*.

Recall that a *(full) tableau* is an $(m+1) \times (n+1)$ matrix organized as follows:

$-c_K u_K$	$\overline{c}_1$	$\cdots$	$\overline{c}_j$	$\cdots$	$\overline{c}_n$
u_{k_1}	γ_1^1	$\cdots$	γ_1^j	$\cdots$	γ_1^n
$\vdots$	$\vdots$		$\vdots$		$\vdots$
u_{k_m}	γ_m^1	$\cdots$	γ_m^j	$\cdots$	γ_m^n

The top row contains the current value of the objective function and the reduced costs, the first column except for its top entry contain the components of the current basic solution u_K, and the remaining columns except for their top entry contain the vectors γ_K^j. Observe that the γ_K^j corresponding to indices j in K constitute a permutation of the identity matrix I_m. A tableau together with the new basis $K^+ = (K - \{k^-\}) \cup \{j^+\}$ contains all the data needed to compute the new u_{K^+}, the new $\gamma_{K^+}^j$, and the new reduced costs $\overline{c}_i - (\gamma_{k^-}^i/\gamma_{k^-}^{j^+})\overline{c}_{j^+}$.

When executing the simplex algorithm, we have $u_k \geq 0$ for all $k \in K$ (and $u_j = 0$ for all $j \notin K$), and the incoming column j^+ is determined by picking one of the column indices such that $\overline{c}_j > 0$. Then the index k^- of the leaving column is determined by looking at the minimum of the ratios $u_k/\gamma_k^{j^+}$ for which $\gamma_k^{j^+} > 0$ (along column j^+).

On the other hand, when executing the dual simplex algorithm, we have $\overline{c}_j \leq 0$ for all $j \notin K$ (and $\overline{c}_k = 0$ for all $k \in K$), and the outgoing column k^- is determined by picking one of the row indices such that $u_k < 0$. The index j^+ of the incoming column is determined by looking at the maximum of the ratios $-\overline{c}_j/\gamma^j_{k^-}$ for which $\gamma^j_{k^-} < 0$ (along row k^-).

More details about the comparison between the simplex algorithm and the dual simplex algorithm can be found in Bertsimas and Tsitsiklis [14] and Papadimitriou and Steiglitz [60].

Here is an example of the the dual simplex method.

Example 11.2. Consider the following linear program in standard form:

$$\begin{aligned} &\text{Maximize} \quad -4x_1 - 2x_2 - x_3 \\ &\text{subject to} \quad \begin{pmatrix} -1 & -1 & 2 & 1 & 0 & 0 \\ -4 & -2 & 1 & 0 & 1 & 0 \\ 1 & 1 & -4 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} -3 \\ -4 \\ 2 \end{pmatrix} \quad \text{and} \quad x_1, x_2, x_3, x_4, x_5, x_6 \geq 0. \end{aligned}$$

We initialize the dual simplex procedure with (u, K) where $u = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -3 \\ -4 \\ 2 \end{pmatrix}$ and $K = (4, 5, 6)$.

The initial tableau, before explicitly calculating the reduced cost, is

0	$\overline{c}_1$	$\overline{c}_2$	$\overline{c}_3$	$\overline{c}_4$	$\overline{c}_5$	$\overline{c}_6$
$u_4 = -3$	-1	-1	2	1	0	0
$u_5 = -4$	-4	-2	1	0	1	0
$u_6 = 2$	1	1	-4	0	0	1

.

Since u has negative coordinates, Case (B) applies, and we will set $k^- = 4$. We must now determine whether Case (B1) or Case (B2) applies. This determination is accomplished by scanning the first three columns in the tableau and observing each column has a negative entry. Thus Case (B2) is applicable, and we need to determine the reduced costs. Observe that $c = (-4, -2, -1, 0, 0, 0)$, which in turn implies $c_{(4,5,6)} = (0, 0, 0)$. Equation $(*_2)$ implies

that the nonzero reduced costs are

$$\overline{c}_1 = c_1 - c_{(4,5,6)} \begin{pmatrix} -1 \\ -4 \\ 1 \end{pmatrix} = -4$$
$$\overline{c}_2 = c_2 - c_{(4,5,6)} \begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix} = -2$$
$$\overline{c}_3 = c_3 - c_{(4,5,6)} \begin{pmatrix} -2 \\ 1 \\ 4 \end{pmatrix} = -1,$$

and our tableau becomes

0	-4	-2	-1	0	0	0
$u_4 = -3$	-1	-1	2	1	0	0
$u_5 = -4$	-4	-2	1	0	1	0
$u_6 = 2$	1	1	-4	0	0	1

.

Since $k^- = 4$, our pivot row is the first row of the tableau. To determine candidates for j^+, we scan this row, locate negative entries and compute

$$\mu^+ = \max\left\{ -\frac{\overline{c}_j}{\gamma_4^j} \;\middle|\; \gamma_4^j < 0,\ j \in \{1,2,3\} \right\} = \max\left\{ \frac{-2}{1}, \frac{-4}{1} \right\} = -2.$$

Since μ^+ occurs when $j = 2$, we set $j^+ = 2$. Our new basis is $K^+ = (2,5,6)$. We must normalize the first row of the tableau, namely multiply by -1, then add twice this normalized row to the second row, and subtract the normalized row from the third row to obtain the updated tableau.

0	-4	-2	-1	0	0	0
$u_2 = 3$	1	1	-2	-1	0	0
$u_5 = 2$	-2	0	-3	-2	1	0
$u_6 = -1$	0	0	-2	1	0	1

It remains to update the reduced costs and the value of the objective function by adding twice the normalized row to the top row.

6	-2	0	-5	-2	0	0
$u_2 = 3$	1	1	-2	-1	0	0
$u_5 = 2$	-2	0	-3	-2	1	0
$u_6 = -1$	0	0	-2	1	0	1

We now repeat the procedure of Case (B2) and set $k^- = 6$ (since this is the only negative entry of u^+). Our pivot row is now the third row of the updated tableau, and the new μ^+

becomes

$$\mu^+ = \max\left\{ -\frac{\bar{c}_j}{\gamma_6^j} \;\middle|\; \gamma_6^j < 0,\ j \in \{1,3,4\} \right\} = \max\left\{ \frac{-5}{2} \right\} = -\frac{5}{2},$$

which implies that $j^+ = 3$. Hence the new basis is $K^+ = (2,5,3)$, and we update the tableau by taking $-\frac{1}{2}$ of Row 3, adding twice the normalized Row 3 to Row 1, and adding three times the normalized Row 3 to Row 2.

6	-2	0	-5	-2	0	0
$u_2 = 4$	1	1	0	-2	0	-1
$u_5 = 7/2$	-2	0	0	$-7/2$	1	$-3/2$
$u_3 = 1/2$	0	0	(1)	$-1/2$	0	$-1/2$

It remains to update the objective function and the reduced costs by adding five times the normalized row to the top row.

17/2	-2	0	0	$-9/2$	0	$-5/2$
$u_2 = 4$	1	1	0	-2	0	-1
$u_5 = 7/2$	-2	0	0	$-\frac{7}{2}$	1	$-3/2$
$u_3 = 1/2$	0	0	(1)	$-1/2$	0	$-1/2$

Since u^+ has no negative entries, the dual simplex method terminates and objective function $-4x_1 - 2x_2 - x_3$ is maximized with $-\frac{17}{2}$ at $(0, 4, \frac{1}{2})$. See Figure 11.5.

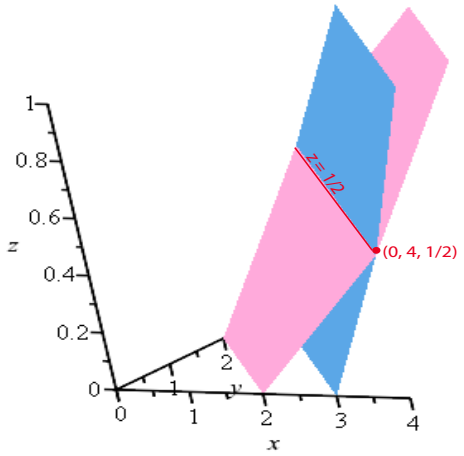


Figure 11.5: The objective function $-4x_1 - 2x_2 - x_3$ is maximized at the intersection between the blue plane $-x_1 - x_2 + 2x_3 = -3$ and the pink plane $x_1 + x_2 - 4x_3 = 2$.

11.6 The Primal-Dual Algorithm

Let $(P2)$ be a linear program in standard form

$$\begin{aligned} &\text{maximize} \quad cx \\ &\text{subject to} \quad Ax = b \text{ and } x \geq 0, \end{aligned}$$

where A is an $m \times n$ matrix of rank m, and (D) be its dual given by

$$\begin{aligned} &\text{minimize} \quad yb \\ &\text{subject to} \quad yA \geq c, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$.

First we may assume that $b \geq 0$ by changing every equation $\sum_{j=1}^n a_{ij}x_j = b_i$ with $b_i < 0$ to $\sum_{j=1}^n -a_{ij}x_j = -b_i$. If we happen to have some feasible solution y of the dual program (D), we know from Theorem 11.13 that a feasible solution x of $(P2)$ is an optimal solution iff the equations in $(*_P)$ hold. If we denote by J the subset of $\{1, \ldots, n\}$ for which the equalities

$$yA^j = c_j$$

hold, then by Theorem 11.13 a feasible solution x of $(P2)$ is an optimal solution iff

$$x_j = 0 \quad \text{for all } j \notin J.$$

Let $|J| = p$ and $N = \{1, \ldots, n\} - J$. The above suggests looking for $x \in \mathbb{R}^n$ such that

$$\begin{aligned} \sum_{j \in J} x_j A^j &= b \\ x_j &\geq 0 \quad \text{for all } j \in J \\ x_j &= 0 \quad \text{for all } j \notin J, \end{aligned}$$

or equivalently

$$A_J x_J = b, \quad x_J \geq 0, \tag{$*_1$}$$

and

$$x_N = 0_{n-p}.$$

To search for such an x, we just need to look for a feasible x_J, and for this we can use the *Restricted Primal* linear program (RP) defined as follows:

$$\begin{aligned} &\text{maximize} \quad -(\xi_1 + \cdots + \xi_m) \\ &\text{subject to} \quad \begin{pmatrix} A_J & I_m \end{pmatrix} \begin{pmatrix} x_J \\ \xi \end{pmatrix} = b \text{ and } x, \xi \geq 0. \end{aligned}$$

Since by hypothesis $b \geq 0$ and the objective function is bounded above by 0, this linear program has an optimal solution (x_J^*, ξ^*).

If $\xi^* = 0$, then the vector $u^* \in \mathbb{R}^n$ given by $u_J^* = x_J^*$ and $u_N^* = 0_{n-p}$ is an optimal solution of (P).

Otherwise, $\xi^* > 0$ and we have failed to solve $(*_1)$. However we may try to use ξ^* to improve y. For this consider the Dual (DRP) of (RP):

$$\begin{aligned} &\text{minimize} && zb \\ &\text{subject to} && zA_J \geq 0 \\ &&& z \geq -\mathbf{1}_m^\top. \end{aligned}$$

Observe that the Program (DRP) has the same objective function as the original Dual Program (D). We know by Theorem 11.12 that the optimal solution (x_J^*, ξ^*) of (RP) yields an optimal solution z^* of (DRP) such that

$$z^* b = -(\xi_1^* + \cdots + \xi_m^*) < 0.$$

In fact, if K^* is the basis associated with (x_J^*, ξ^*) and if we write

$$\widehat{A} = \begin{pmatrix} A_J & I_m \end{pmatrix}$$

and $\widehat{c} = [0_p^\top \quad -\mathbf{1}^\top]$, then by Theorem 11.12 we have

$$z^* = \widehat{c}_{K^*} \widehat{A}_{K^*}^{-1} = -\mathbf{1}_m^\top - (\overline{c}_{K^*})_{(p+1,\ldots,p+m)},$$

where $(\overline{c}_{K^*})_{(p+1,\ldots,p+m)}$ denotes the row vector of reduced costs in the final tableau corresponding to the last m columns.

If we write

$$y(\theta) = y + \theta z^*,$$

then the new value of the objective function of (D) is

$$y(\theta) b = yb + \theta z^* b, \tag{$*_2$}$$

and since $z^* b < 0$, we have a chance of improving the objective function of (D), that is, decreasing its value for $\theta > 0$ small enough if $y(\theta)$ is feasible for (D). This will be the case iff $y(\theta) A \geq c$ iff

$$yA + \theta z^* A \geq c. \tag{$*_3$}$$

Now since y is a feasible solution of (D) we have $yA \geq c$, so if $z^* A \geq 0$, then $(*_3)$ is satisfied and $y(\theta)$ is a solution of (D) for all $\theta > 0$, which means that (D) is unbounded. But this implies that (P) is not feasible.

Let us take a closer look at the inequalities $z^*A \geq 0$. For $j \in J$, since z^* is an optimal solution of (DRP), we know that $z^*A_J \geq 0$, so if $z^*A^j \geq 0$ for all $j \in N$, then $(P2)$ is not feasible.

Otherwise, there is some $j \in N = \{1, \ldots, n\} - J$ such that

$$z^*A^j < 0,$$

and then since by the definition of N we have $yA^j > c_j$ for all $j \in N$, if we pick θ such that

$$0 < \theta \leq \frac{yA^j - c_j}{-z^*A^j} \quad j \in N,\ z^*A^j < 0,$$

then we decrease the objective function $y(\theta)b = yb + \theta z^*b$ of (D) (since $z^*b < 0$). Therefore we pick the best θ, namely

$$\theta^+ = \min\left\{ \frac{yA^j - c_j}{-z^*A^j} \;\middle|\; j \notin J,\ z^*A^j < 0 \right\} > 0. \tag{$*_4$}$$

Next we update y to $y^+ = y(\theta^+) = y + \theta^+ z^*$, we create the new restricted primal with the new subset

$$J^+ = \{j \in \{1, \ldots, n\} \mid y^+A^j = c_j\},$$

and repeat the process.

Here are the steps of the primal-dual algorithm.

Step 1. Find some feasible solution y of the Dual Program (D). We will show later that this is always possible.

Step 2. Compute

$$J^+ = \{j \in \{1, \ldots, n\} \mid yA^j = c_j\}.$$

Step 3. Set $J = J^+$ and solve the Problem (RP) using the simplex algorithm, starting from the optimal solution determined during the previous round, obtaining the optimal solution (x_J^*, ξ^*) with the basis K^*.

Step 4.

If $\xi^* = 0$, then stop with an optimal solution u^* for (P) such that $u_J^* = x_J^*$ and the other components of u^* are zero.

Else let

$$z^* = -\mathbf{1}_m^\top - (\overline{c}_{K^*})_{(p+1,\ldots,p+m)},$$

be the optimal solution of (DRP) corresponding to (x_J^*, ξ^*) and the basis K^*.

If $z^*A^j \geq 0$ for all $j \notin J$, then stop; the Program (P) has no feasible solution.

Else compute

$$\theta^+ = \min\left\{ -\frac{yA^j - c_j}{z^*A^j} \;\middle|\; j \notin J,\ z^*A^j < 0 \right\}, \qquad y^+ = y + \theta^+ z^*,$$

and

$$J^+ = \{j \in \{1, \ldots, n\} \mid y^+A^j = c_j\}.$$

Go back to Step 3.

The following proposition shows that at each iteration we can start the Program (RP) with the optimal solution obtained at the previous iteration.

Proposition 11.14. *Every $j \in J$ such that A^j is in the basis of the optimal solution ξ^* belongs to the next index set J^+.*

Proof. Such an index $j \in J$ correspond to a variable ξ_j such that $\xi_j > 0$, so by complementary slackness, the constraint $z^*A^j \geq 0$ of the Dual Program (DRP) must be an equality, that is, $z^*A^j = 0$. But then we have

$$y^+A^j = yA^j + \theta^+ z^*A^j = c_j,$$

which shows that $j \in J^+$. □

If (u^*, ξ^*) with the basis K^* is the optimal solution of the Program (RP), Proposition 11.14 together with the last property of Theorem 11.12 allows us to restart the (RP) in Step 3 with $(u^*, \xi^*)_{K^*}$ as initial solution (with basis K^*). For every $j \in J - J^+$, column j is deleted, and for every $j \in J^+ - J$, the new column A^j is computed by multiplying $\widehat{A}_{K^*}^{-1}$ and A^j, but $\widehat{A}_{K^*}^{-1}$ is the matrix $\Gamma^*[1{:}m; p+1{:}p+m]$ consisting of the last m columns of Γ^* in the final tableau, and the new reduced $\overline{c}_j$ is given by $c_j - z^*A^j$. Reusing the optimal solution of the previous (RP) may improve efficiency significantly.

Another crucial observation is that for any index $j_0 \in N$ such that $\theta^+ = (yA^{j_0} - c_{j_0})/(-z^*A^{j_0})$, we have

$$y^+A_{j_0} = yA_{j_0} + \theta^+ z^*A^{j_0} = c_{j_0},$$

and so $j_0 \in J^+$. This fact that be used to ensure that the primal-dual algorithm terminates in a finite number of steps (using a pivot rule that prevents cycling); see Papadimitriou and Steiglitz [60] (Theorem 5.4).

It remains to discuss how to pick some initial feasible solution y of the Dual Program (D). If $c_j \leq 0$ for $j = 1, \ldots, n$, then we can pick $y = 0$. If we are dealing with a minimization problem, the weight c_j are often nonnegative, so from the point of view of maximization we will have $-c_j \leq 0$ for all j, and we will be able to use $y = 0$ as a starting point.

Going back to our primal problem in maximization form and its dual in minimization form, we still need to deal with the situation where $c_j > 0$ for some j, in which case there

may not be any obvious y feasible for (D). Preferably we would like to find such a y very cheaply.

There is a trick to deal with this situation. We pick some very large positive number M and add to the set of equations $Ax = b$ the new equation

$$x_1 + \cdots + x_n + x_{n+1} = M,$$

with the new variable x_{n+1} constrained to be nonnegative. If the Program (P) has a feasible solution, such an M exists. In fact it can shown that for any basic feasible solution $u = (u_1, \ldots, u_n)$, each $|u_i|$ is bounded by some expression depending only on A and b; see Papadimitriou and Steiglitz [60] (Lemma 2.1). The proof is not difficult and relies on the fact that the inverse of a matrix can be expressed in terms of certain determinants (the adjugates). Unfortunately, this bound contains $m!$ as a factor, which makes it quite impractical.

Having added the new equation above, we obtain the new set of equations

$$\begin{pmatrix} A & 0_n \\ \mathbf{1}_n^\top & 1 \end{pmatrix} \begin{pmatrix} x \\ x_{n+1} \end{pmatrix} = \begin{pmatrix} b \\ M \end{pmatrix},$$

with $x \geq 0, x_{n+1} \geq 0$, and the new objective function given by

$$\begin{pmatrix} c & 0 \end{pmatrix} \begin{pmatrix} x \\ x_{n+1} \end{pmatrix} = cx.$$

The dual of the above linear program is

$$\begin{array}{ll} \text{minimize} & yb + y_{m+1}M \\ \text{subject to} & yA^j + y_{m+1} \geq c_j \quad j = 1, \ldots, n \\ & y_{m+1} \geq 0. \end{array}$$

If $c_j > 0$ for some j, observe that the linear form $\widetilde{y}$ given by

$$\widetilde{y}_i = \begin{cases} 0 & \text{if } 1 \leq i \leq m \\ \max_{1 \leq j \leq n}\{c_j\} > 0 & \end{cases}$$

is a feasible solution of the new dual program. In practice, we can choose M to be a number close to the largest integer representable on the computer being used.

Here is an example of the primal-dual algorithm given in the Math 588 class notes of T. Molla [57].

Example 11.3. Consider the following linear program in standard form:

$$\begin{array}{ll} \text{Maximize} & -x_1 - 3x_2 - 3x_3 - x_4 \\ \text{subject to} & \begin{pmatrix} 3 & 4 & -3 & 1 \\ 3 & -2 & 6 & -1 \\ 6 & 4 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \text{ and } x_1, x_2, x_3, x_4 \geq 0. \end{array}$$

The associated Dual Program (D) is

$$\begin{aligned} &\text{Minimize} && 2y_1 + y_2 + 4y_3 \\ &\text{subject to} && \begin{pmatrix} y_1 & y_2 & y_3 \end{pmatrix} \begin{pmatrix} 3 & 4 & -3 & 1 \\ 3 & -2 & 6 & -1 \\ 6 & 4 & 0 & 1 \end{pmatrix} \geq \begin{pmatrix} -1 & -3 & -3 & -1 \end{pmatrix}. \end{aligned}$$

We initialize the primal-dual algorithm with the dual feasible point $y = (-1/3 \;\; 0 \;\; 0)$. Observe that only the first inequality of (D) is actually an equality, and hence $J = \{1\}$. We form the Restricted Primal Program $(RP1)$

$$\begin{aligned} &\text{Maximize} && -(\xi_1 + \xi_2 + \xi_3) \\ &\text{subject to} && \begin{pmatrix} 3 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 \\ 6 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \text{ and } x_1, \xi_1, \xi_2, \xi_3 \geq 0. \end{aligned}$$

We now solve $(RP1)$ via the simplex algorithm. The initial tableau with $K = (2, 3, 4)$ and $J = \{1\}$ is

	x_1	ξ_1	ξ_2	ξ_3
7	12	0	0	0
$\xi_1 = 2$	3	1	0	0
$\xi_2 = 1$	(3)	0	1	0
$\xi_3 = 4$	6	0	0	1

.

For $(RP1)$, $\hat{c} = (0, -1, -1, -1)$, $(x_1, \xi_1, \xi_2, \xi_3) = (0, 2, 1, 4)$, and the nonzero reduced cost is given by

$$0 - \begin{pmatrix} -1 & -1 & -1 \end{pmatrix} \begin{pmatrix} 3 \\ 3 \\ 6 \end{pmatrix} = 12.$$

Since there is only one nonzero reduced cost, we must set $j^+ = 1$. Since $\min\{\xi_1/3, \xi_2/3, \xi_3/6\} = 1/3$, we see that $k^- = 3$ and $K = (2, 1, 4)$. Hence we pivot through the red circled 3 (namely we divide row 2 by 3, and then subtract $3\times$ (row 2) from row 1, $6\times$ (row 2) from row 3, and $12\times$ (row 2) from row 0), to obtain the tableau

	x_1	ξ_1	ξ_2	ξ_3
3	0	0	-4	0
$\xi_1 = 1$	0	1	-1	0
$x_1 = 1/3$	1	0	$1/3$	0
$\xi_3 = 2$	0	0	-2	1

.

At this stage the simplex algorithm for $(RP1)$ terminates since there are no positive reduced costs. Since the upper left corner of the final tableau is not zero, we proceed with Step 4 of the primal dual algorithm and compute

$$z^* = \begin{pmatrix} -1 & -1 & -1 \end{pmatrix} - \begin{pmatrix} 0 & -4 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 3 & -1 \end{pmatrix},$$

$$yA^2 - c_2 = (-1/3 \;\; 0 \;\; 0)\begin{pmatrix} 4 \\ -2 \\ 4 \end{pmatrix} + 3 = \frac{5}{3}, \qquad z^*A^2 = -(-1 \;\; 3 \;\; -1)\begin{pmatrix} 4 \\ -2 \\ 4 \end{pmatrix} = 14,$$

$$yA^4 - c_4 = (-1/3 \;\; 0 \;\; 0)\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} + 1 = \frac{2}{3}, \qquad z^*A^4 = -(-1 \;\; 3 \;\; -1)\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = 5,$$

so

$$\theta^+ = \min\left\{ \frac{5}{42}, \frac{2}{15} \right\} = \frac{5}{42},$$

and we conclude that the new feasible solution for (D) is

$$y^+ = (-1/3 \;\; 0 \;\; 0) + \frac{5}{42}(-1 \;\; 3 \;\; -1) = (-19/42 \;\; 5/14 \;\; -5/42).$$

When we substitute y^+ into (D), we discover that the first two constraints are equalities, and that the new J is $J = \{1, 2\}$. The new Reduced Primal $(RP2)$ is

$$\begin{aligned} &\text{Maximize} \quad -(\xi_1 + \xi_2 + \xi_3) \\ &\text{subject to} \quad \begin{pmatrix} 3 & 4 & 1 & 0 & 0 \\ 3 & -2 & 0 & 1 & 0 \\ 6 & 4 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \quad \text{and} \quad x_1, x_2, \xi_1, \xi_2, \xi_3 \geq 0. \end{aligned}$$

Once again, we solve $(RP2)$ via the simplex algorithm, where $\hat{c} = (0, 0, -1, -1, -1)$, $(x_1, x_2, \xi_1, \xi_2, \xi_3) = (1/3, 0, 1, 0, 2)$ and $K = (3, 1, 5)$. The initial tableau is obtained from the final tableau of the previous $(RP1)$ by adding a column corresponding the the variable x_2, namely

$$\widehat{A}_K^{-1}A^2 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1/3 & 0 \\ 0 & -2 & 1 \end{pmatrix} \begin{pmatrix} 4 \\ -2 \\ 4 \end{pmatrix} = \begin{pmatrix} 6 \\ -2/3 \\ 8 \end{pmatrix},$$

with

$$\bar{c}_2 = c_2 - z^*A^2 = 0 - \begin{pmatrix} -1 & 3 & -1 \end{pmatrix} \begin{pmatrix} 4 \\ -2 \\ 4 \end{pmatrix} = 14,$$

and we get

	x_1	x_2	ξ_1	ξ_2	ξ_3
3	0	14	0	-4	0
$\xi_1 = 1$	0	(6)	1	-1	0
$x_1 = 1/3$	1	$-2/3$	0	$1/3$	0
$\xi_3 = 2$	0	8	0	-2	1

.

Note that $j^+ = 2$ since the only positive reduced cost occurs in column 2. Also observe that since $\min\{\xi_1/6, \xi_3/8\} = \xi_1/6 = 1/6$, we set $k^- = 3$, $K = (2, 1, 5)$ and pivot along the red 6 to obtain the tableau

	x_1	x_2	ξ_1	ξ_2	ξ_3
2/3	0	0	$-7/3$	$-5/3$	0
$x_2 = 1/6$	0	1	1/6	$-1/6$	0
$x_1 = 4/9$	1	0	1/9	2/9	0
$\xi_3 = 2/3$	0	0	$-4/3$	$-2/3$	1

.

Since the reduced costs are either zero or negative the simplex algorithm terminates, and we compute

$$z^* = (-1 \ \ -1 \ \ -1) - (-7/3 \ \ -5/3 \ \ 0) = (4/3 \ \ 2/3 \ \ -1),$$

$$y^+ A^4 - c_4 = (-19/42 \ \ 5/14 \ \ -5/42) \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} + 1 = 1/14,$$

$$z^* A^4 = -(4/3 \ \ 2/3 \ \ -1) \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = 1/3,$$

so

$$\theta^+ = \frac{3}{14},$$

$$y^+ = (-19/42 \ \ 5/14 \ \ -5/42) + \frac{5}{14}(4/3 \ \ 2/3 \ \ -1) = (-1/6 \ \ 1/2 \ \ -1/3).$$

When we plug y^+ into (D), we discover that the first, second, and fourth constraints are equalities, which implies $J = \{1, 2, 4\}$. Hence the new Restricted Primal $(RP3)$ is

$$\text{Maximize} \quad -(\xi_1 + \xi_2 + \xi_3)$$

$$\text{subject to} \quad \begin{pmatrix} 3 & 4 & 1 & 1 & 0 & 0 \\ 3 & -2 & -1 & 0 & 1 & 0 \\ 6 & 4 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_4 \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \quad \text{and} \quad x_1, x_2, x_4, \xi_1, \xi_2, \xi_3 \geq 0.$$

The initial tableau for $(RP3)$, with $\hat{c} = (0, 0, 0, -1, -1, -1)$, $(x_1, x_2, x_4, \xi_1, \xi_2, \xi_3) = (4/9, 1/6, 0, 0, 0, 2/3)$ and $K = (2, 1, 6)$, is obtained from the final tableau of the previous $(RP2)$ by adding a column corresponding the the variable x_4, namely

$$\widehat{A}_K^{-1} A^4 = \begin{pmatrix} 1/6 & -1/6 & 0 \\ 1/9 & 2/9 & 0 \\ -4/3 & -2/3 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/3 \\ -1/9 \\ 1/3 \end{pmatrix},$$

with

$$\bar{c}_4 = c_4 - z^* A^4 = 0 - \begin{pmatrix} 4/3 & 2/3 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = 1/3,$$

and we get

	x_1	x_2	x_4	ξ_1	ξ_2	ξ_3
$2/3$	0	0	$1/3$	$-7/3$	$-5/3$	0
$x_2 = 1/6$	0	1	$1/3$	$1/6$	$-1/6$	0
$x_1 = 4/9$	1	0	$-1/9$	$1/9$	$2/9$	0
$\xi_3 = 2/3$	0	0	$1/3$	$-4/3$	$-2/3$	1

.

Since the only positive reduced cost occurs in column 3, we set $j^+ = 3$. Furthermore since $\min\{x_2/(1/3), \xi_3/(1/3)\} = x_2/(1/3) = 1/2$, we let $k^- = 2$, $K = (3, 1, 6)$, and pivot around the red circled $1/3$ to obtain

	x_1	x_2	x_4	ξ_1	ξ_2	ξ_3
$1/2$	0	-1	0	$-5/2$	$-3/2$	0
$x_4 = 1/2$	0	3	1	$1/2$	$-1/2$	0
$x_1 = 1/2$	1	$1/3$	0	$1/6$	$1/6$	0
$\xi_3 = 1/2$	0	-1	0	$-3/2$	$-1/2$	1

.

At this stage there are no positive reduced costs, and we must compute

$$z^* = (-1 \;\; -1 \;\; -1) - (-5/2 \;\; -3/2 \;\; 0) = (3/2 \;\; 1/2 \;\; -1),$$

$$y^+ A^3 - c_3 = (-1/6 \;\; 1/2 \;\; -1/3) \begin{pmatrix} -3 \\ 6 \\ 0 \end{pmatrix} + 3 = 13/2,$$

$$z^* A^3 = -(3/2 \;\; 1/2 \;\; -1) \begin{pmatrix} -3 \\ 6 \\ 0 \end{pmatrix} = 3/2,$$

so

$$\theta^+ = \frac{13}{3},$$
$$y^+ = (-1/6 \;\; 1/2 \;\; -1/3) + \frac{13}{3}(3/2 \;\; 1/2 \;\; -1) = (19/3 \;\; 8/3 \;\; -14/3).$$

We plug y^+ into (D) and discover that the first, third, and fourth constraints are equalities. Thus, $J = \{1, 3, 4\}$ and the Restricted Primal $(RP4)$ is

$$\text{Maximize} \quad -(\xi_1 + \xi_2 + \xi_3)$$

$$\text{subject to} \quad \begin{pmatrix} 3 & -3 & 1 & 1 & 0 & 0 \\ 3 & 6 & -1 & 0 & 1 & 0 \\ 6 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_3 \\ x_4 \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix} \quad \text{and} \quad x_1, x_3, x_4, \xi_1, \xi_2, \xi_3 \geq 0.$$

The initial tableau for $(RP4)$, with $\hat{c} = (0, 0, 0, -1, -1, -1)$, $(x_1, x_3, x_4, \xi_1, \xi_2, \xi_3) = (1/2, 0, 1/2, 0, 0, 1/2)$ and $K = (3, 1, 6)$ is obtained from the final tableau of the previous $(RP3)$ by replacing the column corresponding to the variable x_2 by a column corresponding to the variable x_3, namely

$$\widehat{A}_K^{-1} A^3 = \begin{pmatrix} 1/2 & -1/2 & 0 \\ 1/6 & 1/6 & 0 \\ -3/2 & -1/2 & 1 \end{pmatrix} \begin{pmatrix} -3 \\ 6 \\ 0 \end{pmatrix} = \begin{pmatrix} -9/2 \\ 1/2 \\ 3/2 \end{pmatrix},$$

with

$$\overline{c}_3 = c_3 - z^* A^3 = 0 - \begin{pmatrix} 3/2 & 1/2 & -1 \end{pmatrix} \begin{pmatrix} -3 \\ 6 \\ 0 \end{pmatrix} = 3/2,$$

and we get

	x_1	x_3	x_4	ξ_1	ξ_2	ξ_3
$1/2$	0	$3/2$	0	$-5/2$	$-3/2$	0
$x_4 = 1/2$	0	$-9/2$	1	$1/2$	$-1/2$	0
$x_1 = 1/2$	1	$1/2$	0	$1/6$	$1/6$	0
$\xi_3 = 1/2$	0	$3/2$	0	$-3/2$	$-1/2$	1

.

By analyzing the top row of reduced cost, we see that $j^+ = 2$. Furthermore, since $\min\{x_1/(1/2), \xi_3/(3/2)\} = \xi_3/(3/2) = 1/3$, we let $k^- = 6$, $K = (3, 1, 2)$, and pivot along the red circled $3/2$ to obtain

	x_1	x_3	x_4	ξ_1	ξ_2	ξ_3
0	0	0	0	-1	-1	-1
$x_4 = 2$	0	0	1	-4	-2	3
$x_1 = 1/3$	1	0	0	$2/3$	$1/3$	$-1/3$
$x_3 = 1/3$	0	1	0	-1	$-1/3$	$2/3$

.

Since the upper left corner of the final tableau is zero and the reduced costs are all ≤ 0, we are finally finished. Then $y = (19/3 \;\; 8/3 \;\; -14/3)$ is an optimal solution of (D), but more

importantly $(x_1, x_2, x_3, x_4) = (1/3, 0, 1/3, 2)$ is an optimal solution for our original linear program and provides an optimal value of $-10/3$.

The primal-dual algorithm for linear programming doesn't seem to be the favorite method to solve linear programs nowadays. But it is important because its basic principle, to use a restricted (simpler) primal problem involving an objective function with fixed weights, namely 1, and the dual problem to provide feedback to the primal by improving the objective function of the dual, has led to a whole class of combinatorial algorithms (often approximation algorithms) based on the primal-dual paradigm. The reader will get a taste of this kind of algorithm by consulting Papadimitriou and Steiglitz [60], where it is explained how classical algorithms such as Dijkstra's algorithm for the shortest path problem, and Ford and Fulkerson's algorithm for max flow can be derived from the primal-dual paradigm.

11.7 Summary

The main concepts and results of this chapter are listed below:

- Strictly separating hyperplane.
- Farkas–Minkowski proposition.
- Farkas lemma, version I, Farkas lemma, version II.
- Distance of a point to a subset.
- Dual linear program, primal linear program.
- Dual variables, primal variables.
- Complementary slackness conditions.
- Dual simplex algorithm.
- Primal-dual algorithm.
- Restricted primal linear program.

11.8 Problems

Problem 11.1. Let $(v_1, \ldots, v_n)$ be a sequence of n vectors in $\mathbb{R}^d$ and let V be the $d \times n$ matrix whose j-th column is v_j. Prove the equivalence of the following two statements:

(a) There is no nontrivial positive linear dependence among the v_j, which means that there is no nonzero vector, $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, with $y_j \geq 0$ for $j = 1, \ldots, n$, so that

$$y_1 v_1 + \cdots + y_n v_n = 0$$

or equivalently, $Vy = 0$.

(b) There is some vector, $c \in \mathbb{R}^d$, so that $c^\top V > 0$, which means that $c^\top v_j > 0$, for $j = 1, \ldots, n$.

Problem 11.2. Check that the dual in maximization form (D'') of the Dual Program (D') (which is the dual of (P) in maximization form),

$$\begin{aligned} &\text{maximize} && -b^\top y^\top \\ &\text{subject to} && -A^\top y^\top \leq -c^\top \text{ and } y^\top \geq 0, \end{aligned}$$

where $y \in (\mathbb{R}^m)^*$, gives back the Primal Program (P).

Problem 11.3. In a General Linear Program (P) with n primal variables $x_1, \ldots, x_n$ and objective function $\sum_{j=1}^n c_j x_j$ (to be maximized), the m constraints are of the form

$$\begin{aligned} \sum_{j=1}^n a_{ij} x_j &\leq b_i, \\ \sum_{j=1}^n a_{ij} x_j &\geq b_i, \\ \sum_{j=1}^n a_{ij} x_j &= b_i, \end{aligned}$$

for $i = 1, \ldots, m$, and the variables x_j satisfy an inequality of the form

$$\begin{aligned} x_j &\geq 0, \\ x_j &\leq 0, \\ x_j &\in \mathbb{R}, \end{aligned}$$

for $j = 1, \ldots, n$. If $y_1, \ldots, y_m$ are the dual variables, show that the dual program of the linear program in standard form equivalent to (P) is equivalent to the linear program whose objective function is $\sum_{i=1}^m y_i b_i$ (to be minimized) and whose constraints are determined as follows:

$$\text{if} \quad \left\{ \begin{matrix} x_j \geq 0 \\ x_j \leq 0 \\ x_j \in \mathbb{R} \end{matrix} \right\}, \quad \text{then} \quad \left\{ \begin{matrix} \sum_{i=1}^m a_{ij} y_i \geq c_j \\ \sum_{i=1}^m a_{ij} y_i \leq c_j \\ \sum_{i=1}^m a_{ij} y_i = c_j \end{matrix} \right\},$$

and

$$\text{if} \quad \left\{ \begin{array}{l} \displaystyle\sum_{j=1}^{n} a_{ij}x_j \leq b_i \\ \displaystyle\sum_{j=1}^{n} a_{ij}x_j \geq b_i \\ \displaystyle\sum_{j=1}^{n} a_{ij}x_j = b_i \end{array} \right\}, \quad \text{then} \quad \left\{ \begin{array}{l} y_i \geq 0 \\ y_i \leq 0 \\ y_i \in \mathbb{R} \end{array} \right\}.$$

Problem 11.4. Apply the procedure of Problem 11.3 to show that the dual of the (general) linear program

$$\begin{array}{ll} \text{maximize} & 3x_1 + 2x_2 + 5x_3 \\ \text{subject to} & \\ & 5x_1 + 3x_2 + x_3 = -8 \\ & 4x_1 + 2x_2 + 8x_3 \leq 23 \\ & 6x_1 + 7x_2 + 3x_3 \geq 1 \\ & x_1 \leq 4,\ x_3 \geq 0 \end{array}$$

is the (general) linear program:

$$\begin{array}{ll} \text{minimize} & -8y_1 + 23y_2 - y_3 + 4y_4 \\ \text{subject to} & \\ & 5y_1 + 4y_2 - 6y_3 + y_4 = 3 \\ & 3y_1 + 2y_2 - 7y_3 = 2 \\ & y_1 + 8y_2 - 3y_3 \geq 5 \\ & y_2,\ y_3,\ y_4 \geq 0. \end{array}$$

Problem 11.5. (1) Prove that the dual of the (general) linear program

$$\begin{array}{ll} \text{maximize} & cx \\ \text{subject to} & Ax = b \text{ and } x \in \mathbb{R}^n \end{array}$$

is

$$\begin{array}{ll} \text{minimize} & yb \\ \text{subject to} & yA = c \text{ and } y \in \mathbb{R}^m. \end{array}$$

(2) Prove that the dual of the (general) linear program

$$\begin{array}{ll} \text{maximize} & cx \\ \text{subject to} & Ax \geq b \text{ and } x \geq 0 \end{array}$$