Analysis, Convexity, and Optimization

Henry C. Pinkham

Henry C. Pinkham

Draft of September 4, 2014.

Contents

Preface

These notes are designed for a course on the mathematics associated to static optimization theory. More specifically, the book serves as an introduction to those concepts in linear algebra, analysis and convexity that are most important in static optimization. We then go on to optimization itself, focusing on examples from economics.

Prerequisites

The prerequisites for reading these lectures are given below. Most of the concepts from the prerequisite courses—especially those central to optimization theory—are reviewed at the beginning of the book.

- A one-semester course in single-variable calculus. The book does not use integrals, but differentiation, and infinite sequences are fundamental. Infinite series only make an appearance via Taylor series. See Lectures 3 and 4,

- A one-semester course in linear algebra. Lectures 5 and 6, and lectures 7, 8, and 9.

- A one-semester course in multivariable calculus. The book assumes the ability to differentiate a function f of several variables, to apply the chain rule in several variables, and to compute the second-order Taylor polynomial. Students should already understand both the graph and the level varieties of the function. Some of the lectures on linear algebra also contains multivariable calculus, of course. The sections devoted to results usually covered in a multivariable calculus course are Lectures 5, 10, 11 and 12.

For the material that is not reviewed, references are given to the elementary texts that are most familiar to me: Stewart [63] for calculus, and Spence-Insel-Friedberg [60] and Strang [68] for linear algebra. References are also given to more specialized books that are roughly at the same level as this one, but with more detail in the areas that they cover. My favorites are:

- Rudin [55] and Strichartz[70] for analysis;
- Serre [57] and Strang [67] for linear algebra;
- Roberts-Varburg [52] and Barvinok [4] for convexity; and
- Franklin [23], Berkovitz [7] and Boyd-Vandenberghe [10] for optimization.

Proofs

Another important skill is the ability to read and master mathematical proofs. Some courses are specifically designed to do that, using a text such as Dumas-McCarthy [21] for this purpose, but most training is acquired through courses such as this one. Although prior experience with proofs is helpful, these lectures are designed to accommodate the readers with no background in reading and writing proofs, and to help them cultivate these fundamental skills.

Here is how I suggest you read the proof of a theorem:

- You should make sure you understand what each individual statement in the proof means, and how each one follows logically from the previous one.
- In some instances you will have to go back to the definitions and prior theorems used in the proof to complete the previous step. Read through them carefully.
- Make sure you understand how each one of the hypotheses are used. For most (but not all) theorems, the result is known to be false when a hypothesis is dropped. Make sure you see why that is true. Construct your own simple examples whenever possible
- See how the theorem is used in the rest of the notes.

There are several kinds of proofs that are used repeatedly in these lectures:

- Proof by induction.
- Proof by contradiction.
- Proving the contrapositive statement. If the theorem claims that statement P implies statement Q, you may prove instead the equivalent statement that $notQ$ implies $notP$. Notice how closely related this is to proof by contradiction.

- Proof by successive approximation. This is referred to as *divide and conquer* in [70] p. 53. This is especially important in these lectures. Instead of attempting to prove the result in one step, one constructs a sequence of better and better approximations to the result, and then shows that the limit of the sequence gives the desired result. The prototypical result is Newton's method for computing the root x^* of a real-valued differentiable function that you may have seen in calculus. See for example Stewart [63], §4.8.

It is also essential to understand the difference between a necessary and a sufficient condition for a result to be true. Here is an example of a necessary condition that is not sufficient: in the era before calculators, all school children used to know this algorithm.

Example: Proof by nine Suppose you are given two positive integers a and b in their decimal representation, say $a = 215738$ and $b = 61922$. Multiply them, and call the answer c. My calculator tells me that the answer is $c = 13358928436$. How to check this answer if one has done the computation by hand? Add the digits of a (in our example we get 26), add the digits again (get 8 in the example), and when there is only one digit left, stop. Do the same for b (in our example, you get 20 and then 2) and for c (get 52 and then 7). The proof by nine tells us that the product of the numbers obtained for a and b must be equal (after reducing the digits once more if necessary) to the number obtained for c. Indeed in our example $8 \cdot 2 = 16$, add the digits to get 7, the same result as for c. The equality of these two numbers is a necessary condition for the multiplication to be correct, but it is not a sufficient condition.

What does this have to do with the number 9? For each one of the three numbers we are computing the *equivalence class* modulo 9 of the number: this is a way of saying we are computing the remainder of the number after division by 9. This is an equivalence relation (see Definition 15.5.2) on integers, and an easy result says that if we represent the remainder classes by $\overline{0}$, $\overline{1}$, $\overline{2}$, ..., $\overline{8}$, then $\overline{a} \cdot \overline{b} = \overline{c}$. The last ingredient of the algorithm is that to compute the remainder of a number under division by 9, you just repeatedly add its digits until you get an integer no larger that 8. You should verify this on your own. The point is simply that the remainder class of a power of 10 modulo 9 is 1.

Note that getting the correct value for the product says that the probability of the computation being correct is $8/9$: in other words there is still nearly a 9% chance that the computation is incorrect. This is just another way of quantifying the fact that the condition is necessary but not sufficient.

Keep this example in mind when you think about necessary and sufficient conditions.

Course Objectives

In mathematical terms, to solve a static optimization problem is to find the maxima and minima of a single real-valued function $f(\mathbf{x})$ of several variables. The first goal of these lectures is to teach those parts of linear algebra, real analysis, and convexity theory that are most relevant to solving such problems. Two key concepts that are not part of the standard math syllabus will emerge: 1) positivity (the variables in applied mathematics are often constrained to be positive), and 2) convexity/concavity of functions. These issues are discussed at length in Lectures 5 through 19.

The book's second purpose is to study the principal results of static optimization theory, using examples from economics. This broad topic is the focus of Lectures 25 through 34. While these lectures are not as proof-oriented as, for example, an introduction to real analysis (e.g., Rudin [55]), the reader is encouraged to master the most important proofs as well as the statements of all the theorems. Through plenty of exercises, the lectures encourage the reader to work out concrete applications of the main results.

More generally, the goal of these notes is to enable the reader to determine the structure of the solution, rather than merely to compute numbers. To quote Richard Bellman (in the introduction to [6], pg. 7),

> Concepts are more important than constants.

This is not to say that computation is unimportant. On the contrary, the field of optimization flourished when computers enabled computations in high dimension. Subsequent editions of these notes will introduce readers to the technology associated with these higher-level computations—namely, symbolic programming in Matlab, Python, and Sage. The mathematical background of these computations will also be discussed.

Comments, corrections, and other suggestions for improving these notes are welcome. Please email them to me at henrypinkham@gmail.com.

HENRY C. PINKHAM
New York, NY
September 4, 2014

Part I

Introduction

Lecture 1

Some Examples

This lecture introduces the key definitions and concepts for optimization and then covers three applied examples that illustrate what comes later: first, two key linear optimization problems: the Diet Problem §1.3 and the Transportation Problem §1.4, and then a convex optimization problem §1.5.

1.1 The Problem

We begin by describing the central problem studied throughout this book. We work with a continuous function $f(x_1, \dots, x_n)$ whose domain contains a region D in $\mathbb{R}^n$. f is a real-valued function, meaning that its range is contained in the real numbers $\mathbb{R}$. While the number n of real inputs can be very large—thousands in some real-life applications—f outputs a single real number.

1.1.1 Definition. This function f, our central concern, is called the *objective function*.[1]

Our task is to *optimize* f over the region D: this means to either maximize or minimize it (or both), depending on the problem at hand.

1.1.2 Definition. The region D plays a central role in our study, and it too has a name: the *feasible set*, or the *constraint set*.

D must be contained in the domain of f (the set where f is defined), but it can be smaller. In economics, for example, problems often make sense only for non-negative values of the variables. Say that f represents the cost of a bundle of goods, and you can't—as is usually the case—buy negative quantities. It is sensible, therefore, to restrict f to the positive quadrant in $\mathbb{R}^n$.

[1] In some minimization problems, f is called the *cost function*.

Most of our examples, like this one, come from economics. We study examples from economics for each of the optimization problems treated in this book. Those problems, defined by the choices for the objective function and the feasible set, can be summarized as follows:

- In Lecture 3, we review the one dimensional case, familiar to you from the max-min problems of single variable calculus.

- In Lecture 7, §7.5, we show that orthogonal projection minimizes the distance of a point to a linear space. We generalize this to the distance of a point to a convex set in Corollary 18.6.4. We apply this technique to least squares in §13.3.

- In Lecture 13, we study the case where the feasible set D is all of $\mathbb{R}^n$. This is *unconstrained optimization*. Most of this material will be familiar to you from multi-variable calculus.

- Lecture 25 considers the case where f is a linear function and the feasible set is constrained by linear equalities and inequalities. This case is known as *linear programming* or linear optimization. Another reference for this material is [23]. Applications are given in Lecture 26.

- In Lecture 28, we use Lagrange Multipliers (28.3.9) to generalize the results of equality-constrained linear optimization to a nonlinear context. While the constraints are all equalities, there is no linearity requirement for either the constaints or the objective function. Most multivariable-calculus texts offer preliminary introductions to this problem (see, e.g., [63], §14).

- Lecture 30 presents *quadratic optimization*, another special case in which the objective function is quadratic and the constraint functions (either equality or inequality) are affine. Other references are [23], II.1 and [20]

- As discussed in Lectures 22, 23, and 33, assuming that the objective function and the constraints are convex also has special consequences. This case therefore also deserves its own name: *convex optimization*. The background material for convex optimization is found in Lecture 18 on convex sets and Lecture 21 on convex functions. Excellent references are [7] and [10], books that are more advanced than these lectures. I recommend [52] as an even more detailed alternative.

- Lectures 31 and 32, finally, discusses the most general nonlinear case, which we call *nonlinear optimization*. This case imposes no limits on the character or quality of either the objective function or the constraints. Many of the books already mentioned cover this case. Additional references include [5], [18], [43], [44], [45] and [71].

You may be curious why some of the books cited here use the word *programming* rather than *optimization* in their title. This terminology reflects the fact that, in most situations, it is not possible to find a closed-form expression for the solution of an optimization problem. Instead, one constructs an iterative algorithm that converges to the solution—in other words, one writes a computer *program* to solve the problem. You may have seen instances of this in calculus—the Newton-Raphson method, for example (see [63], §4.9).

1.1.3 Definition.

- To minimize $f(\mathbf{x})$ on D, we look for values $\mathbf{a} \in D$ where the value $f(\mathbf{a})$ is a minimum, formally:

  $$f(\mathbf{a}) \leq f(\mathbf{x}), \text{ for all } \mathbf{x} \in D. \tag{1.1.4}$$

  A value $\mathbf{a} \in D$ where (1.1.4) holds is called a *minimizer*. If $\mathbf{a}$ is a minimizer, then we say that the value $f(a)$ is a *minimum* for f.

- To maximize $f(\mathbf{x})$ on D, we look for values $\mathbf{b} \in D$ where the value $f(\mathbf{b})$ is maximum, in other words:

  $$f(\mathbf{b}) \geq f(\mathbf{x}), \text{ for all } \mathbf{x} \in D. \tag{1.1.5}$$

  A value $\mathbf{b} \in D$ where (1.1.5) holds is called a *maximizer*. Correspondingly, if $\mathbf{b}$ is a maximizer, then we say that the value $f(b)$ is a *maximum*.

When we are agnostic between a maximum or a minimum, we use the term *extremum* (pl. extrema).

1.1.6 Proposition (Minimizing and maximizing)**.** *Suppose we want to* maximize *the objective function f on the feasible set D. This is equivalent to* minimizing *the function $-f$ on the feasible set D.*

Proof. For a vector $\mathbf{a} \in D$ to be a maximizer for the function f on D means that (1.1.5) holds (substituting $\mathbf{a}$ for $\mathbf{b}$). Multiplying the inequality by -1 reverses the direction of the inequality, giving (1.1.4). Thus, $\mathbf{a}$ is a minimizer for $-f$. □

1.1.7 Remark. Proposition 1.1.6 shows that we really only need to consider either minimization or maximization problems, and not both. These lectures focus on minimization.

1.1.8 Exercise. Keep the same objective function f, but change the feasible set D to a larger feasible set D^*: $D \subset D^*$. If $\mathbf{a}$ is a minimizer for f on D, and $\mathbf{a}^*$ is a minimizer for f on D^*, show that $f(\mathbf{a}^*) \leq f(\mathbf{a})$. Also show that if $f(\mathbf{a}^*) < f(\mathbf{a})$, then $\mathbf{a}^* \notin D$.

A minimization problem admits a solution only if a minimizer exists. However it might not. Three things could go wrong:

1. The feasible set could be empty.

2. The values of f could get arbitrarily negative on D, meaning that they approach $-\infty$.

3. Even if the values of f do not tend to $-\infty$, a minimizer might still elude us if, as explained in §16.2, D is not a compact set. This property—defined in §15.1—play an important role in Weierstrass's Theorem 16.2.2, which is a sufficient condition on D for the existence of a minimizer $\mathbf{a}$ and a minimum.

We consider two types of minima: the one in Definition 1.1.3 is known as a a *global minimizer*[2], because the function's value there is less than or equal to than its value at any other point on the domain. We are also interested in local minimizers, which can be defined as follows:

1.1.9 Definition. A point $\mathbf{a} \in D$ is a *local minimizer* for f if there exists a small enough neighborhood U of $\mathbf{a}$ in D such that

$$f(\mathbf{a}) \leq f(\mathbf{x}), \text{ for all } \mathbf{x} \in U.$$

Then $f(\mathbf{a})$ is a *local minimum.*

The notion of *neighborhood* will be made mathematically precise in Definition 14.4.1; for now, just think of it as a small region around a point. In $\mathbb{R}$, a neighborhood of a point a is an interval of the form $(a - \epsilon, a + \epsilon)$, for any positive number ϵ.

A global minimum is always a local minimum, but a local minimum need not be a global minimum. Optimization often uses differential calculus to find minima, but this tool set only helps us find local minima. We are more interested, meanwhile, in global minima—or reasonable approximations of global minima. In the case of linear optimization (Lecture 25), we can always find the global minimum via the simplex method (see Lecture 27). When the objective function is convex, convex optimization usually produces the global minimum (Lecture 23). In less-friendly optimization problems, we sometimes have to resort to approximation methods using computer programs. This will not be discussed in this book.

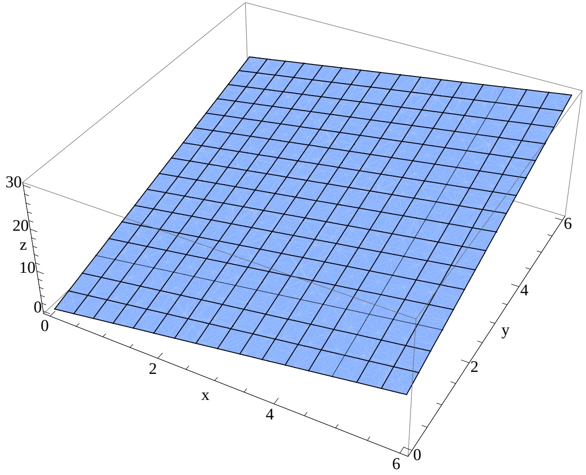
[2] The term absolute minimizer is the term used instead, for example, in [63], §14.7.

1.1.10 Exercise. Find an objective function of one variable $f(x)$ and a feasible set D such that f has a global minimum on D; then find an f and D where there is a local minimum that is not a global minimum. Finally find an f and D where there is no minimizer. You may want to refresh your memory by looking at Chapter 3.3.

1.2 Examples in Several Variables

The next few examples require some vector notation. We write $\mathbf{x}$ for the vector $[x_1, x_2, \dots, x_n]$. When we have only two variables, we sometimes write $[x, y]$.

1.2.1 Example ($f(x, y)$ linear)**.** Consider the objective function $f(x, y) = 2x+3y$. The feasible set D is the positive quadrant $x \geq 0, y \geq 0$. We want to minimize f on this feasible set. Because the function is strictly increasing in both the x and the y directions, the minimum value is attained at the lowest possible values for x and y, that is, $x = y = 0$. The minimum value $f(0, 0) = 0$. This solution can also be illustrated graphically:



1.2.2 Definition. An *extreme point* of the set D is a point $\mathbf{p}$ that satisfies the following property: let L be any line in the plane passing through $\mathbf{p}$. Then $L \cap D$ does not contain an interval around $\mathbf{p}$.

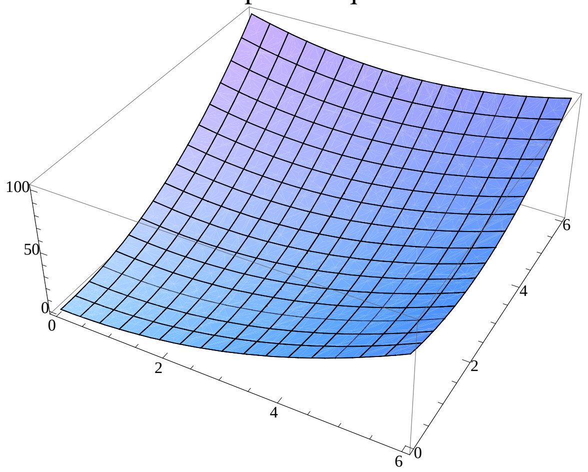
What are the extreme points of D? We will generalize the notion of extreme point to any convex set in Definition 18.1.10.

1.2.3 Exercise. We minimize $f(x, y) = ax + by + c$ on the positive quadrant D as before, where a, b, c are real constants. Why does the value of c not matter? For which choices of a and b does a minimizer exist? If a minimizer $\mathbf{p}$ does exist, show that it always occurs at an extreme point of D. As discussed in Lecture 25, this result is a consequence of the fact that a linear function has no *critical* points, meaning points where its gradient is the zero vector.

1.2.4 Exercise ($f(x, y)$ quadratic)**.** Suppose now that f is a quadratic polynomial in two variables:

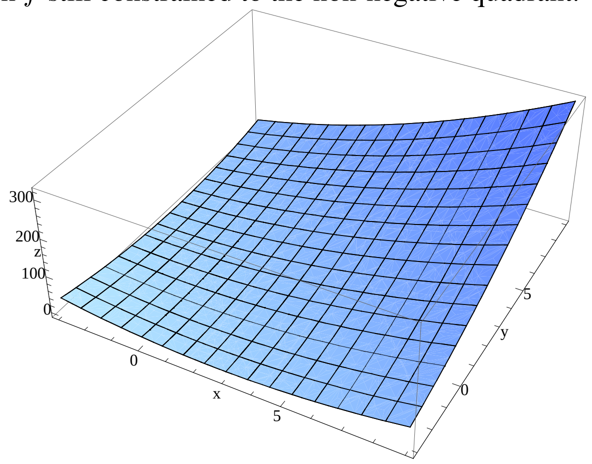
$$f(x, y) = x^2 - 2xy + 3y^2 - y$$

again constrained to the positive quadrant.



Does f have a minimum on the non-negative quadrant $D = \{(x, y) \mid x \geq 0, y \geq 0)$. Does it have a maximum there? (Hint: by completing the square on two groups of terms, notice that you can write $f(x, y) = (x-y)^2 + 2(y-1/4)^2 - 1/8$. We will use this technique in $\S$8.6 .)

1.2.5 Exercise. Consider the quadratic function $f(x, y) = x^2 + 2xy + y^2 + x - y$, with f still constrained to the non-negative quadrant.



In terms of minimization, what is the key difference between this exercise and the previous one?

Now suppose the feasible set is all of $\mathbb{R}^2$. Is there a minimum or a maximum? Explain.

1.2.6 Exercise ($f(x_1, x_2, \dots, x_n)$ quadratic)**.** In the quadratic case, we often deal with a large, often unspecified, number of variables. We use n to denote the number of variables and use summation notation to write

$$f(x_1, x_2, \dots, x_n) = \sum_{i=1}^{n} \sum_{j=i}^{n} a_{ij} x_i x_j + \sum_{k=1}^{n} b_k x_k \tag{1.2.7}$$

where a_{ij}, $1 \le i \le j \le n$ and b_k , $1 \le k \le n$ are real constants. Note that the first term, $a_{ij} x_i x_j$, is quadratic, while the second term, $b_k x_k$, is linear.

Write out the summation when $n = 3$. How many a_{ij} are there?

The question of when the objective function in (1.2.7) has a maximum or a minimum will occupy us for a large part of this course. The key is to notice that the quadratic coefficients a_{ij} can be organized into a matrix. This will be studied in Lecture 8.

For the rest of this lecture we will look at some applied problems that will be solved later using the math techniques we develop. The first two are linear optimization problems that have appeared in textbooks at least since Gale's very readable text [24].

1.3 The Diet Problem

The Diet Problem, originally formulated in a 1945 paper by economist George Stigler [64], is an amusing example of linear optimization, which we will study in detail in §26.1. Treatments of this problem can be found in all books convering linear optimization. It is also dealt with in some linear algebra books: see for example [39], p.175. The question posed by the Diet Problem is this: What is the minimal cost of a nutritionally adequate diet? We assume that human nutritional requirements, nutritive content of foods, and cost of foods are known quantities.

Assume we have n foods, labeled $1, 2, \dots, n$. The cost per unit of food j we call c_j. We also have m nutrients, labeled $1, 2, \dots, m$. The minimum daily requirement for each nutrient we denote by b_i, where i ranges over the set of nutrients, 1 through m. Finally, let a_{ij} be the amount of nutrient i in one unit of food j.

For variables, let x_j denote the quantity of food j purchased. Then the objective function is

$$f(x_1, \dots, x_n) = c_1 x_1 + c_2 x_2 + \dots + c_n x_n,$$

the cost of purchasing x_j units of food j, for $1 \leq j \leq n$. It is subject to constraints, meaning that the feasible set does not include all possible values of the x_j. To begin with, we have n positivity constraints given by $x_j \geq 0$ for $1 \leq j \leq n$: we can only purchase non-negative qualities of food. More subtly, we also need to constrain the feasible set such that the daily minimum requirement of each nutrient has been met. Thus each nutrient i, for $1 \leq i \leq m$, must satisfy the inequality:

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n \geq b_i$$

Indeed $a_{ij}x_j$ is the amount of nutrient i in x_j units of food j. Since we have m nutrients, we have m nutrient constraints.

This is a typical linear optimization problem. The following example is simple enough to be solved by hand.

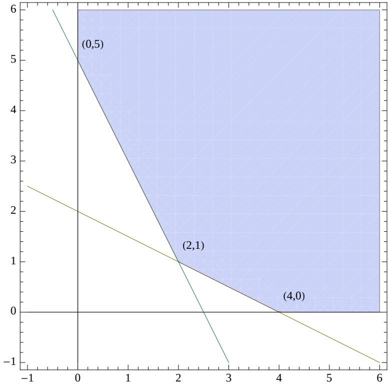
1.3.1 Example. Suppose that $m = n = 2$, so we have two foods and two nutrients, and the constraints are provided by the quantity of nutrients needed:

$$\begin{aligned} x_1+2x_2 &\geq 4 \\ 2x_1+\ x_2 &\geq 5. \end{aligned} \tag{1.3.2}$$

so

$$\begin{aligned} a_{11} &= 1, \quad a_{12}= 2 \\ a_{21} &= 2, \quad a_{22}= 1 \\ b_1 &= 4, \quad b_2\ = 5 \end{aligned} \tag{1.3.3}$$

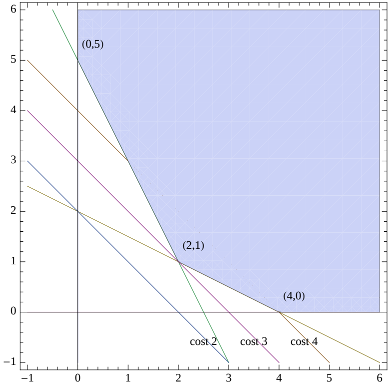
We now draw the (shaded) feasible set F in the first quadrant. Note that it contains all the rest of the first quadrant that is not shown.



It is given by the two inequalities of (1.3.2). plus the two positivity constraints $x_1 \geq 0$, $x_2 \geq 0$. Its boundary is polygonal with vertices $\mathbf{a} = (0, 5)$, $\mathbf{b} = (2, 1)$ and

$\mathbf{c} = (4,0)$. The sides of the polygon are the vertical x_2-axis, the segment $[\mathbf{a},\mathbf{b}]$, the segment $[\mathbf{b},\mathbf{c}]$ and the horizontal x_1-axis.

The cost function is $c_1x_1+c_2x_2$, and we assume that both c_1 and c_2 are positive, so that neither food is free. Then the level curves of fixed cost are lines with negative slope $-\frac{c_1}{c_2}$. The slope depends on the relative cost of the two foods. The line L of minimal cost γ has equation $c_1x_1 + c_2x_2 - \gamma = 0$ for some γ. As we shall see later, L is a *supporting line* (see Equation 18.6.10) for the *convex set* F, such that F is entirely contained in the closed halfspace delimited by L where the function $c_1x_1 + c_2x_2 - \gamma$ is non negative. Depending on its slope, the line L will intersect the feasible set at the vertex $\mathbf{a}$, or along the segment $[\mathbf{a},\mathbf{b}]$, or at the vextex $\mathbf{b}$, or along the segment $[\mathbf{b},\mathbf{c}]$ or at the vextex $\mathbf{c}$.



If we take $c_1 = c_2 = 1$, then we add to the previous graph three level curves for costs 2, 3, and 4. The minimum total cost is 3, and in this case the feasible set meets the level line in just one point.

1.3.4 Exercise. In the spirit of the previous example, let the objective function be $c_1x_1 + c_2x_2$, where c_1 and c_2 are positive, and the constraints be:

$$\begin{aligned} x_1+2x_2 &\geq 4,\\ 2x_1+5x_2 &\geq 9. \end{aligned}$$

and $x_1 \geq 0$, $x_2 \geq 0$. Draw the feasible set: first graph the lines that give the boundary of the feasible set. For all choices of (c_1,c_2), find the solution to this minimization problem. Show graphically that your solutions are correct.

Because we can see geometrically what the feasible set looks like, we can solve the problem readily. As soon as $n > 3$, we can no longer be guided by the geometric picture, and we will have to develop a general algorithm for solving the problem. This is done in Lectures 25 and 27.

1.4 The Transportation Problem

Here is a second linear optimization problem, introduced by Hitchcock [31] in 1941 and studied further by Koopmans [34] in 1949. It is one of the most studied of all linear optimization problems.

A commodity (say oil) is produced by a company at m plants P_i, $1 \le i \le m$ and is shipped to n markets M_j, $1 \le j \le n$. We let c_{ij} be the known cost of shipping a barrel of oil from P_i to M_j, and we let x_{ij} denote the unknown quantity of oil (in barrels) shipped from P_i to M_j. With these definitions, the total cost C of shipping the oil from all plants to all markets is

$$f(x_{11}, \dots, x_{mn}) = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij} \tag{1.4.1}$$

This is our objective function. It is a linear function in the mn variables x_{ij}. We wish to minimize it subject to some constraints. Before doing that we must introduce some additional constants: the supply of oil at plant P_i is s_i barrels, and the demand for oil at market M_j is d_j barrels. The company wishes to satisfy the demand at each market.

What is the feasible set, using these constants? First we can only ship non-negative amounts of oil from each plant, so

$$x_{ij} \ge 0, \quad 1 \le i \le m, 1 \le j \le n.$$

This gives $m \times n$ positivity constraints. Next the total amount of oil shipped out of each plant P_i cannot exceed the supply s_i of oil at that plant:

$$\sum_{j=1}^{n} x_{ij} \le s_i, \quad 1 \le i \le m. \tag{1.4.2}$$

This gives an additional m inequality constraints. Note that we are summing over the markets. Finally the company wants to satisfy the demand at each market M_j:

$$\sum_{i=1}^{m} x_{ij} \ge d_j, \quad 1 \le j \le n. \tag{1.4.3}$$

This gives an additional n inequality constraints. Here we are summing over the plants.

We have our objective function and our feasible set. Our minimization problem is

1.4.4 Problem (The Transportation Problem)**.** Supply all the markets with a sufficient amount of oil at minimal cost, in other words: Minimize $f(x_{ij})$ subject to the three sets of constraints given above.

Here we will only answer the question: is the feasible set non-empty? For that to be true, it is clear that to satisfy (1.4.3) we must assume that the total supply of oil is at least as great as the total demand for oil. In other words,

$$\sum_{i=1}^{m} s_i \geq \sum_{j=1}^{n} d_j. \tag{1.4.5}$$

Since one can change the order of summation in the double summation, the constraints imply

$$\sum_{i=1}^{m} s_i \geq \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{m} x_{ij} \geq \sum_{j=1}^{n} d_j \tag{1.4.6}$$

so that (1.4.5) is a consequence of the constraints. This says that (1.4.5) is a necessary condition for the feasible set to be non-empty.

Assuming (1.4.5) is satisfied, we simplify the problem by introducing new variables called *slack variables* to enable us to replace inequalities by equalities. This standard trick will be used repeatedly in these notes. Define the quantity $d_0 = \sum_{i=1}^{m} s_i - \sum_{j=1}^{n} d_j$ to be the *excess supply*. By hypothesis it is non-negative. If it is positive, we invent a new market M_0 we call the dump, with all transportation costs to the dump $c_{i0} = 0$, and with demand d_0. We have m new variables associated to the dump: the number x_{i1} of barrels of oil transported from plant P_i to the dump. With this additional market, we now have an exact match between supply and demand. This forces the inequalities in (1.4.6) to be equalities and therefore the ones in (1.4.2) and (1.4.3) to be equalities too. This suggests that we consider a new problem, called the canonical transportation problem.[3] For simplicity of notation, we simply increase m by 1, and let the dump be the last market.

1.4.7 Problem (The Canonical Transportation Problem)**.** Minimize $f(x_{ij})$ subject

[3] Some authors call this the standard transportation problem instead. See Definitions 25.1.5 and 25.1.6.

to the constraints

$$x_{ij} \geq 0, \quad 1 \leq i \leq m, 1 \leq j \leq n.$$

$$\sum_{j=1}^{n} x_{ij} = s_i, \quad 1 \leq i \leq m. \tag{1.4.8}$$

$$\sum_{i=1}^{m} x_{ij} = d_j, \quad 1 \leq j \leq n. \tag{1.4.9}$$

where $\sum_{i=1}^{m} s_i = \sum_{j=1}^{n} d_j$.

The construction of the dump shows that a point in the feasible set D of the ordinary transportation problem yields a point in the feasible set D' of the canonical transportation problem. Conversely, forgetting the dump moves you from a point in the feasible set of the canonical transportation problem D' to a point in D. Furthermore, as we will see in Theorem 25.2.2 a minimum for one problem will yield a minimum for the other.

Notice that we have nm variables that must satisfy $n + m$ affine[4] equations - as well as positivity constraints that we ignore for the time being. To understand the situation we use linear algebra to determine the rank of the system of affine equations. It is at most $n + m$, since that is the number of equations, but in fact it is never more than $n + m - 1$, since the sum of all the supply equations is equal to the sum of all the demand equations, as you should check.

Let us now examine the simplest cases, continuing to assume that the supply equals the demand.

1.4.10 Remark. If there is only one plant ($m = 1$) and there are n markets, there is a unique solution $x_{1j} = d_j$, so the $m + 1$ affine equations impose m linear conditions. If there are m plants and only one market ($n = 1$), then $x_{i1} = s_i$, so again there is a unique solution.

We conclude this introduction to the transportation problem by proving the following theorem:

1.4.11 Theorem. *Condition* (1.4.5) *insures that the feasible set is non-empty.*

Proof. It is enough to consider the canonical transportation problem. By Remark 1.4.10 there is a solution if either m or n is equal to 1, so we may assume that they are both strictly greater than 1 and proceed by induction on $n + m$.

Consider the x_{ij} as the entries of a $m \times n$ matrix X. By hypothesis, we know that the i-th row sum of X is s_i and the j-th column sum is d_j. Consider any entry of the matrix X, say x_{11}. Then either

[4] Affine equations are linear equation with a constant term: we study them in §18.2.

1. $s_1 \leq d_1$, in which case we let $x_{11} = s_i$, which means that we use the full supply of plant P_1 for market M_1, or

2. $s_1 > d_1$, in which case we let $x_{11} = d_i$, which means that we fulfill the whole demand at market M_1 using only oil from plant P_1.

In case 1, all the other entries in the first row of X must be 0. Let d'_1 be the residual demand at M_1, namely $d'_1 = d_1 - s_1$, which is non-negative by hypothesis. The entire supply of P_1 goes to M_1, so we can remove P_1, and replace the demand at M_1 by the residual demand. We now have a system with $m - 1$ plants and n markets. So we can proceed by induction.

In case 2, all the other entries in the first column of X must be 0. Let $s'_1 = s_1 - d_1$, the residual supply at P_1. This is positive by hypothesis. The entire demand at M_1 comes from P_1, so we can eliminate M_1 from the system and replace the . We now have a system of m plants and $n - 1$ markets. Again we can proceed by induction. □

1.4.12 Example. Let $m = 2$ and $n = 3$, so we have two plants P_1 and P_2, and three markets M_1, M_2, and M_3. Organize the transportation costs c_{ij} as the 2×3 matrix

$$\begin{bmatrix} 2 & 3 & 4 \\ 1 & 4 & 6 \end{bmatrix}$$

so $c_{11} = 2$, $c_{23} = 6$, etc. The supply at the two plants is $(4, 6)$ (in millions of barrels), so $s_1 = 4$, $s_2 = 6$, and the demand at the three markets are $(2, 3, 5)$ millions of barrels, so $d_1 = 2$, etc. Note that supply equals demand. The goal is to follow the algorithm proposed in the proof of Theorem 1.4.7.

1. Fill the entire demand at M_1 using the oil at P_1. We can then remove M_1, and there are only 2 millions of barrels left at P_1.

2. Use all the remaining oil at P_1 in M_2, reducing the demand there to 1.

3. The only oil left is at plant P_2: we use 1 to satisfy the residual demand at M_2, and then the remaining oil exactly satisfies the demand at M_3, so we are done.

What is the total cost of this solution? We read this off from the cost matrix:

$$f(2, 2, 0, 0, 1, 1) = 2c_{11} + 2c_{12} + 1c_{22} + 5c_{23} = 4 + 2 + 4 + 30 = 40.$$

If instead we first fill the entire demand at M_1 using P_2, and then use the remaining oil at P_2 in M_3, we get instead

$$f(0, 3, 1, 2, 0, 4) = 2c_{21} + 4c_{23} + 3c_{12} + 1c_{13} = 2 + 24 + 9 + 4 = 39,$$

so we have reduced the cost slightly. Can you do better?

Thus we have a method for finding an element in the feasible set. It will usually not be a minimal for the objective function. We will show how to find the minimal cost, and the minimizer later in the course: §26.2. It is a consequence of the duality theorem of Lecture 25. When the number of plants and markets is small, we can find the minimum by using the supply and demand equations, as described below.

Later we will determine the structure of the feasible set. It is the intersection of an affine set (see §18.2) with the positive orthant. In fact it is a compact set, a notion we will study in Lecture 14. Then one of the most important theorems of this course, the Weierstrass Theorem 16.2.2 says that there always is a minimum. However the theorem is a pure existence theorem and it does not say how to find the minimizer or the minimum value. That is the hard work we postpone to §26.2.

We can organize the computation systematically using linear algebra.

1.4.13 Exercise. View the variables x_{ij} as the entries of an unknown mn vector $\mathbf{x}$, where we list the entries in the order

$$(x_{11}, x_{12}, \dots, x_{1n}, \dots, x_{m1}, \dots, x_{mn})$$

so x_{ij} comes before x_{kl} if $i < k$ or $i = k$ and $j < l$.

We now form a $(m + n) \times mn$ matrix A specifically designed so that

$$A\mathbf{x} = \mathbf{b}$$

where $\mathbf{b} = (s_1, \dots, s_m, d_1, \dots, d_n)$. In other words the appropriate row of this equation is one of the m supply equations 1.4.8 first, followed by the n demand equations 1.4.9. The entries of A are either 0 or 1: using i (resp. $m + j$) as row index for the m (resp. last n) rows of A, and ij as the column index, we see that $a_{i,ij} = 1$ and $a_{m+j,ij} = 1$

Note that the matrix A does not depend on the problem, but only on the number of plants and the number of markets. If $m = 2$ and $n = 2$, the matrix A is the 4×4 matrix

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \tag{1.4.14}$$

This matrix will reappear when we study doubly stochastic matrices in §18.8. If

$m = 2$ and $n = 3$, the matrix A is the 5×6 matrix

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Show that the matrix A has rank $m + n - 1$ in the two examples above. This means that we will be able to eliminate $m + n - 1$ of the mn variables. In the case $m = 2$ and $n = 3$ that we have been considering, this means being able to eliminate 5 of 6 variables. Notice also how simple the three demand equations are are: $x_{2j} = d_j - x_{1j}$.

Extra credit: prove that A has rank in $m + n - 1$ in the general case. See Theorem

1.4.15 Exercise. Use the notation of Example 1.4.12, with $m = 2$ and $n = 3$, and transportation costs:

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{bmatrix}$$

The supply at the two plants is $(4, 6)$ (in millions of barrels), and the demand at the three markets are $(2, 3, 5)$ millions of barrels, so supply matches demand. Find a feasible solution using the method described above. Explain how to improve your solution, if possible. Consider using the technique of Exercise 1.4.13.

1.4.16 Exercise. Find a point in the feasible set for the following transportation problem. There are two plants P_1 and P_2, and three markets M_1, M_2, and M_3. The transportation costs c_{ij} are given as the entries of the 2×3 matrix

$$\begin{bmatrix} 2 & 3 & 4 \\ 1 & 2 & 5 \end{bmatrix}$$

The supply at the two plants is $(4, 6)$ (in millions of barrels), so $s_1 = 4$, $s_2 = 6$, and the demand at the three markets are $(2, 3, 5)$ millions of barrels, so $d_1 = 2$, $d_2 = 3$, $d_3 = 5$. By thinking through what the transportation costs are, try to lower the cost of the solution you have already found. Can you determine the minimum cost?

1.4.17 Exercise. Continuing with Exercise 1.4.16, use the matrix of Exercise 1.4.13 to solve this transportation problem as follows. Use the four linear equations to solve for x_{21}, x_{22}, x_{23} and finally x_{13}. What do you get for the remaining two variables? What does this mean? Find the minimizers for the x_{ij}, remembering that they are non-negative, and find the minimum cost.

1.5 Maximizing Utility Functions

We assume a universe where n different goods are available. We label them by the index j, $1 \leq j \leq n$, and use the variable $\mathbf{x} = (x_1, \dots, x_n)$ to denote the quantity of the goods. We let p_j be the price of the j-th good, so we have a price n-vector $\mathbf{p}$. Prices will always be nonnegative.

A consumer (or household) has a *utility function* concerning the acquisition of these goods. This is a real-valued function $u(\mathbf{x})$ from the non-negative quadrant in $\mathbb{R}^n$, and its main property is this: Given two different "bundles" of goods $\mathbf{x}^1$ and $\mathbf{x}^2$, the consumer strictly prefers $\mathbf{x}^1$ if $u(\mathbf{x}^1) > u(\mathbf{x}^2)$. In other words, we assume the consumer wants to maximize the utility function. Generally the consumer will have a budget constraint M, meaning that only bundles of total price $p_1x_1 + \dots + p_nx_n \leq M$ can be considered. Therefore we have an optimization problem:

Maximize $u(\mathbf{x})$ subject to $\mathbf{p} \cdot \mathbf{x} \leq M$ and $\mathbf{x} \succeq \mathbf{0}$.

We want to solve this problem for functions u satisfying very general assumptions that make sense in economics.

One candidate for a utility function is the Cobb-Douglas function: see Example 13.4.1 and its interpretation in Example 13.4.7. For example,

$$\frac{\partial u}{\partial x_j} > 0, \text{ for all } j, \tag{1.5.1}$$

so that u is strictly increasing as a functon of each of its variable. Economists say that the *marginal utility* of each good is strictly positive. A second property is that

$$\frac{\partial^2 u}{\partial x_j^2} < 0, \text{ for all } j, \tag{1.5.2}$$

so that the marginal utility decreases as consumption increases.

Another closely related assumption is that u is concave or at least quasiconcave. See Lectures 21, 22, and 23.

Lecture 2

Mathematics Terminology

This short lecture starts with a section on the language of mathematics, for reference. Some readers will have already seen much of this. Then we discuss what remains of the notion of order on $\mathbb{R}$ that we study in §14.1 when one passes to $\mathbb{R}^n$, $n > 1$: it is called a partial order. We then discuss binary relations in more detail than usual. One concept that may be new to you is duality, defined in subsection 2.3. I recommend that you only study the material of this lecture when you need it later in this course.

2.1 The Language of Mathematics

In this section we briefly list some notation and a collection of facts about the underpinning of mathematics. Most of this is probably familiar to the experienced reader. In any case it should only be skimmed until the material is needed.

2.1.1 Statements and Quantifiers

A mathematical statement is usually written P and Q. Most of the statements of concern to us will depend on a variable, say, x. Then we write $P(x)$ is the statement is true.

Typically we will need to write

- that $P(x)$ is true for all x. We can write this using the *universal quantifier* $\forall$.
- or that there is a x for which $P(x)$ is true. We write this using the *existential quantifier* $\exists$.

In the beginning of the text, the words 'for all' and 'there exists' are written out. It is a good exercise to replace the written out statement with the qualifiers. For example, the Definition 3.1.1 of the limit q of a function $f(x)$ at a point p can be written: The function $f(x)$ has limit q at the point p if

$$\forall\epsilon > 0, \exists\delta > 0 \text{ such that}\big(|x-p| < \delta, x \neq p\big) \Rightarrow |f(x) - q| < \epsilon.$$

As you know, the order in which the different quantifiers are listed in a mathematical statement is critically important.

2.1.2 Logical Connectors and Truth Tables

The two logical connectors are

- *and*, written $\wedge$. The symbol was deliberately chosen to suggest an intersection, since the locus of values x for which both $P(x)$ and $Q(x)$ are true is the intersection of the loci where $P(x)$ is true and where $Q(x)$ is true.
- *or*, written $\vee$. This symbol was chosen because it suggests a union. It is important to note that in mathematics *or* always means that either $P(x)$ is true or $Q(x)$ is true or both are true. In common language *or* sometimes means that either either $P(x)$ is true or $Q(x)$ is true , but not both. This is called the *exclusive or* and in our notation is written $P(x) \vee Q(x) \setminus P(x) \wedge Q(x)$: we remove the locus $P(x) \wedge Q(x)$ from the locus $P(x) \vee Q(x)$ in which it is contained.

  We will use the 'exclusive or' on at least two occasions: in Corollary 7.2.4 in the context of linear algebra, and in the important Farkas Alternative 19.5.1 dealing with linear inequalities.

2.1.3 Negation, If Then, and the Contrapositive

The statement that $P(x)$ is false is written $\neg P(x)$. Note that $\neg\forall x P(x)$ is equivalent to $\exists x \neg P(x)$.

The statement that 'if P is true, then Q is true' is written $P \Rightarrow Q$, and can be read P implies Q. If $P \Rightarrow Q$ and $Q \Rightarrow P$, we write $P \iff Q$, and we say that P is true if and only if (abbreviated *iff*) Q is true.

This is discussed in the language of necessary and sufficient conditions in Remark 3.3.3, which is followed by the important Theorem 3.3.4 that illustrates its use. To translate into the language used here.

A 'necessary condition for P is that Q' means that $P \Rightarrow Q$.

A 'sufficient condition for P is that Q' means that $Q \Rightarrow P$. An important method of proof in these lectures is the *contrapositive*, namely, to prove that $P \Rightarrow Q$, we prove $\neg Q \Rightarrow \neg P$. In other words, to prove that P implies Q, we prove that Q false implies that P is false.

2.1.4 De Morgan's Law

Let $\mathcal{S}$ be a fixed set, and I an index set, finite or infinite. Consider a collection of subsets S_i of $\mathcal{S}$ indexed by $i \in I$. Then, as you know, we write the union of the sets S_i as $\cup_{i \in I} S_i$ and the intersection $\cap_{i \in I} S_i$. In terms of our quantifiers, we see that

$$\cup_{i \in I} S_i = \{s \in \mathcal{S} \mid \exists i \in I, s \in S_i\},$$

and

$$\cap_{i \in I} S_i = \{s \in \mathcal{S} \mid \forall i \in I, s \in S_i\}.$$

The complement of S in $\mathcal{S}$ is written S^c. So

$$S^c = \{s \in \mathcal{S} \mid s \notin S\}.$$

2.1.1 Theorem (De Morgan's Law)**.** *The complement of an intersection is the union of the complements:*

$$\big(\cap_{i \in I} S_i\big)^c = \cup_{i \in I} S_i^c.$$

It is a good exercise to work this out when there are just two subsets A and B: $\big(A \cap B\big)^c = A^c \cup B^c$.

2.1.5 Index Notation

In these lectures we will often use index notation. It is important to understand that the notation, say, x_i, $i \in \mathbb{N}$, is simply a way of writing a function from the natural numbers $\mathbb{N}$ to whatever set the x_i belong to: in this course, typically $\mathbb{R}$. For examples, see §10.1.

2.2 Binary Relations

2.2.1 Ordered Pairs, Correspondences and Functions

Given two sets S and T, the cartesian product $S \times T$ is simply the set of pairs (s, t), where $s \in S$ and $t \in T$.

2.2.1 Definition. A *binary relation* R, sometimes called a *correspondence* on $S \times T$ is a subset R of $S \times T$. For $s \in S$ and $t \in T$, we write sRt to indicate that the pair (s, t) is in R.

One often takes the set T to be the same as S. When that is the case, we can define some interesting kinds of binary relations: this is done in Definitions 2.2.5 and 2.2.8.

The word correspondence is used to designate binary relations that are like functions. Then one speaks of the domain of the correspondence: it is the collection of $s \in S$ such that there is at least one $t \in T$ with sRt. The range of the correspondence is the collection of $t \in T$ such that there is at least one $s \in S$ such that sRt. In particular, a *function* is a correspondence such that for all s in the domain of R, there is a unique t in the range.

2.2.2 Example. Let S be the non-negative real numbers, and T be $\mathbb{R}$. Then the set of pairs $(r, \pm\sqrt{r})$ gives a correspondence that is not a function.

2.2.3 Exercise. In this language, write down the statement that a function has an inverse.

2.2.2 Special Kinds of Binary Relations

This section is devoted to studying the properties of different kinds of binary relations. First some examples.

In §14.1 we note that $\mathbb{R}$ is ordered. This means that there is a binary relation, called an *order*, denoted $>$ on $\mathbb{R}$ with the properties given by Definition 14.1.1. There is no such order on the complex numbers $\mathbb{C}$. But something does remain when one considers $\mathbb{C}$, $\mathbb{N}^n$, or $\mathbb{R}^n$: a binary relation called a *partial order*.

2.2.4 Definition (Partial Order)**.** For two vectors $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$, we write

- $\mathbf{x} \succeq \mathbf{y}$, if $x_i \geq y_i$ for $1 \leq i \leq n$.
- $\mathbf{x} \succneqq \mathbf{y}$, if $x_i \geq y_i$ for $1 \leq i \leq n$, and $\exists i$ such that $x_i > y_i$.
- $\mathbf{x} \succ \mathbf{y}$, if $x_i > y_i$ for $1 \leq i \leq n$.

This is called a partial order because it is not always possible to order two elements. For example, in $\mathbb{R}^2$ it is not possible to say that $(1, -1)$ is bigger or smaller than $(-1, 1)$. They cannot be compared.

Now some general definitions about binary relations.

2.2.5 Definition. Let R be a binary relation on a set S. Then

1. R is *reflexive* when xRx for all $x \in S$.
2. R is *symmetric* when xRy implies yRx.
3. R is *antisymmetric* when xRy and yRx implies $y = x$.
4. R is *complete* when for all x and y we have either xRy or yRx.
5. R is *transitive* when xRy and yRz implies xRz

Note that the order $\leq$ on $\mathbb{R}$ is reflexive, antisymmetric, complete and transitive. The partial order on $\mathbb{R}^n$ is reflexive, antisymmetric and transitive, but not complete.

This partial order has an extra property inherited from the vector space structure of $\mathbb{R}$:

2.2.6 Definition. A binary relation R on $\mathbb{R}^n$ is *linear* if

$$\mathbf{x}R\mathbf{y} \text{ implies } (c\mathbf{x}+\mathbf{z})R(c\mathbf{y}+\mathbf{z}) \text{ for all } c > 0 \text{ in } \mathbb{R} \text{ and all } \mathbf{z} \in \mathbb{R}^n.$$

2.2.7 Example. We can draw the binary relation R on $\mathbb{R}$ given by $x \geq y$: we plot the first term of the binary relation on the x-axis of $\mathbb{R}^2$, and the second term on the y-axis. Then the points in the binary relation are those on or below the 45 degree line through the origin in $\mathbb{R}^2$. You should check that it is the case.

For the partial order $\succeq$ on $\mathbb{R}^2$ one would need four dimensions to draw the full set $\mathcal{R}$. Instead, for a fixed $\mathbf{x}^* \in \mathbb{R}^2$, one can draw in $\mathbb{R}^2$ the set of $\mathbf{y}$ such that $\mathbf{x}^* \succeq \mathbf{y}$. What is it?

2.2.8 Definition. Here are the definitions in terms of the properties of Definition 2.2.5.

1. A binary relation that is reflexive, antisymmetric, complete and transitive is an *order*.
2. A binary relation that is reflexive, antisymmetric and transitive is a *partial order*.
3. A binary relation that is reflexive, symmetric and transitive is an *equivalence relation*.

You have probably already come across equivalence relations. For example, equality is an equivalence relation on any set. A key fact about an equivalence relation on a set S is that it partitions S into non-overlapping *equivalence classes*.

A partition of a set S is a collection of non-overlapping subsets S_i, called equivalence classes, whose union is S. Thus for any two i and j in I, $S_i \cap S_j$ is empty, and $\cup_{i \in I} S_i = S$. A partition defines a binary relation R on $S \times S$, whose domain and range is all of S: sRt if s and t are in the same subset S_i. You should check that R is an equivalence relation. Conversely any equivalence relation R defines a partition of S: start with any element $s \in S$

2.2.9 Example. Congruence modulo an integer k is an equivalence relation on the set of integers. Each equivalence class contains all the integers whose remainder modulo division by k is a fixed integer. Thus there are k equivalence classes.

We will meet three equivalence relations on matrices later in this course: see definitions 7.7.1, 7.7.8 and 8.4.1.

2.2.10 Example (Rays)**.** We will also meet a geometric equivalence relation on $\mathbb{R}^n$ with the origin removed. We say that two non-zero vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$ are equivalent if there is a $a > 0$ in $\mathbb{R}$ such that $\mathbf{x} = a\mathbf{y}$. In geometric terms they belong to the same *ray* through the origin. This equivalence relation is important because the set of equivalence classes is the set of points on the unit sphere S in $\mathbb{R}^n$. Indeed, the ray given by an equivalence class intersects the unit sphere in exactly one point. Indeed the map sending $\mathbb{R}^n \setminus \mathbf{0}$ to S is continuous and onto. This is why one removes $\mathbf{0}$, which would otherwise form its own equivalence class: there is nowhere to send it in S.

In Definition 15.5.2, we will put an equivalence relation on Cauchy sequences of rational numbers. The equivalence classes for this relation are the real numbers, as noted in §15.2.

Here is a more involved equivalence relation on functions.

2.2.11 Example (Preference Functions)**.** Consider a real-valued function $u(\mathbf{x})$ of n variables, which we will call a *utility* function. Given two inputs $\mathbf{a}$ and $\mathbf{b}$, we are only interested in comparing the real numbers $u(\mathbf{a})$ and $u(\mathbf{b})$, which are supposed to measure the preferences of a consumer. There are three possible outcomes:

- If $u(\mathbf{a}) > u(\mathbf{b})$, $\mathbf{a}$ is preferred to $\mathbf{b}$.
- If $u(\mathbf{a}) < u(\mathbf{b})$, $\mathbf{b}$ is preferred to $\mathbf{a}$.
- If $u(\mathbf{a}) = u(\mathbf{b})$, the consumer is indifferent between $\mathbf{a}$ and $\mathbf{b}$.

For every constant c in the range of u, the level sets $L_c = \{\mathbf{x} | u(\mathbf{x}) = c\}$ partition the domain of u into set of points between which the consumer is indifferent. As we learned above , this partition gives rise to an equivalence relation on the points in the domain of u. We say that u orders the elements $\mathbf{x}$ in its domain.

Now take two real-valued functions $u(\mathbf{x})$ and $v(\mathbf{x})$ with the same domain D in $\mathbb{R}^n$. Then u and v are equivalent if there exists a strictly increasing function $f \colon \mathbb{R} \to \mathbb{R}$ whose domain includes the range of u, such that

$$f(u(\mathbf{x})) = v(\mathbf{x}).$$

The point of this definition is that $f(u(\mathbf{x}))$ gives the same ordering on elements $\mathbf{a}$ and $\mathbf{b}$ as does $u(\mathbf{x})$.

You should check that this defines an equivalence relation on functions with the same domain. The key point is that since f is strictly increasing, it has an inverse function f^{-1}, so that if $f(u(\mathbf{x})) = v(\mathbf{x})$, then $u(\mathbf{x}) = f^{-1}(v(\mathbf{x}))$.

A property that is true for all functions in the same equivalence class of utility functions, is called an *ordinal* property of the utility function. It is important to determine the ordinal properties for certain classes of functions.

For example, restrict to $\mathcal{C}^1$-functions $u(\mathbf{x})$ defined on the positive quadrant, and $\mathcal{C}^1$-functions f that have an inverse on their entire range. By the inverse function theorem f^{-1} is also $\mathcal{C}^1$, so $f'(z)$ is never zero. The property

$$\frac{\partial u}{\partial x_i}(\mathbf{x}) > 0,$$

is an ordinal property: if $v(\mathbf{x}) = f(u(\mathbf{x})$, with f strictly increasing, then by the chain rule

$$\frac{\partial v}{\partial x_i}(\mathbf{x}) = f'(\mathbf{x})\frac{\partial u}{\partial x_i}(\mathbf{x})$$

so the left-hand side is strictly positive as required. In other words, more is better than less is an ordinal property.

Define the *marginal rate of substitution* between x_i and x_j at any point $\mathbf{x}$ to be

$$MRS(\mathbf{u}, i, j) = \frac{\partial u}{\partial x_i}(\mathbf{x}) / \frac{\partial u}{\partial x_j}(\mathbf{x}),$$

assuming the denominator does not vanish. Then the property that $MRS(\mathbf{u}, i, j) > 0$ is an ordinal property. The same chain rule computation gives the result.

2.3 Duality

We will spend a lot of time discussing duality in this course, so it is perhaps useful to give a general mathematical framework for duality immediately.

Suppose that $k(x, y)\colon A \times B \to C$ is a function of two variables. Then for each $a \in A$ we get a function $g_a(y)\colon B \to C$, defined by $g_a(y) = k(a, y)$; for each $b \in B$ we get a function $f_b(x)\colon A \to C$, defined by $f_b(x) = k(x, b)$. To know all the functions $g_a(y)$ is the same as knowing $k(x, y)$, which in turn is the same as knowing the all the functions $f_b(x)$.

2.3.1 Definition. The collection of functions $g_a(y)$ and $f_b(x)$ are *dual*.

Here is the key example. It requires the linear algebra developed in Lecture 6.

2.3.2 Example. Take a $m \times n$ matrix A. This yields a bilinear form

$$k(\mathbf{x}, \mathbf{y}) = \mathbf{y}^T A\mathbf{x}.$$

The function $g_{\mathbf{a}}(\mathbf{y})$ is then the linear function $k(\mathbf{a}, \mathbf{y}) = \mathbf{y}^T A\mathbf{a}$ and $f_{\mathbf{b}}(\mathbf{x})$ in the linear function $k(\mathbf{x}, \mathbf{b}) = \mathbf{b}^T A\mathbf{x}$.

In linear algebra, the dual of a vector space V is the vector space of functionals, namely linear maps from V to the scalars, here $\mathbb{R}$. The dual vector space is written V^*. If $\mathbf{v} \in V$ and $\varphi \in V^*$, then we have a pairing

$$k(\mathbf{v}, \varphi) \colon V \times V^* \to \mathbb{R} \text{ given by } k(\mathbf{v}, \varphi) = \varphi(\mathbf{v}).$$

So we are in the framework given above.

Part II

Optimization of Functions in One Variable

Lecture 3

Calculus in One Variable

In this lecture we assume that $f(x)$ is a function of a single real variable x, and review the parts of single variable calculus that are useful in optimization, focusing on the key definitions and conceptual understanding. Many of these concepts are covered in the multivariable setting latter in the course, but it is useful to see the single variable case, which is conceptually easier, first. A key ingredient is the Mean Value Theorem 3.2.1, which we will use repeatedly. Its relies on the Maximum Theorem 3.1.6, proved in full generality in Lecture 16. Then we review optimization in one variable: §3.3, and we conclude with the Intermediate Value Theorem 3.4.3.

3.1 Review of Single Variable Calculus

First, the most important definition of calculus:

3.1.1 Definition. The real-valued function $f(x)$ is defined on an open interval (a, b) in $\mathbb{R}$. We let p be a point in the closed interval $[a, b]$. We say that $f(x)$ approaches q, as x approaches p, if, for all $\epsilon > 0$, there exists a $\delta > 0$ such that when $|x - p| < \delta$, $x \neq p$, and $x \in (a, b)$, then $|f(x) - q| < \epsilon$. We write $\lim_{x \to p} f(x) = q$, and call q the limit of the function $f(x)$ at p. If there is no value q that works, we say the limit does not exist.

In other words, no matter how small an interval V (given by ϵ) you take around q, by taking a suitably small interval U (given by δ) around p and contained in (a, b), you can guarantee that if $x \in U$ and $x \neq p$, then $f(x) \in V$. In a calculus text such as [63], this definition is usually called the *precise* definition of a limit (see [63] definition 2.4.6). Note that p need not be in the domain of f, so we may not be able to evaluate f at p.

3.1.2 Remark. This definition allows for the possibility that the point p at which we are taking the limit be an endpoint of the interval, so that the limit can only be taken from one side, so that we get what is called a *one-sided limit*. We do this to prepare for the multivariable generalization in Definition 11.1.1.

3.1.3 Theorem. *If $f(x)$ has a limit at p, it is unique.*

Proof. This is a good exercise in providing a proof by contradiction. Assume that there are two distinct limits q_1 and q_2. Then pick an ϵ so small that the interval $|y - q_1| < \epsilon$ and the interval $|y - q_2| < \epsilon$ do not overlap. Then for that value ϵ, to say that q_1 (resp. q_2) is a limit is to say that for x close enough to p, $|f(x) - q_1| < \epsilon$ (resp. $|f(x) - q_2| < \epsilon$). But this is impossible since the intervals do not overlap. □

Next a result we will use repeatedly in this course: the inequalities $\geq$ and $\leq$ are preserved in the limit.

3.1.4 Theorem. *Assume that $f(x)$ has a limit at p, and that there is an $\epsilon > 0$ such that for all $x \neq p$ in $|x - p| < \epsilon$, $f(x) \geq r$. Then $\lim_{x \to p} f(x) \geq r$.*

Proof. This is another good exercise, best attempted after reading §14.2. It is easiest doing this by contradiction, like the previous theorem. Indeed, the proof is almost identical. □

It is not true that strict inequalities are preserved in the limit. Indeed just take the absolute value function $y = |x|$ for $p = 0$. For all x near 0, $|x| > 0$, and yet in the limit as one approaches 0, one gets the value 0.

The subsequent definitions build on the definition of a limit.

3.1.5 Definition. The real-valued function $f(x)$ defined on a closed interval $[a, b]$ in $\mathbb{R}$ is *continuous* at a point $p \in [a, b]$ if $\lim_{x \to p} f(x) = f(p)$. The function f is continuous on $[a, b]$ if it is continuous at all $p \in (a, b)$.

The central result we will need concerning continuous functions is

3.1.6 Theorem. *Let $f(x)$ be a continuous function defined on the closed (and bounded) interval $[a, b]$. Then there is a point x_0 in the interval where f attains its minimum, meaning that for all $x \in [a, b]$, $f(x) \geq f(x_0)$.*

The analogous result holds for maxima, of course. This is a significantly deeper result that the results proved up to now. The general case, for f is a function of several variables, is called the Weierstrass Theorem 16.2.2. More details are given there, but it still depends on a key property of the real numbers developed in §14.2. You could read that section now if you wish, as well of the results of §14.1.

3.1.7 Definition. The real-valued function $f(x)$ is defined on the open interval (a, b) in $\mathbb{R}$. For a fixed point $x \in (a, b)$, define the *Newton quotient* $\varphi(t)$ by

$$\varphi(t) = \frac{f(t) - f(x)}{t - x}, \quad \text{for all } t \in (a, b)\ , t \neq x.$$

Then if $\lim_{t \to x} \varphi(t)$ exists, it is called the *derivative* of f at x, and written $f'(x)$. The function f is then said to be *differentiable* at x. It is differentiable on (a, b) if it is differentiable at each point of (a, b).

This allows us to define the tangent line to $f(x)$ at a point x_0 where $f(x)$ is differentiable.

3.1.8 Definition. The tangent line to the graph $y = f(x)$ in the plane at the point with coordinates $(x_0, f(x_0))$ is the graph $y = \ell(x)$ by the affine function:

$$\ell(x) = f(x_0) + f'(x_0)(x - x_0).$$

Affine refers to the fact that we get the graph of a line that does not necessarily go through the origin. We will look at affine functions in §18.2.

Here is how we use this definition. Rewrite the limit defining the derivative as

$$\lim_{x \to x_0} \Big(\frac{f(x) - f(x_0)}{x - x_0} - f'(x_0) \Big) = 0. \tag{3.1.9}$$

Consider the term inside the limit, and multiply it by $x - x_0$ to get

$$f(x) - f(x_0) - f'(x_0)(x - x_0) = f(x) - \ell(x).$$

Applying Definition 3.1.1 to (3.1.9), for all $\epsilon > 0$, there exists a $\delta > 0$ such that when $|x - x_0| < \delta$, $x \neq x_0$, then

$$\Big| \frac{f(x) - f(x_0)}{x - x_0} - f'(x_0) \Big| < \epsilon. \tag{3.1.10}$$

3.1.11 Theorem. *The tangent line $\ell(x) = f(x_0) + f'(x_0)(x - x_0)$ approximates $y = f(x)$ near the point $(x_0, f(x_0)$ better than any other line in the plane. In other words,*

$$\lim_{x \to x_0} \frac{f(x) - \ell(x)}{x - x_0} = 0 \tag{3.1.12}$$

while for any other line, this limit is not 0.

Proof. The limit (3.1.9) gives (3.1.12). Now let $g(x) = a(x - x_0) + b$ be any other line in the plane. If $b \neq f(x_0)$, the desired limit does not even exist, so we may assume $b = f(x_0)$ and $a \neq f'(x_0)$. Then

$$\begin{aligned} |f(x) - g(x)| &= |f(x) - \ell(x) + \ell(x) - g(x)| \\ &\geq |\ell(x) - g(x)| - |f(x) - \ell(x)|, && \text{by the triangle inequality} \\ &\geq |f'(x_0) - a||x - x_0| - \epsilon|x - x_0|, && \text{if } |x - x_0| < \delta. \end{aligned}$$

In the last step, for the given $\epsilon > 0$, use the $\delta > 0$ given by (3.1.10). If we take $\epsilon < |a - f'(x_0)|/2$,

$$|f(x) - g(x)| \geq \frac{|a - f'(x_0)|}{2}|x - x_0|$$

so dividing by $|x - x_0|$ and taking the limit as $x \to x_0$, we get in the limit an expression bounded below by

$$\frac{|a - f'(x_0)|}{2} > 0,$$

so we are done. □

We will discuss higher dimensional analogs of the tangent line in §17.3, and approximations of higher order in §4.3.

3.1.13 Definition. To clarify the next results, we make some local and some less local definitions:

- We say that $f(x)$ *increases* (resp. *increases strictly*) at $x_0 \in (a, b)$ if on a small enough open interval $(\alpha, \beta) \subset (a, b)$ containing x_0, for all $y \in (\alpha, \beta)$ with $x_0 < y$, $f(x_0) \leq f(y)$ (resp. $f(x_0) < f(y)$), and for all $y \in (\alpha, \beta)$ with $y < x_0$, $f(y) \leq f(x_0)$ (resp. $f(y) < f(x_0)$) .
- We say that $f(x)$ *increases* (resp. *increases strictly*) on $[a, b]$ if for all $x < y$ in $[a, b]$, $f(x) \leq f(y)$ (resp. $f(x) < f(y)$).

We leave it to the reader to formulate the analogous definitions for *decreasing* and *strictly decreasing*.

We now derive an easy local result from the definition of the derivative at a point. Notice how weak the hypothesis is: we only need differentiability at the one point x_0. As we will see, since the result only compares the value of $f(x_0)$ with that of a point nearby, it does not imply that the function $f(x)$ is monotonically increasing or decreasing in any neighborhood of x_0, no matter how small.

3.1.14 Theorem. *Let $f(x)$ be defined on the closed interval $[a, b]$, and differentiable at $x_0 \in (a, b)$.*

- *If $f'(x_0) > 0$, then f is strictly increasing at x_0. If $f'(x_0) < 0$, then f is strictly decreasing at x_0.*
- *If $f(x)$ is increasing at x_0, then $f'(x_0) \geq 0$. If $f(x)$ is decreasing at x_0, then $f'(x_0) \leq 0$*

Proof. The key idea, to be reused many times in this course, is this: If $f'(x_0) > 0$, then for a small enough δ and any x with $|x - x_0| < \delta$, the Newton quotient

$$\frac{f(x) - f(x_0)}{x - x_0} > 0$$

and that in itself shows that $f(x)$ is strictly increasing at x_0. For the second statement, if $f(x)$ is increasing at x_0, then the Newton quotient at x_0 is ≥ 0 for x sufficient close to x_0. Then we just apply Theorem 3.1.4 that says that inequalities are preserved in the limit to get $f'(x_0) \geq 0$.[1] □

The missing case $f'(x_0) = 0$ in Theorem 3.1.14 is handled in the next theorem, the prototype of many theorems in this course. This result was proved by Fermat in 1638, and is probably the first result connecting extrema to differentiability. See [12], Theorem 3.9 and [28], chapter II for the history.

3.1.15 Theorem (Fermat's Theorem)**.** *Let $f(x)$ be defined on the closed interval $[a, b]$. Assume f has a local minimum at a point x_0 in (a, b). If $f(x)$ is differentiable at x_0, then $f'(x_0) = 0$.*

Proof. Since x_0 is a local minimum, so that $f(x) - f(x_0)$ is non-positive for all x close enough to x_0, the Newton quotient $(f(x_0) - f(x))/(x_0 - x)$ considered only for x such that $x_0 - x < 0$, must be non-positive. On the other hand, for x greater than x_0, the Newton quotient is non-negative. Thus, since these must agree in the limit, the derivative, which is the common limit of these quotients, must be 0. We have used Theorem 3.1.4 again. □

There is of course a similar theorem for local maxima. We need vocabulary to account for all cases, and here it is:

3.1.16 Definition. A *critical point* of $f(x)$ is a solution x to the equation

$$f'(x) = 0 \tag{3.1.17}$$

A critical point that is neither a local maximum nor a local minimum is an *inflection point*.

[1] Additional details can be found in [70], Theorem 5.2.1, for example.

3.1.18 Example. In Theorem 3.1.14, the hypothesis $f'(x_0) > 0$ does not imply that $f(x)$ is strictly increasing (or decreasing) in a neighborhood of x_0, no matter how small: it only says that it increases or decreases at x_0. Indeed, consider the function

$$f(x) = \begin{cases} x + x^2 \sin 1/x^2, & \text{if } x \neq 0; \\ 0, & \text{if } x = 0. \end{cases}$$

It is differentiable everywhere, even at 0, and the derivative at 0 is 1, and yet, as we now see, it is not monotonically increase on any neighborhood of 0. Since $\sin 1/x^2$ oscillates between -1 and 1, $f(x)$ oscillates between the parabola $y = x - x^2$ and the parabola $y = x + x^2$, which are both tangent at $x = 0$ to the line $y = x$. We can show that $f(x)$ is everywhere differentiable, as follows. A direct Newton quotient computation shows that $f'(0) = 1$. For $x \neq 0$, a standard derivative computation using the product rule and the chain rule says

$$f'(x) = 1 + 2x \sin \frac{1}{x^2} - \frac{2}{x} \cos \frac{1}{x^2}.$$

In particular $f'(x)$ fails to be continuous at $x = 0$. Note that $f'(x)$ takes on both positive and negative values in arbitrarily small neighborhoods of 0. This shows that f is neither increasing or decreasing in any neighborhood of 0, no matter how small. We revisit the techniques of this example in Example 3.4.5. The Intermediate Value Theorem 3.4.3 explains why examples of this nature must be rather complicated. For more details and graphs of both f and f', see [28], remark III.6.5 p. 237.

3.2 The Mean Value Theorem

We conclude this survey of single-variable calculus with one of its most important and deepest theorems, which we will use below in Corollary 3.2.4 and in Lecture 12 on Taylor's theorem. Its proof requires the Weierstrass Theorem 16.2.2, which we stated above in the case of a single variable: Theorem 3.1.6.

3.2.1 Theorem (Mean Value Theorem)**.** *Let $f(x)$ be a real-valued function that is continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b). Then there exists a point $c \in (a, b)$ such that*

$$f(b) - f(a) = f'(c)(b - a) \tag{3.2.2}$$

This theorem is due to Lagrange, and is often called the "finite increment theorem".[2] We will prove this theorem shortly. But first, we introduce the following special case, in which f takes the same value at a and b:

[2] In French, *le théorème des accroissements finis.* See [74] §5.3.2.

3.2.3 Theorem (Rolle's Theorem)**.** *Let $f(x)$ be a real-valued function that is continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b). If $f(a) = f(b)$, there is a point $c \in (a, b)$ where the derivative of f vanishes: $f'(c) = 0$.*

Proof. If f is the constant function $f(x) = v$, then the theorem is trivially true: for all $c \in (a, b)$, $f(c) = v$. Otherwise we can turn to the Weierstrass Theorem 16.2.2, applied to the continuous function $f(x)$ on the closed interval $[a, b]$. Since f is not constant, there is a point $c \in (a, b)$ where $f(c)$ is either greater than v or less than v. If $f(c) > v$, the maximum of f on $[a, b]$ must occur on the interior of the interval, where f is differentiable. At this maximizer d, $f'(d) = 0$, as we know from Theorem 3.1.15. If $f(c) < v$, the minimizer must occur on the interior, and we get the same conclusion. □

As promised, we now prove the Mean Value Theorem.

Proof of the Mean Value Theorem. The slope of the line connecting the points $(a, f(a))$ and $(b, f(b))$ in $\mathbb{R}^2$ is

$$s = \frac{f(b) - f(a)}{b - a}$$

If we replace f by the function $g(x) = f(x) - sx$, we see that $g(a) = g(b)$. This equality allows us to apply Rolle's theorem to $g(x)$, from which we know that there is a point $c \in (a, b)$ such that $g'(c) = 0$. Meanwhile, calculating the derivative of $g(x) = f(x) - sx$ produces $g'(x) = f'(x) - s$. Since $g'(c) = 0$, $f'(c) = s$. □

The Mean Value Theorem gives us a global analog of theorems 3.1.14 and 3.1.15, which follows immediately from (3.2.2).

3.2.4 Corollary. *Let $f(x)$ be a real-valued function that is continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b).*

- *If $f'(x) > 0$ for all $x \in (a, b)$, then f is strictly increasing on $[a, b]$;*
- *If $f'(x) \geq 0$ for all $x \in (a, b)$, then f is increasing on $[a, b]$;*
- *If $f'(x) = 0$ for all $x \in (a, b)$, then f is constant on $[a, b]$;*
- *If $f'(x) \leq 0$ for all $x \in (a, b)$, then f is decreasing on $[a, b]$*
- *If $f'(x) < 0$ for all $x \in (a, b)$, then f is strictly decreasing on $[a, b]$;*

The following generalization of the Mean Value Theorem is due to Cauchy.

3.2.5 Theorem (Cauchy's Mean Value Theorem)**.** *Let $f(x)$ and $g(x)$ be real-valued functions that are continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b). Then there exists a point $c \in (a, b)$ such that*

$$g'(c)(f(b) - f(a)) = f'(c)(g(b) - g(a)) \tag{3.2.6}$$

Proof. Just apply Rolle's theorem to the function

$$g(x)(f(b) - f(a)) - f(x)(g(b) - g(a)).$$

□

Cauchy ([16], p. 243) stated the result for a function g, such that $g'(x)$ is non -zero on the interval, so that we get:

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(c)}{g'(c)} \tag{3.2.7}$$

from which the Mean Value Theorem follows by setting $g(x) = x$.

For further results in this direction, see §3.4.

3.3 Optimization in One Variable

Differential calculus provides tools needed to find local minima for $f(x)$, as we have seen from the theorems in the preceding two sections. It this section we state two theorems (3.3.4 and 3.3.5) that follow readily from the results of the previous sections. We will generalize these theorems to functions of several variables in later lectures: see §13.1. Then we look at examples, and write the general algorithm for finding the minima and maxima on the interval.

To distinguish local maxima from minima and inflection points, we introduce the second derivative $f''(x)$. Thus we assume not only that $f(x)$ is differentiable, but that its derivative $f'(x)$ is also differentiable. We say that $f(x)$ is *twice differentiable*. Furthermore if we require that the second derivative $f''(x)$ itself be continuous: we say $f(x)$ is twice continuously differentiable, or $\mathcal{C}^2$. More generally

3.3.1 Definition. The function $f\colon \mathbb{R} \to \mathbb{R}$ is said to be $\mathcal{C}^k$ in a neighborhood of a if the derivatives $f^{(i)}(x)$, $1 \le i \le k$ exist and are continuous functions in a neighborhood of a. If $k = 1$, we say that f is *continuously differentiable*. To be $\mathcal{C}^0$ is just to be continuous.

3.3.2 Example. the function of Example 3.1.18 is differentiable, but not $\mathcal{C}^1$ at $x = 0$.

Note that if f has a k-th derivative on a neighborhood, it is automatically $\mathcal{C}^{k-1}$, but need not be $\mathcal{C}^k$. Except when we come to Taylor's Theorem in Chapter 4, we will rarely have to worry about conditions beyond $\mathcal{C}^1$ and $\mathcal{C}^2$.

The following two theorems are our main tools (also see [63], p.301). They illustrate another repeated theme in this book: the distinction between necessary and sufficient conditions.

3.3.3 Remark. A set of conditions for a particular conclusion are *necessary*, if the conclusion cannot be true unless all of the conditions are satisfied; however, the conditions being satisfied does not guarantee the conclusion to be true. On the other hand, a set of conditions for a particular conclusion are *sufficient*, if, when all the conditions are satisfied, the conclusion is true; however, the conclusion may still be true notwithstanding that some or all of the conditions are unsatisfied.

For a different approach to necessary and sufficient conditions see §2.1.3.

3.3.4 Theorem (Necessary conditions for a local extremum, single-variable case)**.** *Let $f(x)$ be a real-valued twice differentiable function defined on an open interval (a, b) in $\mathbb{R}$ containing the point x_0. Assume that $f''(x)$ is continuous near x_0. If f has a local minimum at x_0, then $f'(x_0) = 0$ and $f''(x_0) \geq 0$. If f has a local maximum at x_0, then $f'(x_0) = 0$ and $f''(x_0) \leq 0$.*

3.3.5 Theorem (Sufficient conditions for a local extremum, single-variable case)**.** *Let $f(x)$ be a real-valued twice differentiable function defined on an open interval (a, b) in $\mathbb{R}$ containing x_0. Assume that $f''(x)$ is continuous near x_0. If $f'(x_0) = 0$ and $f''(x_0) > 0$, then f has a strict local minimum at x_0. If $f'(x_0) = 0$ and $f''(x_0) < 0$, then f has a strict local maximum at x_0.*

Proof. We start with Theorem 3.3.5. Assume that $f'(x_0) = 0$ and $f''(x_0) > 0$. We need to show that there is a small enough interval (a, b) around x_0, such that if $x \in (a, x_0)$, then $f(x) > f(x_0)$; and if $x \in (x_0, b)$, then $f(x) > f(x_0)$ also. This is the meaning of *strict* minimum. We first work with the function $f'(x)$. The hypothesis $f''(x_0) > 0$ and Corollary 3.2.4 applied to $f'(x)$ show that $f'(x)$ is strictly increasing on a small enough interval (α, β) containing x_0. Indeed, since $f''(x_0) > 0$ and $f''(x)$ is continuous in a neighborhood of x_0, it remains positive in a neighborhood of x_0, so Corollary 3.2.4 can be applied to $f'(x)$.

Then, since $f'(x_0) = 0$, $f'(x)$ must be negative on the interval (α, x_0) and positive on the interval (x_0, β). So by Corollary 3.2.4 applied to the intervals (α, x_0) and (x_0, β), we see that x_0 is a strict minimizer for f on the interval (α, β). The second statement of Theorem 3.3.5 is proved in exactly the same way. □

Now we turn to the proof of Theorem 3.3.4.

Proof. We will prove the second statement, namely

$$f \text{ has a local maximum at } x_0 \;\Rightarrow f'(x_0) = 0 \text{ and } f''(x_0) \leq 0$$

We prove this by establishing the contrapositive, which is logically the same.[3] Since we know that $f'(x_0) = 0$, the contrapositive reads:

$$f'(x_0) = 0 \text{ and } f''(x_0) > 0 \Rightarrow f \text{ does not have a local maximum at } x_0.$$

This is true by Theorem 3.3.5, since the hypotheses imply that f has a strict local minimum at x_0, and a strict local minimum can never be a local maximum. Notice that this last argument fails if we only have a local minimum, since a local minimum can be a local maximum when the function is locally constant. □

The proof shows that Theorem 3.3.4 is a logical corollary of Theorem 3.3.5: it is implied by it just using elementary logic. As the careful reader will have noticed, the only circumstance in which the necessary conditions are satisfied but the sufficient conditions are not satisfied is when $f'(x_0) = 0$ and $f''(x_0) = 0$. Here is a refinement of Theorem 3.3.5 that closes most of the ground between the necessary and sufficient condition.

3.3.6 Theorem. *Let $f(x)$ be a real-valued function, at least n-times differentiable, defined on an open interval (a, b) in $\mathbb{R}$ containing x_0. Assume that $f^{(n)}(x)$ is continuous in a neighborhood of x_0, that the first $n-1$ derivatives of f at x_0 are zero: $f'(x_0) = 0$, $f''(x_0) = 0, \dots, f^{(n-1)}(x_0) = 0$, and that the n-th derivative of f as x_0 is not zero: $f^{(n)}(x_0) \neq 0$.*

- *Assume n is even. If $f^{(n)}(x_0)$ is positive, then f has a strict local minimum at x_0. If $f^{(n)}(x_0)$ is negative, then f has a strict local maximum at x_0.*
- *If n is odd, then f has an inflection point at x_0.*

Proof. The proof follows that of 3.3.5: start with $f^{(n-1)}(x)$, draw the usual conclusion about $f^{(n-2)}(x)$, and notice that the final conclusion just depends on the parity of n. □

At this point you could jump ahead to §4.2 to see how the mean value theorem is used to establish Taylor's theorem.

Now some examples.

[3] See §2.1.3 is you are unclear about the meaning of this sentence.

3.3.7 Example. Consider $f(x) = x^n$ near the point $x_0 = 0$. Its first $n-1$ derivatives at x_0 vanish, and its n-th derivative at 0 is $f^{(n)}(0) = n!$. So Theorem 3.3.6 tells us that when n is even, f has a strict minimum at 0, and when n is odd, an inflection point. Note that this also follows directly from consideration of the sign of x^n.

We now review and generalize the algorithm for finding the global minimum of one-variable functions. Let's assume that the feasible set is the closed interval $[\alpha, \beta]$. We assume that f is twice differentiable.

3.3.8 Algorithm.

Step 1. Compute the first derivative f' of f. This usually presents no difficulty.

Step 2. Solve the equation $f'(x) = 0$ to find the critical points. If $f'(x)$ is sufficiently complicated, this cannot be done in closed form and requires instead a process of approximation (e.g., Newton-Raphson, [63], §4.9). For polynomials in one variable of degree at least 5, for example, there is no closed-form formula for the roots.

Step 3. Determine whether the critical points are local maxima or local minima using the second-derivative test (Theorem 3.3.5) at each critical point. If any of the second derivatives are equal to 0, the test fails. Alternatively, one could just evaluate f at each critical point and pick the smallest value.

Step 4. Evaluate the function at each end point, getting $f(\alpha)$ and $f(\beta)$.

The smallest of all the values found is the global minimizer.

3.3.9 Example. Let's find the critical points of the function $f(x) = 2x^3 - 9x^2 + 12x$ on the feasible set $x \geq 0$. Looking to (3.1.17), we compute the derivative $f'(x)$ and set it equal to zero:

$$\begin{aligned} f(x) &= 2x^3 - 9x^2 + 12x \\ f'(x) &= 6x^2 - 18x + 12 && \text{so, setting it to 0,} \\ 0 &= x^2 - 3x + 2 = (x-2)(x-1). \end{aligned}$$

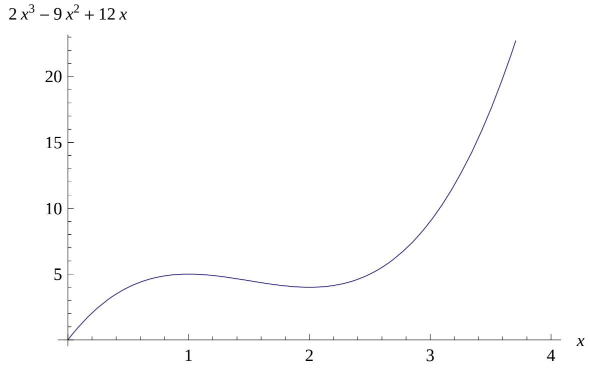
Thus, the function $f(x)$ has critical points at $x = 1$ and $x = 2$. Note that both of these values are feasible for the constraint $x \geq 0$. Because $f(1) = 5$ and $f(2) = 4$, we can write the corresponding points on the graph of $y = 2x^3 - 9x^2 + 12x$ using their coordinates in the plane: $(1, 5)$ and $(2, 4)$.

We do not yet know whether these critical points are maxima, minima, or inflection points. This often requires a second derivative computation. For this simple example, though, just consider that f goes to $+\infty$ as $x \to +\infty$ and f goes to

$-\infty$ as $x \to -\infty$. This indicates that $(1,5)$ is a local maximum and $(2,4)$ is a local minimum. Why? Since there are no other critical points, the function cannot change direction at any other point.

We have found the local extrema on f that are not at the end points of the interval (in this example only one end point). At the end point $x = 0$ we have $f(0) = 0$.

So there are only three local extrema. Because f goes to $+\infty$ as $x \to +\infty$, there is no global maximum. To find the local minimum we need only compare the values $(2,4)$ and $(0,0)$. Thus the global minimizer on this feasible set is $x = 0$, and the minimum value $y = 0$. These findings are reflected on the graph.



Let's apply the second derivative test in Example 3.3.9 .

3.3.10 Example. Since f is twice differentiable, we can take the second derivative

$$f''(x) = 12x - 18$$

Applying the second-derivative test, we evaluate $f''(x)$ at the critical points:

$$f''(1) = -6$$
$$f''(2) = 6$$

Thus Theorem 3.3.5 tells us that f has a local maximum at $x = 1$ and a local minimum at $x = 2$, confirming our previous argument.

Next consider the simplest example of an inflection point , as illustrated by the function $f(x) = x^3$. In this case, $f'(0) = f''(0) = 0$, so at $x = 0$ the necessary conditions are satisfied. The sufficiency conditions are not satisfied, however, and in fact we do not have a minimum. This is illustrated by a graph of $f(x) = x^3$:

3.3.11 Example. We start with the general affine function

$$f(x) = ax + b$$

on the interval $[\alpha, \beta]$. Since the constant b just shifts the graph up or down, we can take $b = 0$. We assume $a \neq 0$, since otherwise we have an uninteresting constant function. $f(x)$ has no critical points, so the minimum must occur on a boundary of the interval $[\alpha, \beta]$. If $a > 0$ (positive slope) the minimum occurs at α; if $a < 0$ (negative slope) the minimum occurs at β. Lectures 19 through 27 will be devoted to generalizing the linear case to a larger number of variables. This is known as *linear programming*.

3.3.12 Example. Now we consider the set of all possible quadratic polynomials:

$$f(x) = ax^2 + bx + c$$

where a, b, c are constants, and where the feasible set D is the set of $x \geq 0$. As with Example 3.3.11, the constant c just shifts the graph of f up or down, so we might as well take $c = 0$. If a were equal to 0, we would end up with the linear case presented in Example 3.3.11, while $a < 0$ would mean that f tends to $-\infty$ as $x \to \pm\infty$ and not have a finite minimum. Since we are looking for a minimum, we take $a > 0$. Dividing by a, we need only consider $f(x) = x^2 + dx$ (where $d = b/a$). Setting the first derivative $f'(x) = 2x + d$ equal to zero produces the equation $-2x = d$. The unique critical point, therefore, occurs at $x = -d/2$. It is necessarily a minimum. However $x = -d/2$ is only in the feasible set if $d \leq 0$. What happens if $d > 0$? The minimum occurs at the end point $x = 0$ of the feasible set.

The key feature of this example is that the coefficient of x^2 is positive. The appropriate generalization in several variables is the notion of positive definite symmetric quadratic form. We will study this topic in detail in Lecture 13, as it is a key ingredient in nonlinear optimization, also known as *nonlinear programming*.

We conclude with some more difficult examples

3.3.13 Example. In this example we use the fact that

$$\lim_{x \to 0} x^n e^{-\frac{1}{x^2}} = 0 \quad \text{for all } n \geq 0.$$

The function

$$f(x) = \begin{cases} e^{-\frac{1}{x^2}}, & \text{if } x \neq 0; \\ 0, & \text{if } x = 0; \end{cases}$$

is infinitely differentiable, even at $x = 0$, and $f^{(n)}(0) = 0$ for all n, so we cannot apply Theorem 3.3.6. Because $f(x)$ is even, it has an extremum at 0, and because the values away from 0 are positive, it is a minimum. For more details see [12], §3.1.

3.3.14 Exercise. Find the minima of $f(x) = x^x$ on the interval $(0, \infty)$. Be sure to determine that your answer is the global minimum. Does the function have local maxima or a global maximum?

Hint: recall from calculus that you can write $x^x = e^{x \ln x}$, so that $f'(x) = x^x(1 + \ln x)$. Note that the first term in this product is always positive.

3.4 The Intermediate Value Theorem

The following theorem, due to Gaston Darboux[4], allows us to prove that certain functions are continuously differentiable: also see Theorem 21.2.17. We give an example 3.4.5 of a function that is not continuously differentiable.

3.4.1 Definition. A real-valued function $f(x)$ satisfies the *intermediate value property* on an open interval S if for all $a < b$ in S, and every v in the open interval bounded by $f(a)$ and $f(b)$, there is a x_0 in the open interval (a, b) with $f(x_0) = v$. If $f(a) = f(b)$, the property is vacuous.

3.4.2 Proposition. *If $f(x)$ increases on the open interval S, and satisfies the intermediate value property on $[a, b]$, for any $a < b$ in S, then f is continuous.*

Proof. We prove continuity of f at any $x_0 \in S$. We will first prove f is continuous as x approaches x_0 from below, and let $f_-(x_0)$ be that limit. A similar argument shows it is continuous from above, and we let $f_+(x_0)$ be that limit. Since f is increasing, $f_-(x_0) \leq f_+(x_0)$, and the intermediate value property forces $f_-(x_0) = f_+(x_0)$, so we are done.

To prove f is continuous as x approaches x_0 from below, we look for a $c \in S$ such that $f(c) < f(x_0)$. If there is no such c, $f(x)$ is constant for $x \leq x_0$, so it is continuous from below. If there is such a c, then for any positive $\epsilon < f(x_0) - f(c)$, the intermediate value property allows us to find a point $c_1 \in (c, x_0)$ such that $f(x_0) - f(c_1) < \epsilon$. Then, since f is increasing, every point $x \in [c_1, x_0]$ satisfies $f(x) - f(c_1) < \epsilon$, so we have established continuity from below. □

3.4.3 Theorem (The Intermediate Value Theorem)**.** *Let $f(x)$ be a differentiable function on the open interval S. Then the derivative $f'(x)$ has the intermediate value property on S.*

So the derivative of f, even though it is not assumed to be continuous, takes on all intermediate values between any two values.

[4] See [12], p. 111 for historical details.

Proof. Pick two distinct point a and b in S, and an arbitrary value v with $f'(a) < v < f'(b)$. We must show there is a x_0 such that $f'(x_0) = v$. For concreteness assume $a < b$. Let $g(x) = f(x) - vx$. Then $g'(a) = f'(a) - v$ is negative, and $g'(b) = f'(b) - v$ is positive. By Theorem 3.1.14, g is strictly decreasing at a and strictly increasing at b, so that the minimum of the continuous function g on the interval $[a, b]$, which exists by the Weierstrass Theorem 3.1.6, must occur on the open interval (a, b). Thus the hypotheses of Theorem 3.1.15 are satisfied, so at the minimizer x_0 for g, we have $g'(x_0) = 0$, so $f'(x_0) = v$. □

3.4.4 Corollary. *Assume $f(x)$ is differentiable and $f'(x)$ is increasing on the open interval S, so that if $a < b$ in S, $f'(a) \leq f'(b)$. Then $f'(x)$ is continuous on S.*

Proof. By Theorem 3.4.3, $f'(x)$ has the intermediate value property. Therefore, since $f'(x)$ is increasing by hypothesis, Proposition 3.4.2 says that f' is continuous. □

This result implies that a derivative cannot have simple 'jump' discontinuities.[5] We will use this result in Theorem 21.2.17 to show that a convex functions $(f(x)$ that are differentiable is continuously differentiable.

This result makes it difficult to produce examples of functions that are differentiable but not continuously differentiable. Here is such an example.

3.4.5 Example. Let α and β be positive real numbers. Consider the function

$$f(x) = \begin{cases} x^\alpha \sin 1/x^\beta, & \text{if } x \neq 0; \\ 0, & \text{if } x = 0. \end{cases}$$

We compute the derivative away from $x = 0$ using the product rule and the chain rule to get

$$f'(x) = \alpha x^{\alpha-1} \sin 1/x^\beta - \beta x^{\alpha-\beta-1} \cos 1/x^\beta \tag{3.4.6}$$

We compute the derivative at 0 directly by using the Newton quotient:

$$f'(0) = \lim_{x\to 0} \frac{x^\alpha \sin 1/x^\beta}{x} = \lim_{x\to 0} x^{\alpha-1} \sin 1/x^\beta. \tag{3.4.7}$$

This limit only exists (and is then equal to 0) if $\alpha > 1$, so that the oscillations of $\sin 1/x^\beta$ are damped out. To show that it is continuously differentiable at 0 we need to show that

$$f'(0) = \lim_{x\to 0} f'(x) \tag{3.4.8}$$

[5] These are defined and called *discontinuities of the first kind* in [55] 4.26.

Comparing (3.4.6) and (3.4.7), this means

$$\lim_{x\to 0} x^{\alpha-\beta-1} \cos 1/x^{\beta} = 0 \tag{3.4.9}$$

and this is true if and only if $\alpha - \beta > 1$. We only required $\alpha > 1$ to get continuity, so this is a new condition.

So for any $\alpha > 1$ and $\beta \geq \alpha - 1$ we get a continuous function that is not continuously differentiable. [6]

[6] See [55], exercise 13, p.115 for a more general exercise, and [70], p. 150 for a discussion. A more elementary reference for the case $\alpha = \beta = 2$ is [12], p. 65.

Lecture 4

Taylor's Theorem in One Variable

In this chapter we review Taylor's Theorem on one variable. This will become an essential tool when we get to the multivariable case, so we give a full account. Our results are based on the Mean Value Theorem 3.2.1. From the mean value theorem, we first get an extended mean value theorem, from which we deduce Taylor's theorem in one variable.[1]

4.1 The Taylor Polynomial in One Variable

First recall what it means for a function to be $\mathcal{C}^k$ in Definition 3.3.1.

4.1.1 Definition. Given a function $f(x)$ of one variable, we usually write $f'(x)$ for the first derivative and $f''(x)$ for the second derivative. In general we use $f^{(k)}$ to denote the k-th derivative of $f(x)$, with the convention that the 0-th derivative $f^{(0)}$ stands for f itself: this is only used in summations.

We start with a one-variable function $f(x)$ defined in the neighborhood of a point a in its domain. We assume that f is N-times-differentiable for some integer $N > 0$.

4.1.2 Definition. For any $n \leq N$, the polynomial

$$P_n(a,x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k \tag{4.1.3}$$

[1] Other references for this material are [55], p. 110, [70], and [74], §5.3

is called the *Taylor polynomial* of degree n of f centered at a. We sometimes suppress the a and write just $P_n(x)$.

The special case where $a = 0$ is called the *Maclaurin* polynomial of degree n of f:

$$P_n(0, x) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!} x^k \tag{4.1.4}$$

Since $P_n(a, x)$ is a polynomial in x of degree $\leq n$, we can easily compute all its derivatives.

4.1.5 Proposition. *When $k \leq n$, the k-th derivative of $P_n(a, x)$ takes the following values:*

$$\begin{aligned} P_n^{(k)}(a, x) &= \sum_{j=k}^{n} \frac{f^{(j)}(a)}{(j-k)!}(x-a)^{j-k} \\ P_n^{(k)}(a, a) &= f^{(k)}(a) \end{aligned}$$

Thus when $k = n$,

$$P_n^{(n)}(a, x) = f^{(n)}(a).$$

When $k > n$, $P_n^{(k)}(a, x) = 0$.

Proof. The first equation is a little exercise in computing the derivative of $(x-a)^j$ and manipulating factorials. Recall that $0! = 1$, and $a^0 = 1$ for any real number a. The rest is clear since $P_n(a, x)$ is a polynomial of degree at most n in x. □

The whole point of defining the Taylor polynomial as above is to obtain the value $f^{(k)}(a)$ for the k-th derivative of $P_n(a, x)$ at a, for $k \leq n$. So f and $P_n(a, x)$ have the same first n derivatives (and same value) at a.

4.1.6 Exercise. Compute the Maclaurin polynomial of any order for the polynomial $f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$.

4.1.7 Example. Take the well-known Maclaurin expansion of $f(x) = \exp x$. Since $f^{(n)}(0) = 1$ for all $n \geq 0$, the Maclaurin polynomial of degree n is

$$P_n(x) = \sum_{k=0}^{n} \frac{1}{k!} x^k.$$

You can easily verify all the assertions of Proposition 4.1.5 in this case. For example,

$$P_n'(x) = \sum_{k=1}^{n} \frac{1}{(k-1)!} x^{k-1}, \text{ and } P_n'(0) = 1.$$

Next we introduce the remainder

$$r_n(a,x) = f(x) - P_n(a,x) \tag{4.1.8}$$

We often write $r_n(x)$ instead of $r_n(a,x)$, to lighten the notation. Using the linearity of differentiation we get the following corollary of Proposition 4.1.5

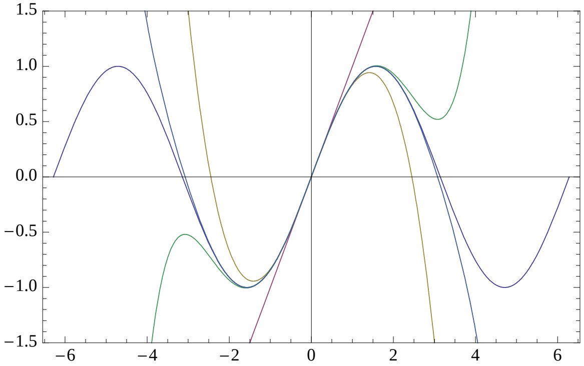
4.1.9 Corollary. *For all $k \leq n$, we have $r_n^{(k)}(a) = 0$. For $k > n$, $r_n^{(k)}(x) = f^{(k)}(x)$. Since $f(x)$ is assumed to be n times differentiable, so is $r_n(a,x)$.*

These derivative computations will be useful in the next section.

4.1.10 Example. Let $f(x) = \sin x$, and $a = 0$. Then

$$\begin{aligned}
P_1(x) &= x\\
P_3(x) &= x - x^3/3!\\
P_5(x) &= x - x^3/3! + x^5/5!\\
P_7(x) &= x - x^3/3! + x^5/5! - x^7/7!
\end{aligned}$$

The graph shows the polynomials P_1, P_3, P_5, and P_7 approximating $\sin x$ more and more accurately around 0.



4.1.11 Exercise. Let $f(x) = a_0 + a_1x + a_2x^2 + a_3x^3$. Compute the Taylor polynomial of f of degree 4 (meaning $n = 4$) centered at the point b.

We next ask: how good an approximation is $P_n(a,x)$ for $f(x)$ in a neighborhood of a?

4.2 A Generalized Mean Value Theorem

We first prove a generalization of Rolle's Theorem 3.2.3.

4.2.1 Theorem. *Let $g(x)$ be a real-valued function that is $\mathcal{C}^{n-1}$ on the closed interval $[a, b]$, and whose n-th derivative exists on the open interval (a, b). Further assume that*

$$g^{(k)}(a) = 0, \quad \textit{for } 0 \leq k < n, \quad \textit{and } g(b) = 0.$$

Then there is a sequence of points $b > c_1 > \cdots > c_n > a$, such that $g^{(k)}(c_k) = 0$, $1 \leq j \leq n$.

Proof. The case $n = 1$ is exactly Rolle's theorem, so we get a point $c_1 \in (a, b)$ with $g'(c_1) = 0$. Since $g'(a) = 0$ also, we can apply Rolle's lemma to the function $g'(x)$ on the interval $[a, c_1]$. This yields a point $c_2 \in (a, c_1)$ with $g''(c_2) = 0$. Continuing inductively in this fashion until $k = n - 1$, we see that on the interval $[a, c_k]$ the hypothesis of Rolle are satisfied for the function $g^{(k)}$, so that there is a $c_{k+1} \in [a, c_k]$ where $g^{(k+1)}(c_{k+1}) = 0$, and the result is proved. □

Note that there is an analogous lemma obtained by interchanging the role of a and b. We trust the reader to formulate it.

We use the generalized Rolle's theorem to prove a generalized mean value theorem. Recall that $P_k(a, x)$ is the k-th Taylor polynomial of f centered at a, and $r_k(a, x)$ is the remainder.

4.2.2 Theorem (Generalized Mean Value Theorem)**.** *Let $f(x)$ be a real-valued function that is $\mathcal{C}^{n-1}$ on the closed interval $[a, b]$, and whose n-th derivative exists on the open interval (a, b). Then there is a point $c_n \in (a, b)$ such that*

$$f(b) = P_{n-1}(a, b) + \frac{f^{(n)}(c_n)}{n!}(b - a)^n$$

Proof. In order to apply the generalized Rolle's theorem, we replace f by

$$g(x) = f(x) - P_{n-1}(a, x) - \frac{r_{n-1}(a, b)}{(b - a)^n}(x - a)^n. \tag{4.2.3}$$

A derivative computation shows that $g^{(k)}(a) = 0$, $0 \leq k < n$, and $g(b) = 0$, so that Rolle's Theorem 4.2.1 applies, so we can find a point $c_n \in (a, b)$ with

$$g^{(n)}(c_n) = 0 \tag{4.2.4}$$

We differentiate (4.2.3) n times:

$$g^{(n)}(x) = f^{(n)}(x) - \frac{r_{n-1}(a, b)}{(b - a)^n} n!$$

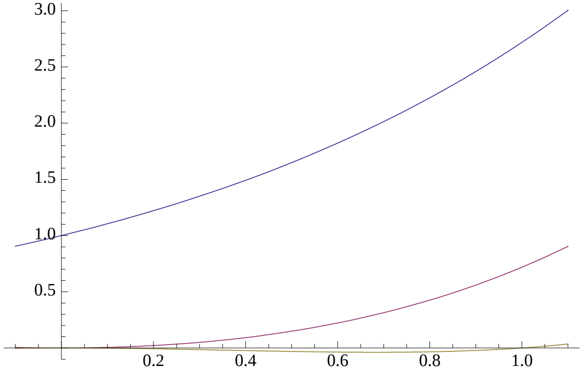
and evaluate at $x = c_n$ using (4.2.4) to get

$$f^{(n)}(c_n) = \frac{r_{n-1}(a,b)}{(b-a)^n} n! \quad \text{or} \quad \frac{r_{n-1}(a,b)}{(b-a)^n} = \frac{f^{(n)}(c_n)}{n!}$$

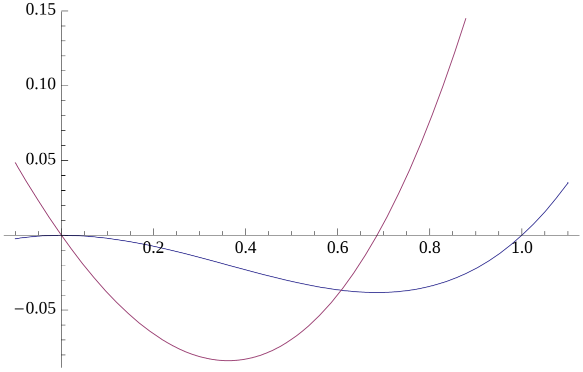
Substitute the left-hand side of the last expression into (4.2.3) and evaluate at $x = b$ to get the result. □

This result is sometimes called Taylor's theorem (for example [55], 5.15) and sometimes the Lagrange form of the remainder of Taylor's theorem (for example [70], §5.4.4). For a more elementary reference see [12], §3.7, where you can also find the Cauchy form of the remainder.

4.2.5 Example. Let's see how the theorem works with $f(x) = e^x$, $a = 0$, $b = 1$ and $n = 2$. Then $P_1(0,x) = 1 + x$, so $r_1(0,1) = f(1) - P_1(0,1) = e - 2$. So $g(x) = e^x - 1 - x - (e-2)x^2$. Note that $g(0) = g'(0) = g(1) = 0$ as expected. Here is the graph for the three functions e^x, $e^x - 1 - x$, and $g(x)$, with $g(x)$ barely distinguishable from the x-axis, but crossing it at $x = 1$:



Now let's plot $g(x)$ and $g'(x)$ magnifying the scale on the vertical axis.



The point c_1 is the intersection of $g'(x)$ with the x-axis (at about 0.69) and the point c_2 is the minimum of $g'(x)$ (at about 0.36).

4.2.6 Remark. There is a nice linear algebra interpretation of the theorem. Consider the set of all polynomials of degree $\leq n$ with real coefficients. This is a vector space of dimension $n+1$. For any number a, the polynomials 1, $x-a$, $(x-a)^2$, $\dots$, $(x-a)^n$ form a basis of this vector space. Write a general polynomial Q in this vector space with the variables z_0, z_1, $\dots$, z_n as coefficients:

$$\sum_{i=0}^{n} z_i(x-a)^i$$

The key idea is that differentiation with respect to x and evaluation at a point imposes a linear condition on the coefficients z_i. For example, saying that the k-th derivative of Q at a is equal to the k-th derivative of f at a is saying that $k!z_k = f^{(k)}(a)$. Furthermore saying that the value of Q at b is equal to $f(b)$ means that

$$\sum_{i=0}^{n} z_i(b-a)^i = f(b)$$

Again, this is a linear condition on the variables z_k. In the generalized MVT we modify f by elements in this vector space in order to satisfy $n+1$ linear conditions, which in fact are independent, meaning that the $(n+1)\times(n+1)$ matrix of coefficents has rank $n+1$. Therefore elementary linear algebra says the conditions can be satisfied uniquely.

Here is one of the main applications of Theorem 4.2.2.

4.2.7 Corollary. *If the n-th derivative of f, $f^{(n)}(x)$, is bounded in absolute value by M on the interval I given by $|x-a| \leq r$, for some positive r, then the difference $|f(x)-P_{n-1}(x)|$ is bounded by $\frac{M}{n!}r^n$ on I.*

Proof. Pick an x in I; it will play the role of b in the theorem. Since c_n is between a and b, the bound for $\left|f^{(n)}(x)\right|$ applies at c_n. So

$$|f(x)-P_{n-1}(x)| = \left|\frac{f^{(n)}(c_n)}{n!}(x-a)^n\right| \leq \left|\frac{f^{(n)}(c_n)}{n!}\right| |(x-a)^n| \leq \frac{M}{n!}r^n$$

as required. □

If n is large and r is small, this expression becomes very small. This shows that the Taylor polynomial can be a good approximation of f. We will quantify this is a neighborhood of a in the next section.

4.2.8 Example. Take $f(x) = \sin x$ and $a = 0$. Since all the derivatives of sine are bounded by 1, the Taylor polynomials approximate f quickly.

4.2.9 Exercise. The convergence is not quite so fast with e^x at $a = 0$, but we still get good convergence if the interval is small enough. Compute the first n terms of the Taylor polynomial of $f(x) = e^x$ at $a = 0$ on the interval $[-1, 1]$. What kind of approximation to f do you get?

4.2.10 Example. We compute the Taylor expansion for the natural logarithm function $\ln x$ around the point $x = 1$. Write $x = 1 + y$. We will mainly look at the case $y \geq 0$. Because the derivative of $\ln$ is $1/x$, we get the following formula, for y such that $|y| < 1$, for the Taylor polynomial of degree $n - 1$ of ln:

$$P_{n-1}(1+y) = y - \frac{y^2}{2} + \frac{y^3}{3} - \frac{y^4}{4} + \cdots + \frac{(-1)^n y^{n-1}}{n-1} \tag{4.2.11}$$

How good is this approximation? The error on the interval $1 \leq x \leq 1 + r$ is by Corollary 4.2.7 bounded by Mr^n/n, where M is the maximum of the n-th derivative of $\ln$ on our interval. Because the absolute value $1/x^n$ of the n-th derivative is a decreasing function, this error is maximum at the left end point $y = 0$, so $M = 1$ and the remainder is bounded by r/n.

In the next example we look at the case $n = 2$.

4.2.12 Example (Rule of 70 for doubling time)**.** Suppose you invest S dollars at p percent a year, compounded once a year. Then your money doubles in approximately $70/p$ years.

To prove this, let $y = p/100$ and apply the previous exercise. At the end of t years, you have $S(1 + y)^t$ dollars. For any y we want to solve for t in:

$$S(1+y)^t = 2S, \quad \text{or} \quad (1+y)^t = 2$$

Take the natural logarithm on both sides to get

$$t \ln{(1+y)} = \ln 2 = 0.693147...$$

As the previous example shows, for y small enough and positive, a reasonable approximation for $\ln{(1 + y)}$ is y. Because the Taylor series of $\ln$ alternates with decreasing terms as per (4.2.11), this is an overestimate. All in all we see that the doubling time is underestimated by $0.7/y = 70/p$, hence the rule of 70. In particular, at 7 percent your money doubles in roughly 10 years.

4.3 Taylor's Theorem in One Variable

We prove a version of Taylor's theorem that follows simply from the generalized mean value theorem 4.2.2. Unlike that theorem, Taylor's theorem is a purely local result. As usual $P_n(a, x)$ is the n-th Taylor polynomial centered at a, and $r_n(a, x)$ the remainder, so that

$$f(x) = P_n(a, x) + r_n(a, x) \tag{4.3.1}$$

We start by assuming that $f(x)$ is $\mathcal{C}^n$ in a neighborhood of the point a, so our hypothesis is the same as that in Theorem 4.2.2, given that we are now taking P_n instead of P_{n-1}.

4.3.2 Theorem. *With this hypothesis,*

$$\lim_{x \to a} \frac{r_n(a, x)}{(x - a)^n} = 0$$

Proof. We only treat the case $x > a$. For any $x \neq a$, we proved in Theorem 4.2.2 that there is a c_x, $a < c_x < x$, such that

$$f(x) = P_{n-1}(a, x) + \frac{f^{(n)}(c_x)}{n!}(x - a)^n \tag{4.3.3}$$

Subtract (4.3.3) from (4.3.1) to get

$$0 = \frac{f^{(n)}(a)}{n!}(x - a)^n + r_n(a, x) - \frac{f^{(n)}(c_x)}{n!}(x - a)^n$$

so

$$r_n(a, x) = \frac{f^{(n)}(c_x) - f^{(n)}(a)}{n!}(x - a)^n$$

Divide by $(x - a)^n$.

$$\frac{r_n(a, x)}{(x - a)^n} = \frac{f^{(n)}(c_x) - f^{(n)}(a)}{n!}$$

As x approaches a, c_x also approaches a, since it is between a and x. The hypothesis that f is $\mathcal{C}^n$ near a tells us that $f^{(n)}$ is continuous, so $\lim_{c_x \to a} f^{(n)}(c_x) = f^{(n)}(a)$, and this shows:

$$\lim_{x \to a} \frac{r_n(a, x)}{(x - a)^n} = \lim_{x \to a} \frac{f^{(n)}(c_x) - f^{(n)}(a)}{n!} = 0$$

and the theorem is proved. □

Not surprisingly we can get the same theorem with a slightly weaker hypothesis, at the cost of a little more work. The key step is the following proposition.

4.3.4 Proposition. *Let $h(x)$ be n-times differentiable at the point a. Furthermore assume that $h(a) = h'(a) = \dots = h^{(n)}(a) = 0$. Then*

$$\lim_{x \to a} \frac{h(x)}{(x-a)^n} = 0$$

Proof. First note that the case $n = 1$ is a special case of (3.1.9), which in turn is just a rewording of the definition of the derivative . This allows us to start the proof by induction on n at $n = 1$. So we may now assume $n \geq 2$. Next note that since $h^{(n)}(a)$ exists, each previous derivative $h^{(k)}(x)$, $1 \leq k < n$ must be defined on an open neighborhood U_k of a. So we work on the open intersection U of these neighborhoods. We continue the induction by assuming the result is true for all functions satisfying the hypotheses for a given $n - 1$, and showing it is true for n. Consider the function $h'(x)$: it is $n - 1$ times differentiable, and its first $n - 1$ derivatives at a are zero, since $(h'(x))^{(k)} = h^{(k+1)}(x)$. So, by induction,

$$\lim_{x \to a} \frac{h'(x)}{(x-a)^{n-1}} = 0$$

Now we apply the mean value theorem to h on the interval $[a, x]$. For concreteness we assume $a < x$. Then there is a c, $a < c < x$ such that

$$h(x) = h(x) - h(a) = h'(c)(x - a)$$

So

$$\left|\frac{h(x)}{(x-a)^n}\right| = \left|\frac{h'(c)}{(x-a)^{n-1}}\right| < \left|\frac{h'(c)}{(c-a)^{n-1}}\right|$$

Now take the limit as x tends to a. Since c is always between a and x, c tends to a too, so

$$\lim_{x \to a} \left|\frac{h(x)}{(x-a)^n}\right| \leq \lim_{c \to a} \left|\frac{h'(c)}{(c-a)^{n-1}}\right| = 0$$

by the induction hypothesis, and we are done. □

Note that in the proof of Theorem 4.3.2, we used the fact that $f(x)$ is $\mathcal{C}^n$. Proposition 4.3.4 shows that we do not need this, so we formulate our final version of Taylor's theorem with the weaker hypothesis.

4.3.5 Theorem (Taylor's Theorem)**.** *Assume that $f(x)$ is n-times differentiable at the point a. Then*

$$\lim_{x \to a} \frac{r_n(a, x)}{(x-a)^n} = 0$$

Furthermore $P_n(a,x)$ is the unique polynomial Q of degree less than or equal to n such that

$$\lim_{x\to a} \frac{f(x)-Q(x)}{(x-a)^n} = 0$$

Proof. We can apply Proposition 4.3.5 to $h(x) = r_n(x)$ by using the derivative computations of Corollary 4.1.9, so the result follows immediately. If you let c_i be the coefficient of $(x-a)^i$ in Q, then we can solve inductively for the c_i by noting that $c_0 = f(a)$ and

$$c_{k+1} = \lim_{x\to a} \frac{f(x) - \sum_{i=0}^{k} c_i(x-a)^i}{(x-a)^{k+1}}$$

This establishes the uniqueness. □

This is the promised generalization of Theorem 3.1.11.

A similar argument, using Cauchy's Mean Value Theorem 3.2.5 gives the following version of L'Hospital's rule:

4.3.6 Theorem. *Let $f(x)$ and $g(x)$ be real-valued functions that are n-times differentiable at the point a. Assume that for some $n > 0$, $f(a) = f'(a) = \cdots = f^{(n-1)}(a) = 0$, $g(a) = g'(a) = \cdots = g^{(n-1)}(a) = 0$, and $g^{(n)}(a) \neq 0$. Then*

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{f^{(n)}(a)}{g^{(n)}(a)}$$

The proof is left as an exercise.

Part III

Linear Algebra

Lecture 5

Real Vector Spaces

Now we review the basic results on the vector space structure $\mathbb{R}^n$. Most of these concepts are covered in multivariable calculus courses, and certainly in linear algebra courses, but sometimes only for $\mathbb{R}^2$ and $\mathbb{R}^3$. The challenge is to generalize these results to $\mathbb{R}^n$ for any positive integer n. This challenge is twofold. First, you must master an expanded notation—namely, summations and double summations—that allows you to represent objects in $\mathbb{R}^n$. Second, you must learn to think geometrically about spaces of dimension higher than 3, for which visuospatial intuition is impossible. The central result is the Cauchy-Schwarz Inequality 5.4.6, sometimes referred to as the most important inequality in mathematics.

5.1 Real Vector Spaces

We summarize the basic facts about the vector space structure of $\mathbb{R}^n$. These generalize the situation in $\mathbb{R}^2$ and $\mathbb{R}^3$ familiar from multivariable calculus (see [63], §12.1-3), as is done in any linear algebra course. We continue with this in Chapter 6. So we start with a definition:

5.1.1 Definition. For any positive integer n, $\mathbb{R}^n$ is the space of all ordered n-tuples of real numbers. A point in $\mathbb{R}^n$ is one such n-tuple.

So $(1.2, \pi, -3, 1)$ is a point in $\mathbb{R}^4$, and $(\pi, 1.2, 1, -3)$ is a different point. We often call points *vectors*, imagining them as directed line segments between the *origin* $(0, 0, \dots, 0)$ and the point.

The space $\mathbb{R}^n$ is a *vector space*, meaning that two operations are defined on it, vector addition and scalar multiplication:

Addition. Adding vectors is accomplished by adding their coordinates. If $\mathbf{x} =$

$(x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, then

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, \ldots, x_n + y_n)$$

and

$$(1.2, \pi, -3, 1) + (\pi, 1.2, 1, -3) = (1.2 + \pi, 1.2 + \pi, -2, -2)$$

Scalar Multiplication. Scalar multiplication of a vector $\mathbf{x}$ by a real number c is given by $c\mathbf{x} = (cx_1, cx_2, \ldots, cx_n)$. The real number c is often called a *scalar*, hence the name of the operation.

Scalar multiplication distributes over addition. So

$$-2(1.2, \pi, -3, 1) = (-2.4, -2\pi, 6, -2)$$

5.1.2 Definition. The *coordinate vectors* (or *basis vectors*) in $\mathbb{R}^n$ are the n vectors $\mathbf{e}_i$, $1 \le i \le n$, with 0 in all positions except the i-th position, where it has 1.

For example, $\mathbf{e}_1 = (1, 0, \ldots, 0)$, $\mathbf{e}_2 = (0, 1, 0, \ldots, 0)$, and $\mathbf{e}_n = (0, \ldots, 0, 1)$. Note that $\mathbb{R}^n$ has n coordinate vectors (see [63], §12.2 or [68], §3.1). This gives us a second way of writing a vector. For example

$$(1.2, \pi, -3, 1) = 1.2\mathbf{e}_1 + \pi\mathbf{e}_2 - 3\mathbf{e}_3 + \mathbf{e}_4$$

5.2 Basis and Dimension

Take an arbitrary vector space V, meaning a set admitting addition and scalar multiplication with the properties described in §5.1. Then

5.2.1 Definition. Vectors $\mathbf{e}_1$, $\mathbf{e}_2$, …, $\mathbf{e}_n$ in a vector space V are *linearly dependent* if there exist real numbers λ_1, λ_2, …, λ_n, not all zero, such that

$$\sum_{j=1}^{n} \lambda_j \mathbf{e}_j = \mathbf{0}.$$

Otherwise they are *linearly independent*. A *basis* for V is a linearly independent set $\{\mathbf{e}_i\}$, $1 \le i \le n$, that spans V, meaning that every element $\mathbf{v}$ in V can be expressed as a linear combination of the $\{\mathbf{e}_j\}$:

$$\mathbf{v} = \sum_{j=1}^{n} x_j \mathbf{e}_j,$$

for uniquely determined real numbers x_j called the *coordinates* of $\mathbf{v}$ relative to this basis. An important theorem in linear algebra proves that any two bases of V have the same number of elements, allowing us to define the *dimension* of V: it is is the number of elements in the basis, n in this case[1].

Then it follows that any real vector space V of dimension n is just $\mathbb{R}^n$, by mapping a point $\mathbf{v}$ to its n-tuple of coordinates $(x_1, \dots, x_n)$.

5.3 Linear Subspaces

A linear subspace W of a vector space V is a subset that is a vector space in its own right, using the operations of V. We also just call is a subspace. To check that W is a subspace, we must show that it is *closed* under the operations of V. In other words,

5.3.1 Definition. A subset W of the vector space V is a subspace of V if

1. For all $\mathbf{v}$ and $\mathbf{w}$ in W, $\mathbf{v} + \mathbf{w}$ is in W;
2. For all real numbers a and all $\mathbf{w} \in W$, then $a\mathbf{w}$ is in W.

This implies that $\mathbf{0}$ is in W, since $\mathbf{0} = 0\mathbf{w}$, for any $\mathbf{w} \in W$.

Note that the vector space consisting just of the origin is a subspace of any vector space. The space V is a subspace of itself. We call both of the subspaces the trivial subspaces of V.

5.3.2 Example. Check that the following subsets are actually subspaces.

- The subset of all triples in $\mathbb{R}^3$ where the last entry is 0: $(v_1, v_2, 0)$.
- The subset of all n-tuples in $\mathbb{R}^n$ where the last entry is 0: $(v_1, \dots, v_{n-1}, 0)$.

5.3.3 Example. In the vector space of polynomials in t over F, consider the subset P_k of polynomials of degree at most k, for any integer k. Show P_k is a subspace of the vector space of polynomials over F. Explain why the polynomials of degree exactly n do not form a subspace.

Because a subspace W in a vector space V is a subspace in its own right, it has a dimension. It is easy to prove that this dimension is no greater than the dimension of V. Moreover if its dimension is equal to the dimension of V, it is V.

[1] We will make parallel definitions later in these lectures for affine independence, convex independence and conical independence.

5.3.4 Proposition. *If U and W are both subspaces of the vector space V, then $U \cap W$ is a subspace of V.*

Proof. This is elementary set theory. If $\mathbf{u}$ is in $U \cap W$, then $\mathbf{u}$ is both in U and in W. Since U is a subspace, $c\mathbf{u}$ is in U for every real number c; since V is a subspace, $c\mathbf{u}$ is in W for every real number c. So $c\mathbf{u}$ is in $U \cap W$.

If $\mathbf{u}$ and $\mathbf{v}$ are in $U \cap W$, then $\mathbf{u}$ is both in U and in W, and $\mathbf{v}$ is both in U and in W. So $\mathbf{u} + \mathbf{v}$ in in U, because U is a subspace, and it is also in W, because W is a subspace. Thus $\mathbf{u} + \mathbf{v}$ is in $U \cap W$. □

5.4 The Distance between Points in Real Vector Spaces

After the vector space structure on $\mathbb{R}^n$ we can now define the *inner product*, the *norm*, and the distance between two points. This last concept will be key.

5.4.1 Definition. These definitions are familiar from any multivariable calculus course:

- The *inner product* on $\mathbb{R}^n$, also called the *dot product* or the *scalar product*, is given by:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i \tag{5.4.2}$$

We sometimes write $\mathbf{x} \cdot \mathbf{y}$ for $\langle \mathbf{x}, \mathbf{y} \rangle$, hence the name dot product. The inner product satisfies the following three properties:

1. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$;
2. $\langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$;
3. $\langle a\mathbf{x}, \mathbf{y} \rangle = a\langle \mathbf{x}, \mathbf{y} \rangle$.

Indeed, these three properties can be used to define an inner product.

5.4.3 Exercise. Show that these properties imply

- $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$;
- $\langle \mathbf{x}, b\mathbf{y} \rangle = b\langle \mathbf{x}, \mathbf{y} \rangle$.

- The *norm*, sometimes called the *length*, of the vector $\mathbf{x}$ is:

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{x_1^2 + \cdots + x_n^2}. \tag{5.4.4}$$

The norm is written with a symbol similar to that of the absolute value, which makes sense because the norm is the absolute value when $n = 1$.

- The *distance* $d(\mathbf{x},\mathbf{y})$ between the two points $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$ is the norm of the difference vector:
$$d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| \tag{5.4.5}$$

5.4.6 Theorem (The Cauchy-Schwarz inequality)**.** *For any two vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *in* $\mathbb{R}^n$*,*
$$|\mathbf{x}\cdot\mathbf{y}| \le \|\mathbf{x}\|\|\mathbf{y}\| \tag{5.4.7}$$
with equality only if one vector is $\mathbf{0}$ *or if the vectors are proportional—namely,* $\mathbf{y} = c\mathbf{x}$ *for a scalar* c*.*

As J. Michael Steele says in [61], “there is no doubt that this is one of the most widely used and most important inequalities in all of mathematics”. Chapter 1 of that book is a good reference to its many proofs. We give a second proof using convexity arguments in Corollary 22.6.13.

In coordinates, the Cauchy-Schwarz inequality says that
$$|x_1y_1+\dots+x_ny_n| \le \sqrt{x_1^2+\dots+x_n^2}\sqrt{y_1^2+\dots+y_n^2}. \tag{5.4.8}$$

Proof. For any scalars a and b, we have:
$$0 \le (a\mathbf{x}+b\mathbf{y})\cdot(a\mathbf{x}+b\mathbf{y}) = a^2\mathbf{x}\cdot\mathbf{x} + 2ab\mathbf{x}\cdot\mathbf{y} + b^2\mathbf{y}\cdot\mathbf{y} \tag{5.4.9}$$

The 0 on the left comes from the fact that the right-hand side is the square of a norm. Now some magic: Let $a = \mathbf{y}\cdot\mathbf{y}$ and $b = -\mathbf{x}\cdot\mathbf{y}$. Substituting these values into (5.4.9), we get

$$\begin{aligned} 0 \le (\mathbf{y}\cdot\mathbf{y})^2\mathbf{x}\cdot\mathbf{x} - 2(\mathbf{y}\cdot\mathbf{y})(\mathbf{x}\cdot\mathbf{y})^2 + (\mathbf{x}\cdot\mathbf{y})^2(\mathbf{y}\cdot\mathbf{y}) \\ = (\mathbf{y}\cdot\mathbf{y})^2\mathbf{x}\cdot\mathbf{x} - (\mathbf{y}\cdot\mathbf{y})(\mathbf{x}\cdot\mathbf{y})^2 \end{aligned} \tag{5.4.10}$$

The theorem is trivial if $\mathbf{y}$ is the $\mathbf{0}$ vector ($\mathbf{0} \le \mathbf{0}$), so we assume that $(\mathbf{y}\cdot\mathbf{y})$ is non-zero. Since $(\mathbf{y}\cdot\mathbf{y})$ is a sum of squares by Definition 5.4.1, it is positive, and we can divide the inequality by $(\mathbf{y}\cdot\mathbf{y})$ and the inequality will not reverse directions:
$$0 \le (\mathbf{y}\cdot\mathbf{y})(\mathbf{x}\cdot\mathbf{x}) - (\mathbf{x}\cdot\mathbf{y})^2 \tag{5.4.11}$$

We move the last term on the right to the other side of the inequality:
$$(\mathbf{x}\cdot\mathbf{y})^2 \le (\mathbf{y}\cdot\mathbf{y})(\mathbf{x}\cdot\mathbf{x}) \tag{5.4.12}$$

and take the square root of both sides—which is legal because they are non-negative. The resulting inequality, $|\mathbf{x}\cdot\mathbf{y}| \le \|\mathbf{x}\|\|\mathbf{y}\|$, is the Cauchy-Schwarz inequality. □

In Definition 5.4.1 we provided $\mathbb{R}^n$ with a notion of distance. We now give a very general definition of distance that works on any set D.

5.4.13 Definition. If D is a set, a *distance function* on D is a function $d(\mathbf{x}, \mathbf{y})$ that assigns to any two elements $\mathbf{x}$ and $\mathbf{y}$ of D a real number satisfying the following four properties:

1. $d(\mathbf{x}, \mathbf{y}) \geq 0$.
2. $d(\mathbf{x}, \mathbf{y}) = 0$ implies that $\mathbf{x} = \mathbf{y}$. That is, the distance between two elements is 0 if and only if the points are the same.
3. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$. The distance between $\mathbf{x}$ and $\mathbf{y}$ is the same as the distance between $\mathbf{y}$ and $\mathbf{x}$.
4. For any three elements $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ in D, the triangle inequality
$$d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \tag{5.4.14}$$
is satisfied.

This allows us to define a general mathematical structure.

5.4.15 Definition. A *metric space* is a set D equipped with a distance function $d(\mathbf{x}, \mathbf{y})$.

5.4.16 Exercise. Show that any subset of a metric space is a metric space.

Because we are thinking geometrically, we often refer to the elements of D as points rather than elements.

We now show that $\mathbb{R}^n$ has the distance function in the above sense, derived from its vector space structure. Thus $\mathbb{R}^n$ is a metric space.

5.4.17 Theorem. *The distance $d(\mathbf{x}, \mathbf{y})$ on $\mathbb{R}^n$ is a distance function. Thus $\mathbb{R}^n$ is a metric space.*

Proof. We need to verify the four conditions of Definition 5.4.13. For (1),

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \sqrt{\langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle} = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

so that we have the square root of a sum of squares, which is non-negative.

For (2), we use a simple fact about real numbers that will serve us many times in this course: For a sum of squares to be zero, each one of the terms in the sum must be zero.

This forces the coordinates of the two points to be equal, which means the points are the same.

For (3), just note that $(x_i - y_i)^2 = (y_i - x_i)^2$.

For (4), we need to show that

$$\|\mathbf{x} - \mathbf{z}\| \leq \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{z}\|$$

This requires the Cauchy-Schwarz inequality 5.4.6 above. First we change variables in the triangle inequality, letting $\mathbf{u} = \mathbf{x} - \mathbf{y}$ and $\mathbf{v} = \mathbf{y} - \mathbf{z}$, so that $\mathbf{u} + \mathbf{v} = \mathbf{x} - \mathbf{z}$. The triangle inequality takes the simpler form:

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|.$$

Square the left-hand side:

$$\|\mathbf{u} + \mathbf{v}\|^2 = (\mathbf{u} + \mathbf{v}) \cdot (\mathbf{u} + \mathbf{v}) = \|\mathbf{u}\|^2 + 2\mathbf{u} \cdot \mathbf{v} + \|\mathbf{v}\|^2.$$

Now we use the Cauchy-Schwarz inequality to replace $2\mathbf{u} \cdot \mathbf{v}$ by the larger term $2\|\mathbf{u}\|\|\mathbf{v}\|$:

$$\|\mathbf{u}\|^2 + 2\mathbf{u} \cdot \mathbf{v} + \|\mathbf{v}\|^2 \leq \|\mathbf{u}\|^2 + 2\|\mathbf{u}\|\|\mathbf{v}\| + \|\mathbf{v}\|^2.$$

We recognize the right-hand side as the square of $\|\mathbf{u}\| + \|\mathbf{v}\|$, so we get

$$\|\mathbf{u} + \mathbf{v}\|^2 \leq (\|\mathbf{u}\| + \|\mathbf{v}\|)^2.$$

Taking the square root of both sides, and reverting to the original variables, we get the triangle inequality. □

5.4.18 Definition. $\mathbb{R}^n$ equipped with this distance function is called *Euclidean space* of dimension n, the distance function above is called the Euclidean distance, and its norm is called the Euclidean norm.

We can put many other distance functions on $\mathbb{R}^n$. We will not use any of them, so we leave it to you to show they are distance functions.

5.4.19 Example. The following three functions in $\mathbb{R}^n$ are all distance functions. In all cases, the proof is easier than for the Euclidean metric.

- The discrete metric:

$$d(\mathbf{x}, \mathbf{y}) = \begin{cases} 0, & \text{if } \mathbf{x} = \mathbf{y}; \\ 1, & \text{otherwise} \end{cases}$$

- The max metric:

$$d(\mathbf{x}, \mathbf{y}) = \max_i |x_i - y_i| .$$

- The L^1 metric

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i| .$$

5.4.20 Exercise. Show that the three functions in Example 5.4.19 are actually distance functions.

5.4.21 Definition. Two non-zero vectors $\mathbf{p}$ and $\mathbf{q}$ in $\mathbb{R}^n$ are *orthogonal* or *perpendicular* if $\mathbf{p} \cdot \mathbf{q} = 0$.

Note that the unit vectors $\mathbf{e}_i$ are mutually perpendicular: $\mathbf{e}_i \cdot \mathbf{e}_j = 0$ when $i \neq j$.

5.4.22 Exercise. Let $\mathbf{v}_1, \dots, \mathbf{v}_k$ be a collection of k non-zero mutually orthogonal vectors, meaning that $\mathbf{v}_i \cdot \mathbf{v}_j = 0$ for $i \neq j$. Then the $\mathbf{v}_i$ are linearly independent.

5.4.23 Exercise. Prove the Pythagorean Theorem: If $\mathbf{p}$ and $\mathbf{q}$ are orthogonal, then

$$\|\mathbf{p} + \mathbf{q}\|^2 = \|\mathbf{p}\|^2 + \|\mathbf{q}\|^2.$$

5.4.24 Exercise. Prove the parallelogram law: For any $\mathbf{p}$ and $\mathbf{q}$, then

$$\|\mathbf{p} + \mathbf{q}\|^2 + \|\mathbf{p} - \mathbf{q}\|^2 = 2\|\mathbf{p}\|^2 + 2\|\mathbf{q}\|^2.$$

5.4.25 Definition. Assume $\mathbf{q} \neq \mathbf{0}$. Then for any $\mathbf{p}$, there is a unique $c \in \mathbb{R}$ with $\mathbf{p} - c\mathbf{q}$ is orthogonal to $\mathbf{q}$. This c is called the *component of* $\mathbf{p}$ *along* $\mathbf{q}$. Note that

$$c = \frac{\mathbf{p} \cdot \mathbf{q}}{\mathbf{q} \cdot \mathbf{q}}.$$

Lecture 6

Matrices

We continue our review of linear algebra started in Lecture 5. We only do linear algebra over the real numbers $\mathbb{R}$.

We first review matrices and matrix multiplication, the key operation. The lecture then reviews standard material on square matrices: we build up to define and study the determinant. Permutations and permutation matrices are introduced in §6.4, as a tool for handling determinants and as an example of orthogonal matrices. We also study orthogonal matrices, which are important in the Spectral Theorem: §9.2. They will reappear when we discuss the convex set of doubly stochastic matrices in §18.8 and steady states for probability matrices in §26.3.

The main material covered in this lecture is standard[1].

The last section does not belong to the main thread of the lecture: Block decomposition of matrices: §6.10, which is a convenient computational tool used in Lectures 19 through 30. It can be postponed until it is needed.

Our notation for vectors and matrices is described in Appendix A.

6.1 Matrix Definitions

A matrix of size $m \times n$ is a collection of mn numbers with double indices i and j written in the following particular way:

$$a_{ij}, 1 \leq i \leq m, 1 \leq j \leq n.$$

These numbers are called the entries of the matrix.

We will write our matrices using capital roman letters, and their entries by the same lower case roman letter, with a double index. So for example, if A is a $m \times n$

[1] References to two introductory linear algebra texts ([60] and [68]), are included.

matrix, we write $A = [a_{ij}]$, where $1 \leq i \leq m$ and $1 \leq j \leq n$. We also write matrices out as rectangular arrays:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \tag{6.1.1}$$

which allows us to talk about the rows and the columns of a matrix. We write the i-th row of the matrix A as $\mathbf{a}^i$ and the j-th column as $\mathbf{a}_j$.

So the 2×3 matrix

$$A = \begin{bmatrix} 1 & 2 & 4 \\ -1 & 3 & 5 \end{bmatrix}$$

has two rows and three columns, and

$$\mathbf{a}^2 = \begin{bmatrix} -1 & 3 & 5 \end{bmatrix} \text{ and } \mathbf{a}_3 = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

A matrix of size $n \times 1$ is called a column vector of length n, or a n-column vector. A matrix of size $1 \times m$ is called a row vector of length m, or a m-row vector.

We can define two simple operations on $m \times n$ matrices A and B.

1. First addition: $A + B = C$ where $C = [c_{ij}]$ is the $m \times n$ matrix with $c_{ij} = a_{ij} + b_{ij}$ for all i, j. Thus the corresponding entries are added.

2. Then multiplication by a number c: $cA = [ca_{ij}]$, so each entry of the matrix A is multiplied by the scalar c.

6.1.2 Definition. Here are some special and important matrices to which we give names. First note that the diagonal of a square matrix A is the set of entries with equal indices: a_{ii}. The remaining elements are the off-diagonal terms.

- The $m \times n$ whose entries are all zero is written 0, or $0_{m \times n}$ if it is important to keep track of its size. The remaining definitions concern square matrices.

- The identity matrix I is the diagonal matrix with all diagonal terms equal to 1. If its size n needs to be recalled we write I_n. We usually write the entries of I as e_{ij}. So $e_{ii} = 1$ for all i, and $e_{ij} = 0$ if $i \neq j$.

- The square matrix A is diagonal if all its off diagonal terms are 0.

- A is *upper-triangular* if all the terms below the diagonal are zero. In other words $a_{ij} = 0$ when $i > j$. Correspondingly A is lower triangular if $a_{ij} = 0$ when $i < j$.

6.2 Matrix Multiplication

The fundamental matrix operation is multiplication of a $m \times n$ matrix A with a column vector $\mathbf{x}$ of length n to yield a column vector of length m.. Here is the all-important formula:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n \end{bmatrix}$$

Note the important special case where A is a row vector:

6.2.1 Definition.

$$\begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = a_1x_1 + a_2x_2 + \dots + a_nx_n.$$

Calling the row vector $\mathbf{a}$ and the column vector $\mathbf{x}$, we get $\mathbf{a} \cdot \mathbf{x}$, the inner product $\langle \mathbf{a}, \mathbf{x} \rangle$, or sot product of the two vectors $\mathbf{a}$ and $\mathbf{x}$.

6.2.2 Definition. The product $C = AB$ of a $m \times n$ matrix A multiplied on the right by a $n \times r$ matrix B is the $m \times r$ matrix $C = [c_{ik}]$, where

$$c_{ik} = a_{i1}b_{1k} + a_{i2}b_{2k} + \dots + a_{in}b_{nk}.$$

Using summation notation, we have

$$c_{ik} = \sum_{j=1}^{n} a_{ij}b_{jk}.$$

Note that as often in such cases we are summing over the repeated index j.

We can only form the product AB of a $m \times n$ matrix A by a $r \times s$ matrix B if $n = r$. In that case the product is a $m \times s$ matrix. This of course still works when B is a column vector, the special case where $s = 1$, in which $C = AB$ is a column vector of length m.

Matrix multiplication is associative: let A be a $m \times n$ matrix, B a $n \times r$ matrix, and C a $r \times s$ matrix. Then $A(BC) = (AB)C$, so the order in which the multiplications are performed does not matter.

Furthermore matrix multiplication distributes over matrix addition, whenever the two operations are possible. Let A and B be $m \times n$ matrices, and let C and D be $n \times r$ matrices, and let $c \in F$ be a number. Then

$$A(C+D) = AC + AD \text{ and } (A+B)C = AC + BC.$$

Also $(cA)(D) = c(AD)$ and $A(cD) = c(AD)$ for any real number c.

Next we review the transpose of a matrix A.

6.2.3 Definition. The transpose of the $m \times n$ matrix $A = [a_{ij}]$ is the $n \times m$ matrix A^T with entries $a^T_{ji} = a_{ij}$. So the transpose of a column vector $\mathbf{x}$ is a row vector.

6.2.4 Example. The transpose of the 2×3 matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \text{ is the } 3 \times 2 \text{ matrix } A^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

The transpose A^T takes $\mathbf{y} \in \mathbb{R}^m$ to $\mathbf{x} = A^T\mathbf{y}$. The transpose is used in the following important definition:

6.2.5 Definition. A square matrix A is *symmetric* if it is equal to its transpose:

$$A^T = A$$

.

We will study symmetric matrices in detail in Lecture 8.

6.2.6 Exercise. Show that for any square matrix A, the matrix $A + A^T$ is symmetric.

6.2.7 Exercise. For any two matrices A and B of the same size, show that

$$(A+B)^T = A^T + B^T$$

Here is a property of transposes that we will use often.

6.2.8 Proposition. *If C is an $m \times n$ matrix, and D an $n \times p$ matrix, then*

$$(CD)^T = D^T C^T$$

Proof. The proof is left as an exercise to the reader. □

6.2.9 Example. Let C and D be the 2×2 matrices

$$\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \text{ and } \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix}$$

Then compute CD, $(CD)^T$, and $D^T C^T$, checking that these last two are the same.

6.2.10 Exercise. Find a simple numerical example where the matrices C and D from the previous example do not commute, meaning that $CD \neq DC$.

6.3 Square Matrices

Not surprisingly, a square matrix is a matrix with the same number of rows as columns. Instead of saying a '$n \times n$ matrix', we will sometimes say a 'square matrix of size n'. The extra feature that arises when dealing with a square matrix A, is that we can form the product of that matrix with itself: AA, which we write A^2.

When A and B are both square matrices of the same size, we can form both products AB and BA.

6.3.1 Remark. Matrix multiplication is not commutative. If A and B are two square matrices of the same size n, so that AB and BA are both square matrices of size n, it is not necessarily the case that $AB = BA$. Give examples of square matrices of size 2 that do not commute.

6.3.2 Exercise. Let A and B be square matrices that commute: $AB = BA$. Show that

$$(A+B)^2 = A^2 + 2AB + B^2 \text{ and } (A+B)(A-B) = A^2 - B^2.$$

This shows that we can do algebra with squares matrices as with numbers, taking account, of course, that matrix multiplication is not generally commutative.

One key feature of multiplication of numbers is that there is a neutral element for multiplication, usually denoted 1. There also is a neutral element for matrix multiplication, the identity matrix I discussed earlier. As we noted, for any square matrix A of size n, $AI = IA = A$.

Continuing the analogy with multiplication of numbers, we may ask if a square matrix A has an inverse, meaning a square matrix B, called the inverse of A, of the same size as A so that

$$AB = I = BA. \tag{6.3.3}$$

It is easy to probe that If A has an inverse, then its inverse B is unique. The unique inverse is written A^{-1}.

A product of invertible matrices A and B is invertible, and $(AB)^{-1} = B^{-1}A^{-1}$, as follows from the associativity of matrix multiplication. Similarly, if a matrix A is invertible, its transpose A^T is too.

6.3.4 Exercise. Show that the inverse of the transpose is the transpose of the inverse: $(A^t)^{-1} = (A^{-1})^t$. Hint: take the transpose of the identity $AA^{-1} = I$ and use the uniqueness of the inverse.

6.4 Permutations

Our next goal is to review the determinant of a square matrix. Permutations are needed when defining the determinant, and permutation matrices form a beautiful subgroup of the orthogonal matrices, which is why we study them here. By subgroup, we mean that the product of any two permutation matrices is a permutation matrix, that permutation matrices are invertible, and that that the inverse of a permutation matrix is a permutation matrix. Finally the identity matrix I corresponds to the identity permutation. These assertions will be established here.

6.4.1 Definition. A permutation σ of a finite set S is a one-to-one map from S to itself.

If S has n elements, we will always use for S the integers $\{1, 2, \dots, n\}$.

6.4.2 Example. The trivial permutation (1) takes any i to itself.

The next simplest permutations are the *transpositions*, which interchange two integers but do not move the others. For example, the permutation σ with values $\sigma(1) = 2$, $\sigma(2) = 1$, and $\sigma(i) = i$ for $i \neq 1, 2$ is a transposition.

There are exactly two permutations on $\{1, 2\}$, the trivial permutation and the transposition exchanging 1 and 2.

A cumbersome way of writing a permutation σ on n elements consists in writing the integers 1 through n on a top row; then beneath each integer i write the value $\sigma(i)$. So, for example

$$\begin{vmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 3 & 1 \end{vmatrix}$$

denotes the permutation σ sending 1 to 2, 2 to 4, 3 to 3 and 4 to 1. In this notation, the fact that a permutation is one-to-one is expressed by the fact that each integer from 1 to n appears exactly once in the second row.

6.4.3 Exercise. Enumerate all the permutations on $\{1, 2, 3\}$, listing the trivial permutation, then all the transpositions, and then the remaining ones.

6.4.4 Exercise. Prove by induction that there are $n!$ different permutations on $\{1, 2, \dots, n\}$.

We can follow a permutation σ by another permutation τ, yielding a third permutation $\tau\sigma$ called the product permutation, or the composition of the two permutations. It sends the element k to $\tau(\sigma(k))$. Furthermore any permutation has an inverse, namely the permutation σ^{-1} that undoes the effect of σ: for all k, $\sigma^{-1}(\sigma(k)) = k$. These rules, plus the associativity of composition, makes the set of permutations on n elements a group. This is a structure defined in Abstract Algebra courses that only comes up peripherally here.

6.4.5 Exercise. Write the inverse of the permutation σ on 3 elements sending $1 \to 2$, $2 \to 3$, $3 \to 1$. Write $\sigma\sigma$.

6.4.6 Exercise. Find two permutations on three letters such that $\sigma\tau \neq \tau\sigma$. We say that permutations are not commutative.

We now exhibit two other ways of writing permutations. The first is by matrices. This will is useful to us later. The second, most compact, notation is the cycle notation. We will not have much use for it, but a quick exposition is given since it is the notation everyone uses.

A compact notation for determinants is the *cycle* notation, which we first exhibit by example. The permutation above is written (124) in cycle notation. The cycle notation for

$$\begin{vmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{vmatrix}$$

is $(12)(34)$.

Here is how to decipher cycle notation. First, note that the integers enclosed in a set of parentheses is called a *cycle*. The representation of a permutation in cycle notation consists in a collection of cycles where each integer appears *at most once*. To find where σ written in cycle notation sends the integer i, scan the representation. If i does not appear, that means that i is not moved by the permutation: $\sigma(i) = i$. Otherwise look at the element j immediately to the right of i, in the same cycle. If there is such an element, then $\sigma(i) = j$. If i is at the right of the cycle, then go all the way to the left inside the same cycle. This gives an algorithm for reconstructing the permutation from the cycle notation.

6.4.7 Example. The cycle notation for the permutation σ in 6.4.5 is (123). The notation for its composition with itself is (132). Note this is also σ^{-1}. We say σ has order 3, since the composition $\sigma\sigma\sigma$, which we write σ^3, is the trivial permutation.

6.4.8 Example. Consider the permutation σ on $\{1, 2, \dots, 6\}$ given in cycle notation by $(154)(26)$. Then $\sigma(3) = 3$, because 3 does not appear in the representation. $\sigma(1) = 5$, $\sigma(5) = 4$, and $\sigma(2) = 6$, (just move right by 1 spot), Where does σ send 4 and 6? They are on the right edge of their respective cycle, so go back to the left: $\sigma(4) = 1$ and $\sigma(6) = 2$. The long representation of this permutation is

$$\begin{vmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 5 & 6 & 3 & 1 & 4 & 2 \end{vmatrix}$$

Take for example two permutations on $\{1, 2, 3, 4\}$: first perform the permutation given by (1342) in cycle notation, and then the permutation $(14)(23)$. To find the composition, write the cycle presentation of the first permutation to be performed to the right of the second:

$$(14)(23)(1432) \tag{6.4.9}$$

This is not the cycle notation for a permutation, because there are numbers that appear more than once. To find the cycle decomposition of the composition, start, say with the element 1. Scan the product of cycle (6.4.9) from the right looking for 1. We see that it appears in (1432), and our algorithm says it goes to 4. Then move to the next cycle to the left, scanning for 4. We find it in (14), which says that 4 goes to 1. So under the composition, we see that 1 is permuted to 1, so it is fixed. We could write a cycle (1), or not write anything at all. Next repeat with 2. We see $2 \to 1 \to 4$. Next repeat with 4: $4 \to 3 \to 2$. So the cycle closes, and we have (24). At this point we know that 3 be fixed, but we check it: $3 \to 2 \to 3$. So the cycle representation of the product is $(1)(24)(3)$, or simply (24), a transposition.

6.4.10 Example. Find the cycle decomposition of these same two permutations but in the other order, namely simplify

$$(1432)(14)(23)$$

You get (13), so reversing the order changes the result.

6.5 Permutation Matrices

Represent the n objects to be permuted by the n unit vectors $\mathbf{e}_j$ in $\mathbb{R}^n$. Given a permutation σ on $\{1, 2, \dots, n\}$, we can then ask for a $n \times n$ matrix P^σ such that matrix multiplication $P^\sigma \mathbf{e}_j$ yields $\mathbf{e}_{\sigma(j)}$. Because $P^\sigma I = P^\sigma$, where I is the $n \times n$ identity matrix, the j-th column of P^σ must be the unit vector $\mathbf{e}_{\sigma(j)}$. In particular P^σ must have exactly one 1 in each row and column with all other entries being 0. We can make the formal definition:

6.5.1 Definition. Given a permutation σ on $\{1, 2, \dots, n\}$, we define the *permutation matrix* $P^\sigma = \begin{bmatrix} p^\sigma_{ij} \end{bmatrix}$ of σ as follows: $p^\sigma_{ij} = 0$ unless $i = \sigma(j)$. Another way of saying this is that the j-th column of P^σ is $\mathbf{e}_{\sigma(j)}$.

6.5.2 Example. In Example 6.4.5 we considered the permutation

$$\begin{vmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{vmatrix}$$

Its permutation matrix is

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

since

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \text{, etc.}$$

Conversely, to each matrix with exactly one 1 in each row and column, and 0 everywhere else, there corresponds a permutation. Indeed if the matrix has a 1 in position (i, j), then the associated permutation σ satisfies $\sigma(j) = i$.

Furthermore

6.5.3 Proposition. *The product of two permutation matrices is a permutation matrix.*

The elementary proof is left to you: just use the definition. We can compare the composition $\tau\sigma$ of two permutations σ and τ with the permutation associated to the product of the permutation matrices $P^{\tau}P^{\sigma}$. Recall that we act with the permutation σ first. Not surprisingly, matrix multiplication of permutation matrices reflects composition of permutations, just as it reflects composition of linear maps, yielding the easy theorem:

6.5.4 Theorem. *If σ and τ are two permutations on $\{1, 2, \dots, n\}$, and P^{σ} and P^{τ} the associated $n \times n$ permutation matrices, then:*

$$P^{\sigma}P^{\tau} = P^{\sigma\tau}.$$

Every permutation σ has an inverse permutation σ^{-1}: the permutation that undoes σ. Its permutation matrix is the transpose of the matrix of σ. Check this for Example 6.5.2, and work out a general proof.

6.5.5 Example. The permutation matrix of the transposition τ exchanging 1 and 2 has matrix

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

6.5.6 Definition. The *sign* of a permutation σ is either 1 or -1. We write it $sgn(\sigma)$. It is 1 if the permutation can be written as the product of an even number of transpositions; -1 if it can be written as a product of an odd number of transpositions.

Since the representation of a permutation as a product of transpositions is not unique, it is not obvious that this definition makes sense.

For any permutations σ and τ, $sgn(\sigma\tau) = sgn(\sigma)sgn(\tau)$.

6.5.7 Example. The permutation matrix associated to the permutation on $\{1, 2, 3\}$ written in cycle notation $(1, 2, 3)$ is given in Example 6.5.2, and has sign 1, while the sign of the permutation in Example 6.5.5 is -1.

6.6 Determinants

Throughout this section, A denotes the $n \times n$ matrix $[a_{ij}]$. The goal of this section is to define the determinant of A and give its properties. The reader is assumed to be familiar with the rudiments of this material, covered for example in [60], chapter 3, [69], chapter 5, or [13], chapter 5.

Using permutations and their sign, we can give the definition of the determinant:

6.6.1 Definition. If A is an $n \times n$ matrix, then

$$\det A = \sum_{\sigma} sgn(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \dots a_{n\sigma(n)}$$

where the sum is over the $n!$ permutations on $\{1, 2, \dots, n\}$.

Thus by definition the permutation matrix P^{σ} has determinant $sgn(\sigma)$. It is not hard to see that $\det A^T = \det A$.

Indeed, each term in Definition 6.6.1 is the determinant of a matrix we call A^{σ} which has $a_{i\sigma(i)}$ in entry $(i, \sigma(i)$ for all i, $1 \leq i \leq n$, and zeroes everywhere else. So

$$A = \sum_{\sigma} A^{\sigma}, \text{ and } \det A = \sum_{\sigma} \det A^{\sigma},$$

where the sums are over all permutations. In the same way $A^T = \sum_{\sigma} (A^{\sigma})^T$. Furthermore $\det((A^{\sigma})^T) = \det A_{\sigma}$, because $sgn(\sigma) = sgn(\sigma^{-1})$, and everything else is the same.

6.6.2 Example. For the matrix A in Example 6.6.4 and the permutation σ of Example 6.5.2 we have

$$A^{\sigma} = \begin{bmatrix} 0 & 0 & 3 \\ 4 & 0 & 0 \\ 0 & 8 & 0 \end{bmatrix}$$

so that the 1s of P^{σ} have been replaced by the appropriate entries of A.

We will need the following important property.

Here is a second definition of the determinant that is sometimes used. It proceeds by induction on n. So we first need to define the determinant of the 1×1 matrix $[a]$: It is, of course, a. Next we define the determinant of an $n \times n$ matrix in terms of certain $(n-1) \times (n-1)$ submatrices.

6.6.3 Definition. The $(n-1) \times (n-1)$ matrix A_{ij} obtained by deleting the i-th row and the j-th column of A is called the ij-th *submatrix* of A, and its determinant m_{ij} is called the ij-th *minor* of A. [2]

6.6.4 Example. If A is the matrix

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

then

$$A_{11} = \begin{bmatrix} 5 & 6 \\ 8 & 9 \end{bmatrix} \quad \text{and } m_{11} = -3;$$

$$A_{12} = \begin{bmatrix} 4 & 6 \\ 7 & 9 \end{bmatrix} \quad \text{and } m_{12} = -6;$$

and

$$A_{13} = \begin{bmatrix} 4 & 5 \\ 7 & 8 \end{bmatrix} \quad \text{and } m_{13} = -3.$$

6.6.5 Definition. For a $n \times n$ matrix A, $n \geq 2$,

$$\det A = a_{11} \det A_{11} - a_{12} \det A_{12} + \dots (-1)^{1+n} a_{1n} \det A_{1n}$$

or, in summation notation,

$$\det A = \sum_{j=1}^{n} (-1)^{1+j} a_{1j} \det A_{1j} = \sum_{j=1}^{n} (-1)^{1+j} a_{1j} m_{1j}.$$

This is called expanding the determinant along the first row. By induction this gives us a definition of the determinant for any n. We could have done something similar along any row and any column, although it is not then clear that the definitions agree. Any such expansion is called a Laplace expansion of the determinant.

You should check that this agrees with Definition 6.6.1 when $n = 2$: just two terms. Then when $n = 3$: 6 terms. In the second case, you need to compute the

[2] Some books use the term minor to describe what we call the submatrix, and some dispense with the term minor completely, preferring to use the *cofactor* $c_{ij} = (-1)^{i+j} m_{ij}$

sign of the 6 elements of the permutation group. Three have sign= 1: they are the identity element, and the two elements written in cyclic notation as (123) and (132). The three permutations with sign -1 are the three transpositions (12), (13) and (23).

The general case is more difficult: one needs to establish that the determinant is a linear function of its columns. For more details see [38], Chapter VI or [39], Chapter 5, for example.

6.6.6 Example. The determinant of the matrix A from Example 6.6.4, following our formula, is given by:

$$1(5\cdot 9-6\cdot 8)-2(4\cdot 9-6\cdot 7)+3(4\cdot 8-5\cdot 7)=-3+12-9=0.$$

6.6.7 Exercise. The analogous formula for the determinant obtained by expanding along the i-th row is

$$\sum_{j=1}^{n}(-1)^{i+j}a_{ij}\det A_{ij}=\sum_{j=1}^{n}(-1)^{i+j}a_{ij}m_{ij},$$

and for expanding along the j-column is

$$\sum_{i=1}^{n}(-1)^{i+j}a_{ij}\det A_{ij}=\sum_{i=1}^{n}(-1)^{i+j}a_{ij}m_{ij}.$$

Notice that only the index of summation has changed. In Example 6.6.4, check that you get the same answer for at least one new row expansion and one column expansion.

We can encode all the row and column expansions into two matrix multiplications. Let M be the square matrix of size n whose (i,j)-th entry is = $(-1)^{i+j}m_{ij}$, where m_{ij} is the ij-th minor of A as before. Its transpose M^T is the *adjoint* of A or the *matrix of cofactors* of A.

6.6.8 Theorem. *If A is any square matrix, and M^T is the adjoint of A, then*

$$M^TA=AM^T=(\det A)I$$

We will not prove this. Note that the fact that all the diagonal terms of the matrix on the right are equal to the determinant says that the expansions by minors along all rows and columns give the determinant. The fact that the off-diagonal terms are zero is equivalent to the fact that all matrices with a repeated row or column have determinant 0. For more details see [68], §4.4, [60], §3.1, or [13],

§5.2. Note that the formula shows that when $\det A \neq 0$, A is invertible: see Theorem 6.6.10 below for a stronger result. It also computes the inverse matrix, albeit very inefficiently, in that case:

$$A^{-1} = \frac{1}{\det A} M^T.$$

A little thought should also convince you that this theorem encodes all the information in Cramer's rule, as the references given above show.

6.6.9 Exercise. Compute the determinant of

$$M = \begin{bmatrix} 1 & -2 & 0 & 0 \\ -3 & 2 & 0 & 0 \\ 0 & 0 & -1 & 3 \\ 0 & 7 & 2 & 1 \end{bmatrix} \text{ and } N = \begin{bmatrix} 1 & -2 & 0 & 0 \\ -3 & 2 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 1 & 1 & 2 & 1 \end{bmatrix}.$$

The following theorems and corollary are the additional key facts we need concerning the determinant. We will not prove them.

6.6.10 Theorem.

1. *If A and B are both $n \times n$ matrices, then* $\det(AB) = \det A \det B$. *The determinant of the identity matrix is 1.*

2. *A square matrix A is invertible if and only if its determinant is non-zero, in which case the determinant of the inverse A^{-1} is*

$$\det A^{-1} = \frac{1}{\det A}$$

6.6.11 Corollary. *Let $A(ij)$ be the matrix obtained by interchanging the i-th and the j-th columns (or rows) of the matrix A. Then* $\det A(ij) = -\det A$.

Proof. $A(ij)$ can be obtained from A by multiplication by a transposition, and the determinant of a transposition is -1. □

6.7 Orthogonal Matrices

6.7.1 Definition. A $n \times n$ matrix Q is *orthogonal* if its columns $\mathbf{q}_i$ have length 1 and are mutually perpendicular, so $\mathbf{q}_i \cdot \mathbf{q}_i = 1$ and $\mathbf{q}_i \cdot \mathbf{q}_j = 0, i \neq j$.

6.7.2 Proposition. *Orthogonal matrices are invertible, and their inverse is equal to their transpose. In other words, $Q^{-1} = Q^T$ for any orthogonal Q. Conversely any invertible matrix Q such that $Q^{-1} = Q^T$ is orthogonal.*

Proof. For any orthogonal matrix Q, $Q^TQ = I$, the identity matrix. Indeed, the ij-th entry of Q^TQ is the dot product of the i-th row of Q^T with the j-th column of Q, so by definition it is equal to 1 if $i = j$ and equal to 0 if $i \neq j$. This equation says that Q is invertible and that the inverse is Q^T. For the converse just reverse this argument. □

6.7.3 Exercise. Prove that the determinant of an orthogonal matrix is ± 1.
Hint: Use induction on n and the Laplace expansion of the determinant.

6.7.4 Exercise. Prove that the product of two orthogonal matrices is orthogonal.
Hint: Take two orthogonal matrices Q and R, so that $Q^T = Q^{-1}$ and $R^T = R^{-1}$. Compute $(QR)^T$ using Proposition 6.2.8. Can you compute $(QR)^{-1}$? Recall that this is the unique matrix S such that $QRS = I$.

6.7.5 Proposition. *Orthogonal matrices Q preserve distance and angle, by which we mean the following: for any two vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *in* V*,*

$$\langle Q\mathbf{a}, Q\mathbf{b}\rangle = \langle \mathbf{a}, \mathbf{b}\rangle \textit{ and therefore } \|Q\mathbf{a}\| = \|\mathbf{a}\|$$

This explains the name “orthogonal”: angle and length are preserved.

Proof. It is enough to prove the first equality. Writing the left-hand side in terms of matrix multiplication, we get

$$\langle Q\mathbf{a}, Q\mathbf{b}\rangle = \mathbf{a}^T Q^T Q\mathbf{b} = \mathbf{a}^T\mathbf{b} = \langle \mathbf{a}, \mathbf{b}\rangle$$

where we used orthogonality of Q to get $Q^TQ = Q^{-1}Q = I$, □

6.7.6 Example (Permutation Matrices)**.** We discussed permutations and permutation matrices in §6.4 . As we noticed there, the transpose of a permutation matrix is its inverse, so permutation matrices are orthogonal matrices.

6.7.7 Example (Rotation Matrices)**.** The matrix

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

is orthogonal, as an easy computation shows. It rotates the plane by angle θ. For our purposes, R is a good source of counterexamples. Indeed, it cannot have real eigenvalues since all directions are moved, so the Spectral Theorem 9.2.1 fails for this matrix, at least when $\sin\theta \neq 0$. This is because R is not symmetric.

6.7.8 Example. Consider any 3×3 orthogonal matrix Q with determinant 1. We show it is a rotation. The characteristic polynomial $p(t)$ (7.3.2) of Q has degree 3. Because the highest degree term of $p(t)$ is t^3, it goes to ∞ as t gets large. Because $\det Q = 1$, $p(0) = -1$. Thus the graph of $p(t)$ crosses the t-axis for a positive value of t. This means Q has a real eigenvector $\mathbf{e}^1$ with associated positive eigenvalue λ, so $Q\mathbf{e}^1 = \lambda\mathbf{e}^1$. Because Q is orthogonal and preserves distance, $\lambda = 1$. Let H be the orthogonal complement of the line L spanned by $\mathbf{e}^1$. Next we show that for all $\mathbf{v} \in H$, then $Q\mathbf{v}$ is perpendicular to $\mathbf{e}^1$, so $Q\mathbf{v} \in H$. Indeed,

$$\begin{aligned} \langle Q\mathbf{v}, \mathbf{e}^1\rangle &= \langle Q\mathbf{v}, Q\mathbf{e}^1\rangle && \text{using the eigenvector equation for } \mathbf{e}^1, \\ &= \langle \mathbf{v}, \mathbf{e}^1\rangle && \text{by Proposition 6.7.5,} \\ &= 0 && \text{since } \mathbf{v} \perp \mathbf{e}^1. \end{aligned}$$

It is then an easy exercise to show that the restriction of A to H is an ordinary rotation in the plane. The line L is called the *axis of rotation*. It is uniquely determined unless the rotation is trivial.

6.8 Principal Submatrices and Principal Minors

The definitions[3] in this short section will be used when we discuss Gaussian elimination in §6.9: see Theorem 6.9.6): and positive definite and positive semidefinite forms: see Theorems 9.4.1 and 9.5.1.

6.8.1 Definition. A *principal submatrix* of size k of an $n \times n$ matrix $A = [a_{ij}]$ is a matrix obtained by removing $n - k$ columns and the corresponding $n - k$ rows from A. Thus if you remove the first column, you must remove the first row, etc.

6.8.2 Remark. An $n \times n$ matrix A has $\binom{n}{k}$ principal submatrices of size k. Indeed, to form the principal submatrices of size k, we are picking k objects from n objects, and the number of ways of doing this is precisely the binomial coefficient $\binom{n}{k}$.

Thus a 3×3 matrix A has 3 principal submatrices of size 1: the three 1×1 matrices $[a_{ii}]$. A has 3 principal submatrices of size 2: the three 2×2 matrices

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \begin{bmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{bmatrix}, \text{ and } \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix}$$

A has only one principal submatrix of size 3—itself.

[3] Principal minors and leading principal minors are also defined in [59], 16.2 p. 381.

6.8.3 Definition. The k-th *leading principal submatrix* of an $n \times n$ matrix $A = (a_{ij})$ is the $k \times k$ matrix A_k, $1 \le k \le n$ with entries a_{ij}, $1 \le i \le k$, $1 \le j \le k$. In other words, A_k is the uppermost and leftmost submatrix of A of size $k \times k$. For this reason, it is sometimes simply called the upper left submatrix (see [67], pg. 331). $D_k = \det(A_k)$ is the kth leading principal minor.

So the 3×3 matrix A above has, as first leading principal submatrix $[a_{11}]$, and as second leading principal submatrix

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

so that the second leading principal minor is $D_2 = a_{11}a_{22} - a_{12}a_{21}$.

6.9 Gaussian Elimination via Matrices

One of the first things one studies in any linear algebra course is Gaussian elimination, the standard method for solving a system of linear equations. When studying the LU decomposition, it is explained that the Gaussian elimination is equivalent to left-multiplying the matrix representing the system by a collection of *elementary matrices*. References are, for example [60], §2.5 or [68], §2.6. Here we will show how this works for a square system, and later specialize to a system where the matrix coefficients are symmetric, the case we will use later. For a more general situation, see the references above, or even more generally, see [38], chapter II, §5. We follow the presentation and largely the notation of Gantmacher [25], chapter II, §1.

So assume we have a system of n equations in n variables, of rank r:

$$A\mathbf{x} = \mathbf{b}.$$

By a suitable reordering of the variables and the equations we can arrange that the leading principal minors D_k of A, $1 \le k \le r$, are all non-zero, while $D_k = 0$ when $k > r$. The last inequalities are implied by the rank of A being r. In particular $D_1 = a_{11} \neq 0$. Then:

6.9.1 Definition. a_{11} is the first *pivot* d_1 of A.

Let E_1 be the elementary $n \times n$ matrix:

$$E_1 = \begin{bmatrix} 1 & 0 & \dots & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}}{a_{11}} & 0 & \dots & 1 \end{bmatrix}$$

So E_1 is invertible with determinant equal to one, since it is lower triangular with ones on the diagonal. Let us write $A^{(1)}$ for the product matrix E_1A. By construction:

$$A^{(1)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & \dots & a_{3n}^{(1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(1)} & a_{n3}^{(1)} & \dots & a_{nn}^{(1)} \end{bmatrix}$$

where the matrix entries without suffix are the entries of A, while those with an upper (1) are just by definition the entries of $A^{(1)}$.

6.9.2 Proposition. *We compute the second diagonal element of the matrix $A^{(1)}$, and show it is non-zero, under the assumption that the rank of A is at least 2, so it will serve as our second pivot d_2:*

$$a_{22}^{(1)} = a_{22} - \frac{a_{12}a_{21}}{a_{11}} = \frac{D_2}{D_1}. \tag{6.9.3}$$

Proof. Because $A^{(1)}$ was obtained from A by adding the first row of A multiplied by a constant, the minors that contain that row (in particular the leading principal minors) do not change when one passes from A to $A^{(1)}$, by the well-known property of the determinant. On the other hand, the second leading principal minor of $A^{(1)}$ is simply $a_{11}a_{22}^{(1)}$, because that principal matrix is triangular. So $a_{11}a_{22}^{(1)} = D_2$, and since $D_1 = a_{11}$, this is what we found by direct computation. This computation establishes the result, since by hypothesis, the leading principal minor D_2 is non-zero. □

This simple but important argument will generalize as we create more zeroes by Gaussian elimination.

6.9.4 Exercise. Write down the definition of E_2 using that of E_1 as a model.

We write $A^{(2)}$ for the matrix $E_2A^{(1)}$. By construction:

$$A^{(2)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \dots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & a_{n3}^{(2)} & \dots & a_{nn}^{(2)} \end{bmatrix}$$

We claim, as before, that if $3 \le r$, where r is the rank of A, then $a_{33}^{(2)} \neq 0$, because

$$a_{33}^{(2)} = \frac{D_3}{D_2}.$$

by the same argument as in (6.9.3). So this gives us the third pivot d_3.

So if $2 < r$ we can continue the elimination process until we reach the rank r of the matrix.

For simplicity, let's first consider the case of maximum rank $r = n$. At each step we get a new non-zero pivot

$$d_k = a_{kk}^{(k-1)} = \frac{D_k}{D_{k-1}}.$$

so in the end we get the upper triangular matrix:

$$A^{(n-1)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1,n-1} & a_{1n} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2,n-1}^{(1)} & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \dots & a_{3,n-1}^{(2)} & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{n-1,n-1}^{(n-2)} & a_{n-1,n}^{(n-2)} \\ 0 & 0 & 0 & \dots & 0 & a_{nn}^{(n-1)} \end{bmatrix}$$

with an accompanying lower triangular matrix $E = E_{n-1}E_{n-2}\cdots E_2E_1$. By construction $A^{(n-1)} = EA$.

Now assume that A is symmetric. Since E is lower triangular, E^T is upper triangular. So $(EA)E^T$, the product of two upper triangular matrices, is upper triangular. But EAE^T is symmetric: just compute its transpose as in Proposition 6.2.8. The only symmetric upper triangular matrices are diagonal, so EAE^T is diagonal and we have achieved the goal of Gaussian elimination without any further computation. We record this special case as a theorem.

6.9.5 Theorem. *Assume A is a symmetric matrix of size n such that all its leading principal minors are non zero. Then Gaussian elimination can be accomplished by left multiplication by an invertible lower triangular matrix E of determinant* 1. *The k-th diagonal element of the diagonal matrix EAE^T is $d_k = \frac{D_k}{D_{k-1}}$, where the D_k, $1 \le k \le n$ are the leading principal minors of A, and $D_0 = 1$ by convention.*

We now generalize the construction to matrices of smaller rank. It will give us one of the characterizations of positive (semi)definite matrices: see Theorem 9.4.1.and 9.5.1. It can also be used to compute the signature of a quadratic form in many cases, as explained in [25], volume 1, p.302.

6.9.6 Theorem. *A is an $n \times n$ symmetric matrix of rank r with non-zero leading principal minors D_k, $1 \le k \le r$. Then Gaussian elimination can be performed to produce zeroes below the first r diagonal elements of the matrix. Denoting the pivots of A by d_k, $1 \le k \le n$, we have*

$$d_k = \frac{D_k}{D_{k-1}} \text{ for } 1 \le k \le n$$

where $D_0 = 1$ by definition.

Proof. After the first $k-1$ columns of A have been cleared by forward elimination, the k-th leading submatrix A_k is upper triangular with the first k pivots on the diagonal. So $D_k = \det(A_k) = \prod_{i=1}^{k} d_i$. Further Gaussian elimination does not modify A_k. Thus, if all leading principal minors of A are non-zero, then so are all the pivots, which means that Gaussian elimination can occur without row exchanges.

□

Then by left multiplication by invertible matrices we get, after $r-1$ steps:

$$A^{(r-1)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1r} & \dots & a_{1n} \\ 0 & a^{(1)}_{22} & a^{(1)}_{23} & \dots & a^{(1)}_{2r} & \dots & a^{(1)}_{2n} \\ 0 & 0 & a^{(2)}_{33} & \dots & a^{(2)}_{3r} & \dots & a^{(2)}_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a^{(r-1)}_{r,r} & \dots & a^{(r-1)}_{r,n} \\ 0 & 0 & 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 & \dots & 0 \end{bmatrix}$$

together with an invertible matrix E so that $EA = A^{(r-1)}$. The non-zero pivots are $\frac{D_k}{D_{k-1}}$, $1 \le k \le k$.

Once you have studied congruent matrices in §8.4, you will see:

6.9.7 Theorem. *The method of Gaussian elimination applied to a symmetric matrix of rank r with non-zero leading principal minors D_k, $1 \le k \le r$, yields a diagonal matrix in the same congruence class as A.*

Proof. This is just the content of Theorem 6.9.6 interpreted in the language of congruence classes of matrices. □

6.10 Block Decomposition of Matrices

A facility with computations using block matrix notation is crucial for matrix computation, which is why we study it here. The simple point is that it is often convenient to think of a matrix as being made up of a grid of smaller submatrices. Here is the general procedure. There is nothing difficult here except the notation. This section can be skipped until it is needed later.

6.10.1 Definition. Let A be a $m \times n$ matrix. Write m as the sum of positive numbers $m_1, \dots, m_s$ and n as the sum of positive integers $n_1, \dots, n_t$.

Then we can write

$$A = \begin{bmatrix} A^{11} & A^{12} & \dots & A^{1t} \\ A^{21} & A^{22} & \dots & A^{2t} \\ \vdots & \vdots & \ddots & \vdots \\ A^{s1} & A^{s2} & \dots & A^{st} \end{bmatrix}$$

where A^{ij} is the $m_i \times n_j$ submatrix of A in the appropriate position. So there are st submatrices.

This is known as partitioning, or decomposing, the matrix into blocks.

By definition all blocks in a given column share the same columns elements of A, while all blocks in a given row share the same row elements.

6.10.2 Example. The 3×4 matrix

$$M = \begin{bmatrix} a_{11} & a_{12} & b_{13} & b_{14} \\ a_{21} & a_{22} & b_{23} & b_{24} \\ c_{31} & c_{32} & d_{33} & d_{34} \end{bmatrix}$$

can be partitioned into the blocks

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where A and B are 2×2 matrices, and C and D are 1×2 matrices. So in this example $s = t = 2$, and $m_1 = n_1 = n_2 = 2$ while $m_2 = 1$. In these lectures we will never exceed the case $s = 3$ and $t = 3$, by the way.

The point of decomposing matrices into blocks is that matrix multiplication behaves nicely with respect to block decomposition, as we will now see. So if some of the blocks are repeated or are simple (for example the identity matrix or the zero matrix) block multiplication can speed up the computation of the matrix product.

6.10.3 Example. Let A be an $m \times n$ matrix and let B be an $n \times p$ matrix. Let C be the product matrix AB of size $m \times p$. We block decompose A with

$$\begin{aligned} m &= m_1 + m_2; \\ n &= n, \end{aligned}$$

so there is no decomposition into columns. We block decompose B with

$$\begin{aligned} n &= n, \\ p &= p_1 + p_2, \end{aligned}$$

so there is no decomposition into rows. So

$$A = \begin{pmatrix} A^{11} \\ A^{21} \end{pmatrix}, B = \begin{pmatrix} B^{11} & B^{12} \end{pmatrix} \tag{6.10.4}$$

Then C can be partitioned according to the partition of the rows of A and the columns of B so that

$$C = \begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix} \tag{6.10.5}$$

with $C^{ij} = A^{i1} B^{1j}$.

We have decomposed the two matrices to be multiplied into blocks in such a way that the blocks are of size suitable for multiplication. All we needed in this example is that the decomposition of the columns of the left hand matrix A is the same as the decomposition of the rows of the right hand matrix B.

6.10.6 Example. If A and B are decomposed in the other direction, with the common index n written as $n_1 + n_2$ for both matrices, and no decomposition of the other indices m and p, then we case write the matrix product as

$$\begin{bmatrix} A^{11} & A^{12} \end{bmatrix} \begin{bmatrix} B^{11} \\ B^{21} \end{bmatrix} = A^{11}B^{11} + A^{12}B^{21}$$

You should check that the matrix multiplications and the matrix addition on the right hand side are well defined.

6.10.7 Exercise. Let

$$A = \begin{bmatrix} 1 & -2 \\ -3 & 2 \\ -1 & 3 \end{bmatrix} \text{ and } B = \begin{bmatrix} 1 & -2 & 1 \\ -3 & 2 & 0 \end{bmatrix}$$

Break A into two blocks

$$A^{11} = \begin{bmatrix} 1 & -2 \\ -3 & 2 \end{bmatrix}, A^{21} = \begin{bmatrix} -1 & 3 \end{bmatrix}$$

Now break B into two blocks so that the decomposition of the column size $(3 = 2 + 1)$ of A agrees with that of the row size $(3 = 2 + 1)$ of B.

$$B^{11} = \begin{bmatrix} 1 & -2 \\ -3 & 2 \end{bmatrix}, B^{12} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

This allows block multiplication. Check that the formula of Example 6.10.3 applies by computing the matrix product two ways.

Here is the main theorem of this section.

6.10.8 Theorem. *Let A be a $m \times n$ matrix block decomposed according to Definition 6.10.1. Let B be a $n \times p$ matrix block decomposed along its rows exactly as A is along its columns, and where $p = p_1 + \cdots + p_u$ is the block decomposition of its columns, so*

$$B = \begin{bmatrix} B^{11} & B^{12} & \dots & B^{1u} \\ B^{21} & B^{22} & \dots & B^{2u} \\ \vdots & \vdots & \ddots & \vdots \\ B^{t1} & B^{s2} & \dots & B^{tu} \end{bmatrix}.$$

Thus B^{jk} is a $n_j \times p_k$ submatrix of B. Then $AB = C$, where the $m \times p$ matrix C can be blocked decomposed as

$$C = \begin{bmatrix} C^{11} & C^{12} & \dots C^{1u} & \\ C^{21} & C^{22} & \dots C^{2t} & \\ \vdots & \vdots & \ddots & \vdots \\ C^{s1} & C^{s2} & \dots C^{su} & \end{bmatrix}$$

where C^{ik} is a $m_i \times p_j$ matrix such that

$$C^{ik} = A^{i1}B^{1k} + A^{i2}B^{2k} + \cdots + A^{it}B^{tk} = \sum_{j=1}^{t} A^{ij}B^{jk} \qquad (6.10.9)$$

Note that (6.10.9) is Definition 6.2.2 with blocks instead of numbers.

Proof. The details of the proof are left to the reader. First notice that the matrices on the right hand side of (6.10.9) are of the appropriate size to be multiplied and added.. Finally just check that for each entry of the matrix C^{ik} you have all the terms of the appropriate entry of C. □

An important special case occurs when the matrices A and B are square, meaning that $m = n = p$, and when the diagonal blocks are also square, implying that $s = t$, and $m_i = n_i$, $1 \leq i \leq n$. In this case, A^{ii} is an $n_i \times n_i$ matrix.

6.10.10 Definition. Assume that the matrix A is square of size n and that its diagonal blocks A^{ii} are square of sizes $n_1, n_2, \dots, n_s$ with $n = n_1 + n_2 + \dots + n_s$.

- Then A is *block diagonal* if A^{ij}, $i \neq j$, is the zero matrix:

$$A = \begin{bmatrix} A^{11} & 0 & \dots & 0 \\ 0 & A^{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A^{ss} \end{bmatrix} \tag{6.10.11}$$

- A is *block upper triangular* if A^{ij}, $i > j$, is the zero matrix:

$$A = \begin{bmatrix} A^{11} & A^{12} & \dots & A^{1s} \\ 0 & A^{22} & \dots & A^{2s} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A^{ss} \end{bmatrix} \tag{6.10.12}$$

In the same way we can define block lower triangular.

6.10.13 Proposition. *Assume A and B are square matrices of size n, and and that blocks are of size $n_1, n_2, \dots, n_s$ with $n = n_1 + n_2 + \dots + n_s$.*

- *If they are both block diagonal, their product $C = AB$ is also block diagonal, with $C^{ii} = A^{ii}B^{ii}$. Furthermore the k-th power of A can be written*

$$A^k = \begin{bmatrix} (A^{11})^k & 0 & \dots & 0 \\ 0 & (A^{22})^k & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (A^{ss})^k \end{bmatrix} \tag{6.10.14}$$

- *If A and B are both block upper triangular, then so it their product.*

Proof. We prove Proposition 6.10.13 using the main theorem. The diagonal case is trivial, so let's just consider the upper triangular case. If $C = AB$ we must show that $C_{ik} = 0$ when $i > k$. By hypothesis $A^{it} = 0$ when $i > t$ and $B^{tk} = 0$ when $t > k$. By (6.10.9) this means that the only non-zero terms in the sum are those with $i \leq t \leq k$. Since $i > k$, there are no such terms. □

6.10.15 Example. A special case that will be of importance to us in the one where A and B are both square of size $n = r + s$ and decomposed as

$$\begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{bmatrix} \text{ and } \begin{bmatrix} B^{11} & B^{12} \\ B^{21} & B^{22} \end{bmatrix}.$$

where A^{11} and B^{11} are $r \times r$ matrices,
A^{12} and B^{12} are $r \times s$ matrices,
A^{21} and B^{21} are $s \times r$ matrices,
A^{22} and B^{22} are $s \times s$ matrices. Then

$$AB = \begin{bmatrix} A^{11}B^{11} + A^{12}B^{21} & A^{11}B^{12} + A^{12}B^{22} \\ A^{21}B^{11} + A^{22}B^{21} & A^{21}B^{12} + A^{22}B^{22} \end{bmatrix}$$

It is easier to check in this special case that the formula is correct.

6.10.16 Example. In this example, if A and B are both block upper triangular, meaning that A^{21} and B^{21} are both the zero matrix, then their product AB is also block upper triangular.

6.10.17 Exercise. Let A be a 4×2 matrix and B be a 2×4 matrix, written in block form as in (6.10.4), where all the blocks are 2×2. Further assume that

$$A^{11} = A^{21} = \begin{bmatrix} 1 & 1 \\ -1 & -2 \end{bmatrix} \text{, and } B^{11} = B^{12} = \begin{bmatrix} 2 & 1 \\ -1 & -1 \end{bmatrix};$$

Write out the matrices A and B, compute the product AB directly, and then compute it by block multiplication.

6.10.18 Exercise. If you have the block decomposition of a matrix A, write a decomposition for its transpose A^T.

Lecture 7

Linear Transformations

We start by considering a $m \times n$ matrix A, which by matrix multiplication gives us a linear map from $\mathbb{R}^n \to \mathbb{R}^m$, sending a point $\mathbf{x} \in \mathbb{R}^n$ to $A\mathbf{x} \in \mathbb{R}^m$. We also consider its transpose A^T, which gives a map in the other direction: $\mathbb{R}^m \to \mathbb{R}^n$. We study the four subspaces associated to this pair.

Linear transformations may be familiar from a linear algebra course, but are worth reviewing. While this course is mainly concerned with matrices, not linear transformations, it is important to see how linear transformations motivate some of the operations we perform on matrices. We start by analyzing the relationship between composition of linear maps and matrix multiplication in Proposition 7.1.6.

The central result of the lecture is Theorem 7.8.10, which says that if you take a linear transformation T from a vector space V to itself, and if you look at the matrix A of T in one basis for V, then at the matrix B of T in another basis, these matrices are *similar*

That explains the importance of similarity studied in §7.7, an equivalence relation on matrices. The important Theorem 7.7.4 shows that similar matrices have the same characteristic polynomial: thus all their other invariants are the same. We analyze the relationship between linear transformations and matrices, showing how the matrix of a linear transformation from a vector space to itself changes when the basis of the vector space changes. As shown in Theorem 7.8.10, the matrix is transformed to a similar matrix, which explains the importance of similarity.

Our notation for vectors and matrices is described in Appendix A.

7.1 Linear Maps

We build on the definition of a vector space reviewed in §5.1.

7.1.1 Definition. A *linear transformation* T from a vector space V to a vector space W is a function such that

$$T(a_1\mathbf{v}_1 + a_2\mathbf{v}_2) = a_1T(\mathbf{v}_1) + a_2T(\mathbf{v}_2)$$

for all real numbers a_1 and a_2 and all vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ in V. On the right-hand side, $T(\mathbf{v}_1)$ and $T(\mathbf{v}_2)$ are elements of W, so the addition sign there denotes vector addition in W.

Here is the key consequence of this definition.

7.1.2 Proposition. *To understand the effect of a linear transformation T on any vector in V, all we need to know is the effect of T on a basis of V.*

Proof. Let n be the dimension of V and m the dimension of W. Pick a basis $\mathbf{e}_1$, …, $\mathbf{e}_n$ of V and a basis $\mathbf{f}_1, \dots, \mathbf{f}_m$ of W. Writing

$$T(\mathbf{e}_j) = \sum_{i=1}^{m} a_{ij}\mathbf{f}_i, \tag{7.1.3}$$

the linear transformation T is determined by the a_{ij}, $1 \le i \le m$, $1 \le j \le n$.[1]

Indeed, an arbitrary vector $\sum_{j=1}^{n} x_j\mathbf{e}_j$ transforms to

$$\begin{aligned} T\big(\sum_{j=1}^{n} x_j\mathbf{e}_j\big) &= \sum_{j=1}^{n} x_jT(\mathbf{e}_j) = \sum_{j=1}^{n} x_j \sum_{i=1}^{m} a_{ij}\mathbf{f}_i \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}x_j\mathbf{f}_i = \sum_{i=1}^{m} y_i\mathbf{f}_i, \end{aligned}$$

where y_i is defined by

$$y_i = \sum_{j=1}^{n} a_{ij}x_j. \tag{7.1.4}$$

□

In the summation of (7.1.3), the index i of $\mathbf{f}_i$ is the first index of a_{ij}. This may surprise you, but the computation shows that this is the right way to set up the notation in order to end up with the standard index order for matrix multiplication in (7.1.4) and therefore when we multiply out

$$\mathbf{y} = A\mathbf{x}. \tag{7.1.5}$$

[1] For more details, see [68], chapter 7, or [60] chapter 2.

This means that we can either talk about the linear transformation T or its matrix A when bases for both V and W have been given. We will determine later how the matrix A varies when the bases of V and W change.

7.1.6 Proposition. *Let U, V, and W be vector spaces of dimensions p, n, and m. Let T is a linear transformation from U to V, and S is a linear transformation from V to W.*

1. *The composite map $S \circ T$ from U to W is also a linear transformation.*
2. *Assume bases for U, V, and W have been selected and that the matrices of S and T in these bases are A and B, so A is an $m \times n$ matrix and B is an $n \times p$ matrix. Then the matrix for the composite transformation $S(T(\mathbf{u}))$ is the matrix product AB.*

Proof. The first point is easy. Let $\mathbf{u}_1$ and $\mathbf{u}_2$ be elements of U. Then, for instance

$$\begin{aligned} S \circ T(\mathbf{u}_1 + \mathbf{u}_2) &= S(T(\mathbf{u}_1 + \mathbf{u}_2) = S(T(\mathbf{u}_1) + T(\mathbf{u}_2)) \\ &= S(T(\mathbf{u}_1) + S(T(\mathbf{u}_2)) = S \circ T(\mathbf{u}_1) + S \circ T(\mathbf{u}_2)). \end{aligned}$$

The second part follows easily from the first part and from Proposition 7.1.2:

$$S \circ T(\mathbf{u}) = S(T(\mathbf{u})) = A(B\mathbf{u}) = (AB)\mathbf{u}$$

which proves the result. □

Often we refer to the linear transformation by its matrix. Then the nullspace $\mathcal{N}(A)$ of A is the set of $\mathbf{x} \in \mathbb{R}^n$ such that $A\mathbf{x} = \mathbf{0}$. The crucial fact is that it is a vector subspace of V of it has a dimension, called the nullity of the linear transformation.

The range of A, denoted $\mathcal{R}(A)$, is the linear subspace of $\mathbb{R}^m$ of elements $\mathbf{y}$ that can be written $A\mathbf{x}$, for all $\mathbf{x} \in \mathbb{R}^n$. The dimension of $\mathcal{R}(A)$ is called the *rank* r of A. The rank r is the maximal number of linearly independent columns of A.

We now get to one of the most important theorems in linear algebra.

7.1.7 Theorem (The Rank-Nullity Theorem). *Let $T\colon V \to W$ is a linear map between finite dimensional vector spaces, with matrix A. If n is the nullity of A, and r its rank, and d is the dimension of V, then*

$$n + r = d.$$

For a proof see see [60],§4.3, p.225.

7.2 The Four Subspaces Associated to a Matrix

This section describes the four key spaces associated to a $m \times n$ matrix A.[2] We write B for the transpose A^T of A, for simplicity of notation. The four subspaces are the range and nullspace of A, and the nullspace and the range of B. Their properties and interconnections are now summarized.

The transpose B of A is an $n \times m$ matrix. An important theorem in linear algebra says that the rank of B is equal to the rank of A; put differently, the row rank of A is equal to the column rank of A. We will establish this important below in the proof of Theorem 7.2.3.

Then the last two subspaces are $\mathcal{R}(B)$ and $\mathcal{N}(B)$. Note that $\mathcal{N}(A)$ and $\mathcal{R}(B)$ are subspaces of $\mathbb{R}^n$; while $\mathcal{N}(B)$ and $\mathcal{R}(A)$ are subspaces of $\mathbb{R}^m$.

We can summarize what we have obtained so far as:

7.2.1 Theorem.

$$\begin{aligned}\dim \mathcal{N}(A) + \dim \mathcal{R}(A) &= n;\\ \dim \mathcal{N}(B) + \dim \mathcal{R}(B) &= m;\\ \dim \mathcal{R}(A) &= \dim \mathcal{R}(B).\end{aligned}$$

In particular, the rank of A (and B) is at most the smaller of m and n,

Proof. The first two equalities are just the Rank-Nullity Theorem for A and B. The last equality says that the row rank and the column rank of a matrix are the same. We prove it by using the standard dot product on $\mathbb{R}^n$ and $\mathbb{R}^m$. First we make a definition:

7.2.2 Definition. Two subspaces V_1 and V_2 of $\mathbb{R}^n$ are *mutually orthogonal* if for any $\mathbf{v}_1 \in V_1$ and any $\mathbf{v}_2 \in V_2$, the standard inner product $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = 0$. We write $V_1 \perp V_2$ to indicate that the spaces are mutually orthogonal. V_1 and V_2 have *complementary dimensions* in $\mathbb{R}^n$ if $\dim V_1 + \dim V_2 = n$.

Take an element $\mathbf{x}_0$ in the nullspace of A, so $A\mathbf{x}_0 = 0$, and an element $\mathbf{x}_1$ in the range of B, so there exists $\mathbf{y}$ such that $\mathbf{x}_1 = B\mathbf{y}$. We compute the dot product

$$\langle \mathbf{x}_1, \mathbf{x}_0 \rangle = \mathbf{x}_1^T \mathbf{x}_0 = (\mathbf{y}^T A)\mathbf{x}_0 = \mathbf{y}^T (A\mathbf{x}_0) = 0$$

so that they are orthogonal and therefore linear independent. Repeating the same argument in $\mathbb{R}^m$ for B we get the following two inequalities:

$$\begin{aligned}\dim \mathcal{N}(A) + \dim \mathcal{R}(B) &\le n;\\ \dim \mathcal{N}(B) + \dim \mathcal{R}(A) &\le m;\end{aligned}$$

[2] An excellent reference for this material is Strang [67], whose entire presentation in §2.4 and §3.1 is organized around this approach (especially 3C on p. 136). Strang [68], §3.6, is also helpful.

Now use the first two equalities of Theorem 7.2.1 to eliminate the $\dim \mathcal{N}(A$ and $\dim \mathcal{N}(B)$. We are left with the inequalities $\dim \mathcal{R}(B) \leq \dim \mathcal{R}(A) \leq \dim \mathcal{R}(B)$, so row rank equals column rank. □

Theorem 7.2.1 shows that $\mathcal{N}(A)$ and $\mathcal{R}(B)$ have complementary dimensions in $\mathbb{R}^n$, and $\mathcal{N}(B)$ and $\mathcal{R}(A)$ have complementary dimensions in $\mathbb{R}^m$.

7.2.3 Theorem (The Four Subspaces Theorem)**.** *As before, B denotes the transpose of the $m \times n$ matrix A.*

- *$\mathcal{N}(A) \perp \mathcal{R}(B)$ in the domain $\mathbb{R}^n$ of A. Thus any element of $\mathbb{R}^n$ can be written uniquely as the sum of a vector of $\mathcal{N}(A)$ and a vector of $\mathcal{R}(B)$.*
- *$\mathcal{N}(B) \perp \mathcal{R}(A)$ in the domain $\mathbb{R}^m$ of B. Thus any vector of $\mathbb{R}^m$ can be written uniquely as the sum of a vector of $\mathcal{N}(B)$ and a vector of $\mathcal{R}(A)$.*

Proof. It is now almost trivial to prove the theorem. It is enough to prove the first statement, since $B^T = A$. The first and the third equations of Theorem 7.2.1 show that a basis for $\mathbb{R}^n$ can be formed by adjoining a basis of $\mathcal{R}(B)$ to a basis of $\mathcal{N}(A)$, if we can show that the basis elements of the two spaces are linearly independent. This is what we just established, so we are done. □

In the next result, note the use of the 'exclusive' *or* (see §2.1.2): either one or the other of the assertions is true, but not both. An inequality version of this result, with added positivity conditions, is known as the Farkas Alternative, and will be studied in §19.5.

7.2.4 Corollary. *Either*

1. *there is a vector $\mathbf{x}$ in $\mathbb{R}^n$ such that $A\mathbf{x} = \mathbf{b}$ has a solution*

   or (exclusive)

2. *there is a vector $\mathbf{y}$ in $\mathbb{R}^m$ with*

$$\mathbf{y}^T A = \mathbf{0} \qquad \textit{and} \qquad \mathbf{y}^T \mathbf{b} \neq 0.$$

Proof. If $A\mathbf{x} = \mathbf{b}$ has a solution, as in case 1, then $\mathbf{b} \in \mathcal{R}(A)$. If $\mathbf{y}^T A = 0$, as in case 2, then $\mathbf{y} \in \mathcal{N}(B)$. According to the Four Subspaces Theorem 7.2.3, $\mathcal{N}(B) \perp \mathcal{R}(A)$. This is contradicted by the condition $\mathbf{y}^T\mathbf{b} \neq 0$ in case 2, so the two cases are mutually exclusive. On the other hand, one of the two cases is always verified. If $\mathbf{b} \in \mathcal{R}(A)$, then we are in case 1. If not, then $\mathbf{b}$ can be written uniquely as $\mathbf{b}' + \mathbf{y}$, with $\mathbf{b}' \in \mathcal{R}(A)$ and $\mathbf{y}$ a non-zero element in $\mathcal{N}(B)$. Furthermore $\mathbf{b}' \perp \mathbf{y}$ by the Four Subspaces Theorem. Then

$$\mathbf{y}^T\mathbf{b} = \mathbf{y}^T(\mathbf{b}' + \mathbf{y}) = \mathbf{y}^T\mathbf{y} \neq 0,$$

so we are in case 2. □

7.2.5 Remark. By replacing $\mathbf{y}$ by a scalar multiple, we can assume that $\mathbf{y}^T\mathbf{b} = 1$ in the alternative of Corollary 7.2.4.

7.2.6 Proposition. *In terms of the dot product $\langle *, * \rangle_1$ on $\mathbb{R}^n$, and the dot product $\langle *, * \rangle_2$ on $\mathbb{R}^m$), we have*

$$\langle \mathbf{y}, A\mathbf{x} \rangle_2 = \langle A^T\mathbf{y}, \mathbf{x} \rangle_1 \tag{7.2.7}$$

Proof.

$$\begin{aligned}
\langle \mathbf{y}, A\mathbf{x} \rangle_2 &= \mathbf{y}^T A\mathbf{x} && \text{(switching to matrix multiplication in } W\text{)},\\
&= (A^T\mathbf{y})^T\mathbf{x} && \text{(by Proposition 6.2.8)},\\
&= \langle \mathbf{A}^T\mathbf{y}, \mathbf{x} \rangle_1 && \text{(switching back to dot product in } V\text{)},
\end{aligned}$$

so we are done. □

We will use an important special case of this result in Theorem 7.4.5.

7.2.8 Example. Assume $m < n$, and let A be a $m \times n$ matrix of rank m, the maximum possible. Then the nullspace of A has dimension $n - m$, and the nullspace of B has dimension 0. This is the case we consider in linear optimization. There, as we will see in Lecture 25, if A is the matrix of the primal problem with variable $\mathbf{x}$, B is the matrix of the dual problem with variable $\mathbf{y}$.

In least squares optimization (§13.3) we consider the case $m > n$ and A of maximal rank.

7.3 The Characteristic Polynomial

Next we consider the characteristic polynomial of a square matrix.[3]

[3] Note that some authors (in particular [60] and [67]) define the characteristic polynomial as $\det(A - tI_n)$, while others (in particular [25], [39], [57], and [13]) use the one given here. It is a simple exercise using properties of the determinant to show that the two possible definitions differ by a factor of $(-1)^n$.

7.3.1 Definition. The *characteristic polynomial* $p(t)$ of the $n \times n$ matrix A is

$$\det(tI_n - A)$$

where, as usual, I_n is the $n \times n$ identity matrix, and t is a variable.

$p(t)$ is easily seen to be a polynomial of degree n in t, which is written

$$p(t) = t^n - p_1 t^{n-1} + p_2 t^{n-2} + \dots + (-1)^n p_n \tag{7.3.2}$$

where p_1 is the *trace* of A, namely the sum of the diagonal elements of A:

$$p_1 = \operatorname{Tr} A = \sum_{i=1}^{n} a_{ii}$$

and

$$p_n = (-1)^n \det A,$$

as we see by setting $t = 0$.

7.3.3 Exercise. Compute the characteristic polynomials of the matrices A and B from Example 7.8.11.

Finally, we consider the trace of a product of $n \times n$ matrices AB. The (i, i)-th entry of $\operatorname{Tr} AB$ is $\sum_k a_{ik} b_{ki}$, so the trace, which is the sum of all the diagonal elements, is

$$\sum_{i=1}^{n} \sum_{k=1}^{n} a_{ik} b_{ki}$$

7.3.4 Theorem.

$$\operatorname{Tr} AB = \operatorname{Tr} BA \tag{7.3.5}$$

Proof.

$$\operatorname{Tr} AB = \sum_{i=1}^{n} \sum_{k=1}^{n} a_{ik} b_{ki} = \sum_{k=1}^{n} \sum_{i=1}^{n} b_{ki} a_{ik} = \operatorname{Tr} BA$$

□

7.3.6 Remark. A word of warning: When you have three matrices A, B, C, Theorem 7.3.4 shows that $\operatorname{Tr} ABC = \operatorname{Tr} BCA = \operatorname{Tr} CAB$. It does not show that $\operatorname{Tr} ABC = \operatorname{Tr} ACB$. Indeed, that is not true in general.

On the other hand, it is obvious that $\operatorname{Tr}(A + B) = \operatorname{Tr} A + TrB$, and that extends to a sum with any number of terms.

7.4 Eigenvalues and Eigenvectors

7.4.1 Definition. A complex number λ is called an *eigenvalue* of A if there is a non-zero complex vector $\mathbf{x}$ such that $A\mathbf{x} = \lambda\mathbf{x}$. The vector $\mathbf{x}$ is an *eigenvector* associated to λ.

7.4.2 Proposition. *An eigenvalue of A is a root of the characteristic polynomial of A.*

Proof. The definition says that $\mathbf{x}$ is a nonzero element in the nullspace of the complex matrix $\lambda I_n - A$, which means that $\lambda I_n - A$ is not invertible, so that its determinant is zero. But the determinant of $\lambda I_n - A$ is the characteristic polynomial of A (up to sign) evaluated at λ. Thus λ is a root of the characteristic polynomial. $\square$

7.4.3 Example. We compute the eigenvalues of the matrix A:

$$\begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}$$

We first compute the characteristic polynomial, namely the determinant of

$$\begin{bmatrix} t-1 & -2 \\ -1 & t-3 \end{bmatrix}$$

getting $t^2 - 4t + 1$. The discriminant of this quadratic is positive, so you get two real eigenvalues (the roots), and then two real eigenvectors. The end of the computation is left to you.

7.4.4 Example. Consider any 3×3 orthogonal matrix Q with determinant 1. We show it is a rotation. The characteristic polynomial $p(t)$ (7.3.2) of Q has degree 3. Because the highest degree term of $p(t)$ is t^3, it goes to ∞ as t gets large. Because $\det Q = 1$, $p(0) = -1$. Thus the graph of $p(t)$ crosses the t-axis for a positive value of t. This means Q has a real eigenvector $\mathbf{e}^1$ with associated positive eigenvalue λ, so $Q\mathbf{e}_1 = \lambda\mathbf{e}_1$. Because Q is orthogonal and preserves distance, $\lambda = 1$. Let H be the orthogonal complement of the line L spanned by $\mathbf{e}_1$. Next we show that for all $\mathbf{v} \in H$, then $Q\mathbf{v}$ is perpendicular to $\mathbf{e}_1$, so $Q\mathbf{v} \in H$. Indeed,

$$\begin{aligned} \langle Q\mathbf{v}, \mathbf{e}_1 \rangle &= \langle Q\mathbf{v}, Q\mathbf{e}_1 \rangle && \text{using the eigenvector equation for } \mathbf{e}_1, \\ &= \langle \mathbf{v}, \mathbf{e}_1 \rangle && \text{by Proposition 6.7.5,} \\ &= 0 && \text{since } \mathbf{v} \perp \mathbf{e}_1. \end{aligned}$$

It is then an easy exercise to show that the restriction of A to H is an ordinary rotation in the plane. The line L is called the *axis of rotation*. It is uniquely determined unless the rotation is trivial.

We conclude this short section with a result concerning the eigenvectors of symmetric matrices.

7.4.5 Theorem. *Let A be a square matrix of size n. A is symmetric if and only if*

$$\langle \mathbf{x}, A\mathbf{y}\rangle = \langle A\mathbf{x}, \mathbf{y}\rangle, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{7.4.6}$$

so that the matrix A can migrate from one side of the dot product to the other.

Proof. By Proposition 7.2.6, this is obviously true if A is symmetric. Now pick for $\mathbf{x}$ the unit vector $\mathbf{e}_i$, and for $\mathbf{y}$ the unit vector $\mathbf{e}_j$. Then $\mathbf{e}_i^T A\mathbf{e}_j = a_{ij}$ and $\mathbf{e}_i^T A^T\mathbf{e}_j = a_{ji}$, so that being symmetric is a necessary condition. □

This allows us to give a simple proof of the following proposition, which we reprove later using the Spectral Theorem 9.2.1.

7.4.7 Proposition. *Assume the matrix A is symmetric. Let $\mathbf{e}_1$ and $\mathbf{e}_2$ be eigenvectors of A with eigenvalue λ_1 and λ_2. Assume $\lambda_1 \neq \lambda_2$. Then $\mathbf{e}_1$ and $\mathbf{e}_2$ are orthogonal, namely: $\langle \mathbf{e}_1, \mathbf{e}_2\rangle = 0$.*

Proof.

$$\langle \mathbf{e}_1, A\mathbf{e}_2\rangle = \langle \mathbf{e}_1, \lambda_2\mathbf{e}_2\rangle = \lambda_2\langle \mathbf{e}_1, \mathbf{e}_2\rangle$$

By Theorem 7.4.5 this is equal to

$$\langle A\mathbf{e}_1, \mathbf{e}_2\rangle = \langle \lambda_1\mathbf{e}_1, \mathbf{e}_2\rangle = \lambda_1\langle \mathbf{e}_1, \mathbf{e}_2\rangle$$

so $\lambda_1\langle \mathbf{e}_1, \mathbf{e}_2\rangle = \lambda_2\langle \mathbf{e}_1, \mathbf{e}_2\rangle$. Since $\lambda_1 \neq \lambda_2$, this can only be true if $\langle \mathbf{e}_1, \mathbf{e}_2\rangle = 0$. □

At this point, we do not know if the λ are real: the Spectral Theorem 9.2.1 will establish that.

7.5 Distance Minimization via Projection

Let us take for V the space $\mathbb{R}^n$ with its dot product. Let K be a subspace of V of dimension k. Let $\mathbf{v}_1, \dots, \mathbf{v}_k$ be an orthogonal basis of K, meaning that $\mathbf{v}_i \cdot \mathbf{v}_j = 0$ whenever $i \neq j$. Such a basis can be found by the Gram-Schmidt orthogonalization process.

Now let $\mathbf{b}$ be any vector in V. Let $c_i \in \mathbb{R}$ be the component of $\mathbf{b}$ along $\mathbf{v}_i$, so that $\mathbf{b} - c_i\mathbf{v}_i$ is orthogonal to $\mathbf{v}_i$: see Definition 5.4.25.

Let $\mathbf{c} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k$. Then

7.5.1 Lemma. $\mathbf{b} - \mathbf{c}$ *is orthogonal to all vectors in K.*

Proof. By linearity it is sufficient to show that it is orthogonal to each basis element of K. Since the basis is orthogonal, we get

$$\langle \mathbf{b} - \sum_{i=1}^{k} c_i \mathbf{v}_i, \mathbf{v}_j \rangle = \langle \mathbf{b} - c_j \mathbf{v}_j, \mathbf{v}_j \rangle = 0$$

by definition of c_j, the component of $\mathbf{b}$ along $\mathbf{v}_j$. □

From this we derive

7.5.2 Theorem. *The point* $\mathbf{c}$ *is the unique point of* K *closest to* $\mathbf{b}$.

Proof. Any point $\mathbf{z} \in K$ can be written

$$\mathbf{z} = x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + \cdots + x_k\mathbf{v}_k$$

for real numbers x_i. Then

$$\begin{aligned} \|\mathbf{b} - \mathbf{z}\|^2 &= \|\mathbf{b} - \mathbf{c} + \mathbf{c} - \mathbf{z}\|^2 \\ &= \|\mathbf{b} - \mathbf{c}\|^2 + \|\mathbf{c} - \mathbf{z}\|^2 \\ &> \|\mathbf{b} - \mathbf{c}\|^2 \end{aligned}$$

unless $\mathbf{z} = \mathbf{c}$. In the second line we used the Pythagorean Theorem 5.4.23, which applies because $\mathbf{c} - \mathbf{z}$ is in K. Since $\|\mathbf{v} - \mathbf{z}\|$ is the distance of $\mathbf{v}$ from $\mathbf{z}$, we are done. □

Thus we have solved this optimization problem in completely elementary fashion. For more in this direction, see §13.3 on Least Squares.

7.6 Example: Orthogonal Projection

We examine what we just did using matrices.

Given an n-dimensional vector space V with an inner product, assume that you have two mutually orthogonal subspaces K and R of complementary dimensions, meaning that if k is the dimension of K and r the dimension of R, then $k + r = n$. See Definition 7.2.2. Then any element $\mathbf{v} \in V$ can be written uniquely as $\mathbf{v} = \mathbf{k} + \mathbf{r}$, where $\mathbf{k} \in K$ and $\mathbf{r} \in R$. Since K and R are orthogonal, $\langle \mathbf{k}, \mathbf{r} \rangle = 0$.

7.6.1 Definition. The *orthogonal projection* of V to R is the linear transformation P from V to V that sends $\mathbf{v}$ to $\mathbf{r}$.

You should check that this is a linear transformation, and this this is what we did in the previous section.

We want to understand the $n \times n$ matrix of P for any choice of basis of V that preserves length and distance.

We start with the basis $\mathbf{e}$ where the first r basis elements are an orthonormal basis for R, and the remaining basis elements form an orthonormal basis for K. Then the matrix for P in this basis can be written in block form (see §6.10) as

$$A_r = \begin{bmatrix} I_r & 0_{rk} \\ 0_{kr} & 0_k \end{bmatrix}$$

where I_r is the $r \times r$ identity matrix, and the other matrices are all zero matrices of size given by the subscripts.

Note, using block multiplication, that $A_r^2 = A$.

Now suppose you take any other matrix representation of the orthogonal projection P. Since we want length and angle to be preserved, we should only allow orthogonally similar matrices, so we should look at matrices $Q^{-1}AQ$, where Q is orthogonal.

7.6.2 Theorem. *The matrix A for an orthogonal projection is symmetric and satisfies $A^2 = A$.*

Conversely, any matrix A that satisfies these two properties is the matrix of the orthogonal projection to the range of A.

Proof. If A is the matrix of an orthogonal projection, it can be written $Q^{-1}A_rQ$, for some r and for an orthogonal matrix Q, so $Q^{-1} = Q^T$. First we show A is symmetric. Note that A_r is symmetric, so $A_r^T = A_r$. So

$$A^T = (Q^{-1}A_rQ)^T = Q^T A_r^T (Q^{-1})^T = Q^{-1}A_rQ = A$$

so that is established.

Next,

$$A^2 = Q^{-1}A_rQQ^{-1}A_rQ = Q^{-1}A_r^2Q = Q^{-1}A_rQ = A$$

as required.

Now the converse. When A is a projection, for any $\mathbf{v} \in V$ the difference $\mathbf{v} - A\mathbf{v}$ is perpendicular to $A\mathbf{w}$, for all $\mathbf{w}$, and this characterizes projections. We can establish this just using our two properties, proving what we want.

$$(\mathbf{v} - A\mathbf{v})^T A\mathbf{w} = \mathbf{v}^T A\mathbf{w} - \mathbf{v}^T A^T A\mathbf{w} = \mathbf{v}^T A^2\mathbf{w} - \mathbf{v}^T A\mathbf{w} = 0.$$

□

Orthogonal projections will be used when we study the method of least squares in §13.3.

7.7 Similar Matrices

You may want to review equivalence relations in Definition 2.2.8 at this point.

7.7.1 Definition. The $n \times n$ matrix A is *similar* to the $n \times n$ matrix B if there is an $n \times n$ invertible matrix C such that

$$B = C^{-1}AC \tag{7.7.2}$$

We will see in Theorem 7.8.10 that two matrices are similar if and only if they are the matrices of the same linear transformation, but viewed in different bases.

We indicate that A is similar to B by $A \sim B$.

7.7.3 Theorem. *Similarity is an equivalence relation on $n \times n$ matrices.*

Proof. To prove that $\sim$ is an equivalence relation, we need to establish the following three points:

- $A \sim A$:

  Use the identity matrix for C.

- if $A \sim B$, then $B \sim A$:

  If $A \sim B$, there is an invertible C such that $B = C^{-1}AC$. Then, multiplying both sides of the equation on the right by C^{-1} and on the left by C, and letting $D = C^{-1}$, we see that $A = D^{-1}BD$, so B is similar to A.

- if $A \sim B$ and $B \sim D$, then $A \sim D$:

  The hypotheses mean that there are invertible matrices C_1 and C_2 such that $B = C_1^{-1}AC_1$ and $D = C_2^{-1}BC_2$, so, substituting from the first equation into the second, we get

  $$D = C_2^{-1}C_1^{-1}AC_1C_2 = (C_1C_2)^{-1}AC_1C_2$$

  so A is similar to D using the matrix C_1C_2.

□

Since similarity is an equivalence relation on $n \times n$ matrices, it partitions symmetric matrices into *equivalence classes*.

Our next goal is to determine the common properties of similar matrices. Here is the key theorem.

7.7.4 Theorem. *Two similar matrices have the same characteristic polynomial and therefore the same trace and determinant. They also have the same eigenvalues (see §7.4), since these are the roots of the characteristic polynomial.*

Proof. If A and B are similar, there is an invertible matrix C such that $B = C^{-1}AC$. Then

$$C^{-1}(tI_n - A)C = C^{-1}(tI_n)C - C^{-1}AC = tI_n - B \tag{7.7.5}$$

Take determinants on both sides, using the theorem that the determinant of a product is the product of the determinants:

$$\det\left(C^{-1}(tI_n - A)C\right) = \det C^{-1} \det\left(tI_n - A\right) \det C \tag{7.7.6}$$

Because $\det C^{-1} = \frac{1}{\det C}$ we are left with $\det\left(tI_n - A\right)$ and therefore taking the determinant on both sides of (7.7.5) we get

$$\det\left(tI_n - A\right) = \det\left(tI_n - B\right) \tag{7.7.7}$$

which shows that the two characteristic polynomials are the same. □

7.7.8 Definition. If A and B are $n \times n$ matrices, then A is *orthogonally similar* to B if there exists an $n \times n$ orthogonal matrix Q such that $B = Q^{-1}AQ$.

7.7.9 Theorem. *Orthogonal similarity is an equivalence relation on $n \times n$ matrices. Furthermore the equivalence classes of orthogonal similarity are finer than those of similarity, meaning that any two matrices that are orthogonally similar are similar, but not conversely.*

7.7.10 Exercise. Prove Theorem 7.7.9 by imitating the proof of Theorem 7.7.3. You will need Exercise 6.7.4.

Given any $n \times n$ matrix A and a permutation σ on $\{1, 2, \dots, n\}$, we define a new matrix A^σ by

$$A^\sigma = (P^\sigma)^{-1}AP^\sigma, \tag{7.7.11}$$

where P^σ is the permutation matrix association to σ: see Definition 6.5.1. Because permutation matrices are orthogonal, we see that A and A^σ are orthogonally similar and therefore similar. So the next result follows from Theorem 7.7.4.

7.7.12 Proposition. *A and A^σ have the same characteristic polynomial and the same eigenvalues.*

Let a^σ_{ij} be the ij-th entry of A^σ.

7.7.13 Exercise. Show that $a^\sigma_{ij} = a_{\sigma(i)\sigma(j)}$

7.7.14 Example. Using the permutation given in cycle notation by (123), and its permutation matrix computed in Example 6.5.2, we multiply out

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} a_{22} & a_{23} & a_{21} \\ a_{32} & a_{33} & a_{31} \\ a_{12} & a_{13} & a_{11} \end{bmatrix}$$

so we get the formula claimed in Exercise 7.7.13.

7.7.15 Exercise. Once you have read the definitions in §6.8, show that every principal submatrix of A is the leading principal submatrix of A^σ for a suitable permutation σ.

7.8 Change of Basis

We now take a linear transformation T from an n-dimensional vector space V to itself, and ask how the matrix of T varies as we vary the basis of V.

So pick a basis $\mathbf{e}^1, \dots, \mathbf{e}^n$ of V. In this basis, we have as per (7.1.3),

$$T(\mathbf{e}_j) = \sum_{i=1}^{n} a_{ij}\mathbf{e}_i,$$

so we get a square $n \times n$ matrix A. We now write in shorthand (we only use this notation in sections involving change of basis):

$$[T]_{\mathbf{e}} = A. \tag{7.8.1}$$

So $[T]_{\mathbf{e}}$ denotes the matrix of the linear transformation T in the $\mathbf{e}$-basis.

Next, if $\mathbf{v} = \sum_{j=1}^n x_j\mathbf{e}_j$, we write, in shorthand, $[\mathbf{v}]_{\mathbf{e}}$ for the column vector $\mathbf{x}$ with entries $x_1, \dots, x_n$. In this notation, (7.1.5) becomes

$$[T(\mathbf{v})]_{\mathbf{e}} = [T]_{\mathbf{e}}[\mathbf{v}]_{\mathbf{e}} = A\mathbf{x}. \tag{7.8.2}$$

Now suppose that we have a second basis $\mathbf{f}_1, \dots, \mathbf{f}_n$ of V. Let B be the matrix of T in this basis, so, using the same notation as in (7.8.1)

$$[T]_{\mathbf{f}} = B. \tag{7.8.3}$$

We can write the same vector $\mathbf{v}$ in terms of this basis, and get $\mathbf{v} = \sum_{j=1}^n z_j\mathbf{f}_j$, so in our shorthand: $[\mathbf{v}]_{\mathbf{f}} = \mathbf{z}$, and exactly as in (7.8.2) we get

$$[T(\mathbf{v})]_{\mathbf{f}} = [T]_{\mathbf{f}}[\mathbf{v}]_{\mathbf{f}} = B\mathbf{z}. \tag{7.8.4}$$

To determine the relationship of the two $n \times n$ matrices A and B, we need to understand the relationship between $[\mathbf{v}]_{\mathbf{f}}$ and $[\mathbf{v}]_{\mathbf{e}}$. Because $\mathbf{f}_1, \dots, \mathbf{f}_n$ is a basis of V, we can write each one of the $\mathbf{e}_j$ uniquely in terms of it, in other words:

$$\mathbf{e}_j = \sum_{i=1}^{n} d_{ij}\mathbf{f}_i. \tag{7.8.5}$$

This gives us a $n \times n$ matrix $D = [d_{ij}]$. D is invertible because the $\mathbf{e}_1, \dots, \mathbf{e}_n$ also form a basis. That is all we can say about D: any invertible matrix can be used to change basis.

To see how the coordinates transform, write as before

$$\mathbf{v} = \sum_{j=1}^{n} x_j \mathbf{e}_j = \sum_{j=1}^{n} z_j \mathbf{f}_j$$

so substituting the values of $\mathbf{e}_j$ from (7.8.5), we get

$$\sum_{j=1}^{n} x_j \mathbf{e}_j = \sum_{j=1}^{n} x_j \sum_{i=1}^{n} d_{ij}\mathbf{f}_i = \sum_{i=1}^{n} \Big(\sum_{j=1}^{n} d_{ij} x_j \Big) \mathbf{f}_i$$

so that $z_i = \sum_{j=1}^{n} d_{ij} x_j$. In matrix notation,

$$\mathbf{z} = D\mathbf{x}. \tag{7.8.6}$$

Because of this equation, we call D the *change of basis matrix* from $\mathbf{e}$-coordinates to $\mathbf{f}$-coordinates. Indeed, the i-th column of D gives the expression of $\mathbf{e}_i$ in the $\mathbf{f}$-basis. In our shorthand, we have

$$[\mathbf{v}]_{\mathbf{f}} = D[\mathbf{v}]_{\mathbf{e}}. \tag{7.8.7}$$

Applying this to $T(\mathbf{v})$ instead of $\mathbf{v}$, we get

$$[T(\mathbf{v})]_{\mathbf{f}} = D[T(\mathbf{v})]_{\mathbf{e}}. \tag{7.8.8}$$

So, combining these equations,

$$\begin{aligned}
D[T]_{\mathbf{e}}[\mathbf{v}]_{\mathbf{e}} &= D[T(\mathbf{v})]_{\mathbf{e}} && \text{by (7.8.2)} \\
&= [T(\mathbf{v})]_{\mathbf{f}} && \text{by (7.8.8)} \\
&= [T]_{\mathbf{f}}[\mathbf{v}]_{\mathbf{f}} && \text{by (7.8.4)} \\
&= [T]_{\mathbf{f}} D[\mathbf{v}]_{\mathbf{e}} && \text{by (7.8.7)}
\end{aligned}$$

Now multiply on the left by the inverse of D, to get

$$[T]_{\mathbf{e}}[\mathbf{v}]_{\mathbf{e}} = D^{-1}[T]_{\mathbf{f}}D[\mathbf{v}]_{\mathbf{e}}.$$

Using (7.8.1) and (7.8.3), this says

$$A\mathbf{x} = D^{-1}BD\mathbf{x}.$$

Since this is true for all $\mathbf{x}$, we have

$$A = D^{-1}BD, \text{ so } B = DAD^{-1}, \tag{7.8.9}$$

the desired result. Notice what this says: to get the matrix B of T in the $\mathbf{f}$-basis, starting with a vector expressed in the $\mathbf{f}$-basis, first convert the vector into $\mathbf{e}$-coordinates using D^{-1}, then apply the matrix A of T in the $\mathbf{e}$-basis, and then convert the resulting expression back to the $\mathbf{f}$-basis using D: reading the right-hand side of (7.8.9) from right to left. We have proved the following theorem.

7.8.10 Theorem. *Two $n \times n$ matrices A and B are similar if and only if they are the matrix of the same linear transformation T from an n-dimensional vector space to itself in different bases.*

Therefore when studying linear transformation, what is important is not the matrix of A, but the similarity class of that matrix. By Theorem 7.7.4, the characteristic polynomial and all the expressions that can be deduced from the characteristic polynomials are invariants of the similarity class.

7.8.11 Example. Suppose that V is two dimensional, with basis $\mathbf{e}_1$,$\mathbf{e}_2$, and that $T(\mathbf{e}_1) = \mathbf{e}_1 + \mathbf{e}_2$, and $T(\mathbf{e}_2) = 2\mathbf{e}_1 + 3\mathbf{e}_2$. Then the matrix $A = [T]_{\mathbf{e}}$ is

$$\begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}$$

Now suppose we have a second basis $\mathbf{f}_1$, $\mathbf{f}_2$, with

$$\begin{aligned} \mathbf{e}_1 &= (\mathbf{f}_1 + \mathbf{f}_2)/2 \\ \mathbf{e}_2 &= (\mathbf{f}_1 - \mathbf{f}_2)/2 \end{aligned}$$

so the matrix D is

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix}$$

We can solve explicitly for the $\mathbf{f}$ in terms of $\mathbf{e}$:

$$\begin{aligned} \mathbf{f}_1 &= \mathbf{e}_1 + \mathbf{e}_2 \\ \mathbf{f}_2 &= \mathbf{e}_1 - \mathbf{e}_2 \end{aligned}$$

so we see directly that D^{-1} is

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

Now

$$\begin{aligned} T(f^1) &= T(\mathbf{e}^1 + \mathbf{e}^2) = 3\mathbf{e}^1 + 4\mathbf{e}^2 = (7f^1 - f^2)/2 \\ T(f^2) &= T(\mathbf{e}^1 - \mathbf{e}^2) = -\mathbf{e}^1 - 2\mathbf{e}^2 = (-3f^1 + f^2)/2 \end{aligned}$$

so the matrix $B = [T]_{\mathbf{f}}$ of T is this basis is

$$\frac{1}{2}\begin{bmatrix} 7 & -3 \\ -1 & 1 \end{bmatrix}$$

and you should check directly that (7.8.9) is verified.

And now for some exercises on a generalization on these concepts.

7.8.12 Exercise. Show that the set of all $n \times m$ real matrices forms a vector space of dimension nm. Write down a basis for this vector space.

Call this vector space $M(n, m)$.

As we saw above, each element $A \in M(n, m)$ represents a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. Imagine you make a change of basis in $\mathbb{R}^n$, and a change of basis in $\mathbb{R}^m$. How does the matrix of the associated linear transformation change?

A change of basis in $\mathbb{R}^n$ is given by an invertible $n \times n$ matrix B. Similarly, a change of basis in $\mathbb{R}^m$ is given by an invertible $m \times m$ matrix C.

7.8.13 Definition. Two $n \times m$ matrices A_1 and A_2 are *equivalent*, written $A_1 \equiv A_2$, if there is a invertible $n \times n$ matrix B and an invertible $m \times m$ matrix C such that

$$A_1 = BA_2C \tag{7.8.14}$$

7.8.15 Exercise. Show that this is an equivalence relation.

7.8.16 Exercise. In which ways does this generalize the equivalence relation on $n \times n$ matrices given in Definition 7.7.1. In which ways is it different?

7.8.17 Exercise. Show that Gaussian elimination and backsubstitution reduces any $n \times m$ matrix A to an equivalent matrix A'.

Lecture 8

Symmetric Matrices

The main object of study in this course is the objective function $f(\mathbf{x})$ of an optimization problem. We want to determine when the point $\mathbf{x} = \mathbf{a}$ is a minimizer for the problem, in other words, when $f(\mathbf{a})$ is the minimum value of f on the feasible set. As always we assume that we are in $\mathbb{R}^n$. We will usually assume that f is sufficiently differentiable, usually meaning at least twice continuously differentiable ($\mathcal{C}^2$).

If we consider only points $\mathbf{a}$ in the interior of the feasible set, then (as we will see in Theorem 13.1.1) $f(\mathbf{a})$ is a local minimum only if the gradient at $\mathbf{a}$ is zero: $\nabla f(\mathbf{a}) = \mathbf{0}$.

Next Taylor's Theorem 12.4.8 tells us that in a neighborhood of any point $\mathbf{a}$ in the interior of its domain, the $\mathcal{C}^2$-function f can be approximated by its second order Taylor polynomial

$$f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot (\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^T F(\mathbf{a})(\mathbf{x} - \mathbf{a})$$

where $F(\mathbf{a})$ is the Hessian of f evaluated at $\mathbf{a}$, thus a symmetric $n \times n$ matrix.

At a point $\mathbf{a}$ where the gradient $\nabla f(\mathbf{a})$ vanishes we are left with

$$f(\mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^T F(\mathbf{a})(\mathbf{x} - \mathbf{a})$$

By making the change of variable $\mathbf{y} = \mathbf{x} - \mathbf{a}$ that brings the point $\mathbf{a}$ to the origin, and ignoring the constant $f(\mathbf{a})$, we get

$$\frac{1}{2}\mathbf{y}^T F(\mathbf{a})\mathbf{y}.$$

This is called a quadratic form, and it gives rise to a symmetric matrix. In fact it gives rise to many symmetric matrices, depending on the basis used for the vector space. All these matrices, which are said to represent the quadratic form, are members of one congruence class (see Definition 8.4.1 and Proposition 8.3.1). Congruence is an important equivalence relation on symmetric matrices for precisely this reason. For optimization it is crucial to understand where a given quadratic form fits in the classification of quadratic forms.

Our first result, Theorem 8.3.2, says that there is a diagonal matrix D in every congruence class, with diagonal elements either 1, 0 or -1. Then Sylvester's Law of Inertia 8.5.5 shows that for a given quadratic form, the number of 1, 0 and -1 in its diagonal representative does not depend on the diagonal matrix found: this yields important invariants for each quadratic form. It is the main result of the lecture.

Section 8.6 are devoted to Gauss-Jordan diagonalization techniques for quadratic forms. These techniques give an efficient computational approach for finding a diagonal matrix representing the quadratic form, as opposed to Theorem 8.3.2, which just gives an existence result; and the Spectral Theorem, where the computation is much less efficient. This is the approach one should use when computing by hand or by machine.

8.1 Quadratic Forms

8.1.1 Definition. A *quadratic form* is a function from $\mathbb{R}^n$ to $\mathbb{R}$ given by

$$q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{n} \Big(\sum_{j=1}^{n} a_{ij} x_i x_j \Big), \tag{8.1.2}$$

where $A = (a_{ij})$ is a $n \times n$ symmetric matrix, so $a_{ij} = a_{ji}$ for all $i \neq j$, and $\mathbf{x} \in \mathbb{R}^n$.

It is called quadratic because $q(t\mathbf{x}) = t^2 q(\mathbf{x})$ for all real number t. In other words q is a homogeneous function of degree 2.

8.1.3 Example. Let $D(d_1, \dots, d_n)$ be a $n \times n$ diagonal matrix. Then

$$q(\mathbf{x}) = d_1 x_1^2 + d_2 x_2^2 + \cdots + d_n x_n^2,$$

as you should check.

8.1.4 Example. Let $q(x_1, x_2, x_3) = x_1^2 + 2x_1x_2 - x_1x_3 - x_2^2 + x_2x_3 + 4x_3^2$. The associated matrix A is

$$\begin{bmatrix} 1 & 1 & -1/2 \\ 1 & -1 & 1/2 \\ -1/2 & 1/2 & 4 \end{bmatrix}$$

as you should check by carrying out the matrix multiplication $\mathbf{x}^T A\mathbf{x}$.

8.1.5 Remark. In this example, note that the off-diagonal terms in the matrix are half the coefficients in the quadratic polynomial. This is because we have not written separate coefficients for $x_i x_j$ and $x_j x_i$, as we have in the sum in (8.1.2) If we wrote the summation differently (check the indexing) we would have:

$$q(\mathbf{x}) = \sum_{i=1}^{n} \Bigl(\sum_{j=i}^{n} b_{ij} x_i x_j \Bigr).$$

For $i \neq j$, $b_{ij} = 2a_{ij}$, while $b_{ii} = a_{ii}$.

Here are the main definitions concerning quadratic forms.

8.1.6 Definition. The quadratic form $q(\mathbf{x})$ with matrix A is *definite* if $q(\mathbf{x}) \neq 0$ for all $\mathbf{x} \neq \mathbf{0}$.

We can refine this classification as follows.

8.1.7 Definition. The quadratic form $q(\mathbf{x})$ with matrix A is

- *Positive definite* if $\forall \mathbf{x} \neq \mathbf{0}$, $q(\mathbf{x}) > 0$, or, equivalently, $\mathbf{x}^T A \mathbf{x} > 0$;
- *Positive semidefinite* if $\forall \mathbf{x}$, $q(\mathbf{x}) \geq 0$, or, equivalently, $\mathbf{x}^T A \mathbf{x} \geq 0$;
- *Negative definite* if $\forall \mathbf{x} \neq \mathbf{0}$, $q(\mathbf{x}) < 0$, or, equivalently, $\mathbf{x}^T A \mathbf{x} < 0$;
- *Negative semidefinite* if $\forall \mathbf{x}$, $q(\mathbf{x}) \leq 0$, or, equivalently, $\mathbf{x}^T A \mathbf{x} \leq 0$;
- *Indefinite* if it does not fall into one of the four previous cases.

8.1.8 Example. The matrix

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

associated to the quadratic form $q = x_1^2 - x_2^2$ is indefinite, because

$$\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 1, \text{ while } \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = -1$$

We pursue this in Example 8.5.1.

We can summarize the definitions above in the language of optimization theory.

8.1.9 Proposition. *We seek to optimize the quadratic form $q(\mathbf{x})$ on $\mathbb{R}^n$. Its critical points are the points $\mathbf{x}$ for which $A\mathbf{x} = \mathbf{0}$, in other words the nullspace of A. The origin is always a critical point. Note that $q(\mathbf{0}) = 0$.*

Then $q(\mathbf{x})$ has a strict minimum at $\mathbf{x} = \mathbf{0}$ if it is positive definite; a minimum if it is positive semidefinite.

It has a strict maximum at $\mathbf{x} = \mathbf{0}$ if it is negative definite; a maximum if it is negative semidefinite.

Finally it has a saddle point*, if it is indefinite, meaning that in a neighborhood of 0 there are both negative and positive values of $q(\mathbf{x})$.*

Determining where a given quadratic form fits in this list is therefore important for optimization.

Where does the quadratic form 8.1.4 fit in the classification? Because $q(\epsilon, 0, 0) = \epsilon^2$, and $q(0, \epsilon, 0) = -\epsilon^2$, we see that there are points arbitrarily close to the origin that take positive values, and points close to the origin where the function takes negative values. So we are at a saddle point.

Going back to the function $f(\mathbf{x})$, we see that if $q(\mathbf{x})$ is the quadratic term of the Taylor expansion of f, the matrix A is half the Hessian $F(\mathbf{a})$ of f at $\mathbf{a}$.

If the matrix A has maximum rank n (which means it is invertible) we will learn in §9.4 that the local behavior of the original function f at a critical point is determined by its quadratic approximation q, and therefore by the matrix A.

For the rest of this lecture we do linear algebra, forgetting about the function f, considering only the quadratic form

$$q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} \tag{8.1.10}$$

where A is a symmetric matrix.

8.2 Symmetric Bilinear Forms

Quadratic forms are naturally paired with symmetric bilinear forms, which we now define.

8.2.1 Definition. A *symmetric bilinear form* is a real valued map from two copies of $\mathbb{R}^n$, the first with variables $\mathbf{x}$, the second with variables $\mathbf{y}$. It is given by

$$b(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T A\mathbf{y},$$

where A is a symmetric $n \times n$ matrix.

8.2.2 Example. As we noted in Definition 5.4.1, the standard inner product on $\mathbb{R}^n$ is an example of symmetric bilinear map. Its matrix A is the identity matrix, and its associated quadratic form is the square of the norm.

8.2.3 Remark. We generalize what we did for inner products in Definition 5.4.1. The form b is symmetric in that $b(\mathbf{x},\mathbf{y}) = b(\mathbf{y},\mathbf{x})$, since $\mathbf{y}^T A\mathbf{x} = \mathbf{x}^T A\mathbf{y}$, as you see by taking the transpose. Bilinear means that it is linear in each set of variables separately. Thus, for a real numbers s,

$$b(s\mathbf{x},\mathbf{y}) = sb(\mathbf{x},\mathbf{y})$$

and

$$b(\mathbf{x}_1 + \mathbf{x}_2,\mathbf{y}) = b(\mathbf{x}_1,\mathbf{y}) + b(\mathbf{x}_2,\mathbf{y});$$

8.2.4 Exercise. Show that the definition implies

$$\begin{aligned} b(\mathbf{x}, s\mathbf{y}) &= sb(\mathbf{x},\mathbf{y}) \\ b(\mathbf{x},\mathbf{y}_1 + \mathbf{y}_2) &= b(\mathbf{x},\mathbf{y}_1) + b(\mathbf{x},\mathbf{y}_2). \end{aligned}$$

by using the symmetry of the form

To each symmetric bilinear form $b(\mathbf{x},\mathbf{y})$ one can associate the quadratic form $q(\mathbf{x}) = b(\mathbf{x},\mathbf{x})$ with the same matrix A. In the other direction, we can construct a symmetric bilinear form b for every quadratic form q by:

$$b(\mathbf{x},\mathbf{y}) = \frac{q(\mathbf{x}+\mathbf{y}) - q(\mathbf{x}-\mathbf{y})}{4}. \tag{8.2.5}$$

This expression is useful, because it shows that one can reconstruct all the entries of the matrix of the quadratic form from the values of q. Indeed,

$$a_{ij} = b(\mathbf{e}^i,\mathbf{e}^j) = \frac{q(\mathbf{e}^i + \mathbf{e}^j) - q(\mathbf{e}^i - \mathbf{e}^j)}{4}.$$

We will use bilinear forms in the proof of Theorem 8.3.2 and in the Law of Inertia.

8.2.6 Exercise. Let $\mathbf{a}_i$, $1 \le i \le n$ be n fixed vectors in an n-dimensional vector space V. For any collection of n real numbers x_i, form the linear function

$$l(\mathbf{x}) = \sum_{i=1}^{n} x_i \mathbf{a}_i.$$

which maps $\mathbb{R}^n$ to V. Show that the vectors $\mathbf{a}_i$ are linearly independent if and only if the null space of l is zero-dimensional. In that case, show that the vectors $\mathbf{a}_i$ form a basis of V.

8.2.7 Example (Hankel Forms)**.** We can produce an $n \times n$ matrix A from $2n-1$ numbers $s_0, \dots, s_{2n-2}$, by letting $a_{ij} = s_{i+j-2}$. Written out, this gives a symmetric matrix

$$A = \begin{bmatrix} s_0 & s_1 & s_2 & \dots & s_{n-1} \\ s_1 & s_2 & s_3 & \dots & s_n \\ \dots & \dots & \dots & \dots & \dots \\ s_{n-1} & s_n & s_{n+1} & \dots & s_{2n-2} \end{bmatrix} \tag{8.2.8}$$

called a *Hankel form.*

Notice that each square submatrix of a Hankel form is again a Hankel form. Hankel forms were investigated by the German mathematician Frobenius in the late nineteenth century: a good reference for his work is Gantmacher [25], V. 1, X.10. Frobenius showed how to compute the rank and the signature of a Hankel form from the signs of its leading principal minors D_k, a preoccupation similar to our preoccupation in this course.

Below we give the first step in Frobenius's approach.

The sequence $\{s_i\}$, $0 \le i \le 2n-2$, remains fixed throughout this discussion. Let us use the notation $[s_k]_{mp}$ for the $m \times p$ matrix M with ij-th entry $m_{ij} = s_{k+i+j-2}$. So the matrix A above is $[s_0]_{nn}$. Since our sequence only goes up to s_{2n-2}, the matrix $[s_k]_{mp}$ is only defined if $k+m+p \le 2n-2$. Every such matrix is a submatrix of A, as long as $m \le n$ and $p \le n$.

8.2.9 Lemma. *Let $\mathbf{a}_j$ be the j-th column of A. Assume that the first h columns of A are linearly independent, but the first $h+1$ columns are dependent. Then the leading principal minor D_h of A is non-zero.*

Note that the matrix formed by the first h columns of A, which in our notation is $[s_0]_{nh}$ has independent columns if and only if one of the $h \times h$ minors is different from 0. The lemma says that with the other hypothesis, the minor formed from the first h rows is always non-zero. This would be false for a more general matrix.

Proof. Since $\mathbf{a}_1, \dots, \mathbf{a}_h$ are linearly independent, the equation of linear dependence between $\mathbf{a}_1, \dots, \mathbf{a}_{h+1}$ can be written:

$$\mathbf{a}_{h+1} = \sum_{j=1}^{h} \alpha_j \mathbf{a}_j \tag{8.2.10}$$

Considering the set of equations (8.2.10) one row at a time, we see that we have:

$$s_{h+k} = \sum_{j=1}^{h} \alpha_j s_{j+k-1}\ , 0 \le k \le n-1. \tag{8.2.11}$$

Now consider the $h \times n$-matrix A' consisting of the first h rows of A. In our notation it is $[s_0]_{hn}$. Equation (8.2.11) says that the $(h+k)$-th column $\mathbf{a}'_{h+k}$ of A' can be written as a linear combination of the previous h columns:

$$\mathbf{a}'_{h+1+k} = \sum_{j=1}^{h} \alpha_j \mathbf{a}'_{j+k}\ , 0 \le k \le n-1.$$

Thus all the columns of A' can be written as a linear combination of its first h columns. Then the first h columns of A' must be linearly independent: if not, the column rank of A' would be less than h, but by hypothesis its row rank is h, so we have a contradiction. Finally notice that the first h columns of A' form a $h \times h$ matrix that is the upper left-hand corner of A, so its determinant is D_h, which is therefore non-zero. □

8.3 The Diagonalization of Quadratic Forms

In this section, we take the point of view of §7.8. We do not want to use $\mathbb{R}^n$ to denote our vector space, since that implies we are working with a given basis and coordinate system. Instead we will use V to denote our n-dimensional vector space. We first give it a basis $\mathbf{e}_1, \dots, \mathbf{e}_n$, and then another basis $\mathbf{f}_1, \dots, \mathbf{f}_n$. In §8.1 we assumed a basis was chosen, and we wrote out the quadratic form in that basis using the coefficients $\mathbf{x}$. Now we can also write $q(\mathbf{v})$, for a vector $\mathbf{v} \in V$. Since we allow arbitrary changes of basis, we will not use the inner product in our arguments.

Using the notation of §7.8, for any $\mathbf{v} \in V$, we write $[\mathbf{v}]_{\mathbf{e}} = \mathbf{x}$, and $[\mathbf{v}]_{\mathbf{f}} = \mathbf{z}$. Recall that this just expresses the fact that the coordinates of $\mathbf{v}$ in the $\mathbf{e}$-basis are $\mathbf{x}$, and the coordinates of $\mathbf{v}$ in the $\mathbf{f}$-basis are $\mathbf{z}$. Then, by the change of basis formula 7.8.7, $[\mathbf{v}]_{\mathbf{f}} = D[\mathbf{v}]_{\mathbf{e}}$ for the invertible change of basis matrix D. In other words, as per (7.8.6), $\mathbf{z} = D\mathbf{x}$, so $\mathbf{x} = D^{-1}\mathbf{z}$. For convenience write E for D^{-1}.

In the $\mathbf{e}$-basis of V, $q(\mathbf{v})$ has value $\mathbf{x}^T A\mathbf{x}$. We want to write a similar expression for $q(\mathbf{v})$ in the $\mathbf{f}$-basis. In the $\mathbf{f}$-basis of V, since $\mathbf{x} = E\mathbf{z}$, $q(\mathbf{v})$ has value

$$(E\mathbf{z})^T AE\mathbf{z} = \mathbf{z}^T E^T AE\mathbf{z}$$

so that the matrix of the quadratic form in the $\mathbf{f}$-basis is $B = E^T AE$.

We say that B *represents* the quadratic form in the $\mathbf{f}$-basis. Note that B is symmetric since A is. Indeed:

$$B^T = (E^T AE)^T = E^T A^T (E^T)^T = E^T AE = B.$$

We have established:

8.3.1 Proposition. *Given a symmetric $n \times n$ matrix A representing q, any other matrix representing q is of the form $E^T AE$, where E is an invertible $n \times n$ matrix.*

Our goal is to find a basis $\mathbf{f}$ on V that leads to the "simplest" possible matrix B representing the quadratic form, and study $\mathbf{z}^T B\mathbf{z}$ instead of $\mathbf{x}^T A\mathbf{x}$. Indeed, we can find a coordinate system such that B is a *diagonal* matrix, so we can write

$$q(\mathbf{z}) = b_{11}z_1^2 + b_{22}z_2^2 \cdots + b_{nn}z_n^2$$

8.3.2 Theorem. *Let A be an $n \times n$ real symmetric matrix. Then there exists an invertible $n \times n$ matrix D such that $B = D^T AD$ is a diagonal matrix with diagonal entries b_i, where each b_i is either 1, 0 or -1. If q is the quadratic form represented by A in the $\mathbf{x}$-coordinate system, then in the coordinate system $\mathbf{z}$ where $\mathbf{x} = D\mathbf{z}$, q is represented by B, so that*

$$q(z_1, z_2, \dots, z_n) = \sum_{i=1}^{n} b_i z_i^2 \tag{8.3.3}$$

Proof. We prove this by induction of the dimension n of the vector space V. We start the induction at $n = 1$. Pick any non-zero element a of V as its basis. Then if $q(a) = r$, for any $b \in \mathbb{R}$, $q(ba) = brb = rb^2$. If $q(a) = 0$, we are done. If $q(a) = r > 0$, then replace the basis a by $a' = a/\sqrt{r}$, to get the result. Indeed, $q(a') = 1$. If $q(a) = -r < 0$, then replace a with $a' = a/\sqrt{r}$ and get $q(a') = -1$.

Next assume that the result is proved for all vector spaces of dimension up to dimension $n-1$. We now prove it for dimension n. The key is to find an element $\mathbf{a}^n$ such that $q(\mathbf{a}_n) \neq 0$. If this cannot be done, (8.2.5) tells us that the bilinear form b associated to q is identically 0, so that the matrix A is 0, and any basis diagonalizes. So we may assume that we can find a $\mathbf{a}_n$ such that $q(\mathbf{a}_n) \neq 0$. Normalizing its length as in the one-dimensional case, we may assume that $q(\mathbf{a}_n) = \pm 1$. Put $\mathbf{y} = \mathbf{a}_n$ in $b(\mathbf{x}, \mathbf{y})$. Then the linear map $L(\mathbf{x}) = b(\mathbf{x}, \mathbf{a}_n)\colon \mathbb{R}^n \to \mathbb{R}$ is non-zero, so that its nullspace $\mathcal{N}$ is $(n-1)$-dimensional by the rank-nullity theorem: see §7.2. Restricting q to $\mathcal{N}$, by our induction hypothesis we can find a basis $\mathbf{a}_1$, $\mathbf{a}_2$, …, $\mathbf{a}_{n-1}$ of $\mathcal{N}$ satisfying the conclusion of the theorem. Furthermore $b(\mathbf{a}_i, \mathbf{a}_n) = 0$ for $1 \le i \le n-1$ by the construction of $\mathcal{N}$. Finally, we need to show that the $\mathbf{a}_i$, $1 \le i \le n$, form a basis for V. Since the first $n-1$ of them are independent, the only way adding $\mathbf{a}_n$ would fail to give a basis is if $\mathbf{a}_n$ is in $\mathcal{N}$. This would mean $L(\mathbf{a}^n) = b(\mathbf{a}^n, \mathbf{a}^n) = 0$, or $q(\mathbf{a}^n) = 0$, which was ruled out.

So we have a basis, and the matrix B representing q in this basis is diagonal with all diagonal elements equal to 0 or ±1, as required. □

Example 8.5.1 shows how this proof works.

We will also find a diagonal B in two other ways: by a Gaussian elimination argument called completing the square: §8.6 and then using Theorem 9.2.5.

8.4 Congruent Matrices

Before continuing with diagonization techniques, we show that there is a new equivalence relation here, this times on symmetric matrices, known as congruence. It is related to the equivalence relation of similar matrices, which applies to all square matrices, not just symmetric ones, studied in §7.7.1.

8.4.1 Definition. The $n \times n$ symmetric matrix A is *congruent* to the symmetric matrix B if there is an invertible matrix C such that $B = C^T AC$.

8.4.2 Example. Let A be the symmetric matrix

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

and C the invertible matrix

$$\begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$$

Then the matrix

$$B = C^T AC = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 14 \end{bmatrix}$$

is congruent to A.

8.4.3 Proposition. *Congruence is an equivalence relation on symmetric matrices.*

Proof. We will use the identity $(C^T)^{-1} = (C^{-1})^T$: the inverse of the transpose is the transpose of the inverse. This is established by taking the transpose of the equation $CC^{-1} = I$, and using the uniqueness of the inverse. The rest of the proof is nearly identical to that of Theorem 7.7.3, and is left to the reader. □

This equivalence relation partitions symmetric matrices into *congruence classes* of congruent matrices. How many different congruence classes are there? This question is answered in Corollary **??**.

Do not confuse congruence with similarity (Definition 7.7.1) , or orthogonal similarity (Definition 7.7.8). Once we have established the Spectral Theorem 9.2.1, we will see the connection between these three equivalence relations.

The following proposition will be useful in the Law of Inertia 8.5.5.

8.4.4 Proposition. *The rank of a symmetric matrix is an invariant of its congruence class: in other words, if two matrices are congruent, they have the same rank.*

Proof. Let $B = C^TAC$ be a matrix congruent to A, so C is invertible. Let $\mathcal{N}$ be the nullspace of A, and $\mathcal{M}$ be the nullspace of B. The subspace $C^{-1}\mathcal{N}$ is contained in $\mathcal{M}$, as one sees by carrying out the multiplication: let $\mathbf{p} \in \mathcal{N}$, so $BC^{-1}\mathbf{p} = C^TACC^{-1}\mathbf{p} = C^TA\mathbf{p} = 0$. By the same argument applied to an element in the nullspace of B, we see that $C\mathcal{M}$ is contained in $\mathcal{N}$. Because C is invertible, the dimension of $C^{-1}\mathcal{N}$ is the same as that $\mathcal{N}$, and the dimension of $C\mathcal{M}$ is the same as that of $\mathcal{M}$. Our two inclusions then say that $\dim \mathcal{N} \le \dim \mathcal{M} \le \dim \mathcal{N}$, so their dimensions are equal. Since A and B have nullspaces of the same dimension, they have the same rank. □

8.5 The Law of Inertia

Now we classify quadratic forms q by looking at the diagonal matrices representing it, namely the diagonal matrices B found in the Diagonalization Theorem 8.3.2. This is done by Sylvester's Law of Inertia 8.5.5. For a given quadratic form, we have not shown that the diagonal matrix B obtained in this way is unique, and in fact it is not. First we examine a simple example, and determine the source of non-uniqueness, using the proof strategy of Theorem 8.3.2.

8.5.1 Example. This is a continuation of Example 8.1.8. Let V be a two-dimensional vector space with basis $\mathbf{e}^1$, $\mathbf{e}^2$, and write an element $\mathbf{v}$ of V as $x_1\mathbf{e}^1 + x_2\mathbf{e}^2$. Assume that the quadratic form q is represented in the $\mathbf{e}$-basis as $q(x_1, x_2) = x_1x_2$, so its matrix is

$$A = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix}.$$

The bilinear form associated to q is

$$b(\mathbf{x}, \mathbf{y}) = \frac{(x_1 + y_1)(x_2 + y_2) - (x_1 + y_1)(x_2 + y_2)}{4} = \frac{x_1y_2 + y_1x_2}{2},$$

by (8.2.5). We construct a diagonalizing basis as in the proof of Theorem 8.3.2: we choose $\mathbf{f}^1 = a_1\mathbf{e}^1 + a_2\mathbf{e}^2$ with $q(\mathbf{f}^1) = a_1a_2 \neq 0$. So both a_1 and a_2 must be non-zero. We could normalize $\mathbf{f}^1$ so that $q(\mathbf{f}^1) = \pm 1$, by dividing by $\sqrt{a_1^2 + a_2^2}$, but we will not bother, to avoid burdening the computation. Then, following the proof of Theorem 8.3.2, we consider the linear form $b(\mathbf{x}, \mathbf{f}^1)$ and find an element $\mathbf{f}^2$ in its nullspace. This means solving for $\mathbf{x}$ in the equation $x_1a_2 + x_2a_1 = 0$. Up to multiplication by a non-zero scalar, we can take $\mathbf{x} = (a_1, -a_2)$, so that the second basis vector $\mathbf{f}^2 = a_1\mathbf{e}^1 - a_2\mathbf{e}^2$. If z_1 and z_2 are the coordinates in the $\mathbf{f}$-basis, the i-th column of the change of basis matrix E (7.8.6) satisfying $\mathbf{x} = E\mathbf{z}$

is the vector of coefficients of $\mathbf{f}^i$ in the $\mathbf{e}$-basis, so

$$E = \begin{bmatrix} a_1 & a_1 \\ a_2 & -a_2 \end{bmatrix}.$$

E is invertible because its determinant $-2a_1a_2 \neq 0$ by our choice of $\mathbf{f}^1$.

Then the matrix representing our quadratic form in the $\mathbf{f}$-basis is

$$B = E^T AE = \begin{bmatrix} a_1 & a_2 \\ a_1 & -a_2 \end{bmatrix} \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix} \begin{bmatrix} a_1 & a_1 \\ a_2 & -a_2 \end{bmatrix} = \begin{bmatrix} a_1a_2 & 0 \\ 0 & -a_1a_2 \end{bmatrix},$$

so, as predicted, it is diagonal, but with entries along the diagonal depending on a_1 and a_2. This shows there are infinitely many bases for V in which the quadratic form is diagonal. Even if one normalizes $\mathbf{f}^1$ and $\mathbf{f}^2$ to have length one, there is more than one choice. Our computation shows that in all of them, one of the diagonal entries is positive and the other is negative. The Law of Inertia 8.5.5 generalizes this computation.

Sylvester's Law of Inertia 8.5.5[1] shows that the following three numbers associated to a diagonal matrix D are congruence invariants of D, even if the diagonal entries d_i themselves are not.

8.5.2 Definition. Let B be an $n \times n$ diagonal matrix with diagonal entries b_1, b_2, $\dots$, b_n. Then

- p is the number of positive $b_i, 1 \leq i \leq n$.
- k is the number of zero $b_i, 1 \leq i \leq n$.
- m is the number of negative $b_i, 1 \leq i \leq n$.

The triple of integers (p, k, m) is called the *inertia* of B.

Note that $p + k + m = n$. The dimension of the nullspace of B is k, so $n - k$ is the rank of B. Proposition 8.4.4 says $n - k$ is an congruence invariant of B, so k is too.

8.5.3 Example. If D is the diagonal matrix D(7, -1, 0, 3, 3, -2), then $p = 3$, $k = 1$, and $m = 2$.

8.5.4 Definition. The *signature* of a diagonal matrix B is the number $p - m$. If $p + m = n$, B is called *non-degenerate* or *non-singular*; if $p + m < n$ is less than n, B is called *degenerate* or *singular*.

[1] Published by J. J. Sylvester in 1852 - [72].

8.5.5 Theorem (Sylvester's Law of Inertia)**.** *Let A be a symmetric $n \times n$ matrix. By Theorem 8.3.2 it is congruent to a diagonal matrix B, which has an inertia. The inertia is a congruence invariant of A: it is the same for any diagonal matrix congruent to A. Conversely any diagonal matrix with the same inertia as B is congruent to B.*

Proof. Assume we have two coordinate systems $\mathbf{e}$ and $\mathbf{f}$ in which the quadratic form q is diagonal. Let V_p, V_k and V_m be the subspaces of V spanned by the basis elements of $\mathbf{e}$ on which the quadratic from is positive, zero and negative, respectively, and let W_p, W_k and W_m be the analogous subspaces for the $\mathbf{f}$-basis. Let p_V, k_V, m_V be the dimensions of V_p, V_k and V_m, and p_W, k_W, m_W the dimensions of W_p, W_k and W_m. Clearly $p_V + k_V + m_V = p_W + k_W + m_W = n$. By Proposition 8.4.4, $k_V = k_W$. We will show that $p_V = p_W$, from which it will follow that $m_V = m_W$.

We claim that the linear subspaces V_p and $W_k + W_m$ of V do not intersect except at the origin. Suppose they did at a point $\mathbf{p} \neq \mathbf{0}$. Because $\mathbf{p} \in V_p$, we have $q(\mathbf{p}) > 0$, but because $\mathbf{p} \in W_k + W_m$, $q(\mathbf{p}) \leq 0$, a contradiction, so the claim is established.

This shows that $p_V \leq n - k_W - m_W = p_W$. Indeed, the $\mathbf{e}$-basis vectors spanning V_p, and the $\mathbf{f}$-basis vectors spanning $W_k + W_m$ can be extended, by the claim, to a basis for V. Indeed, suppose not: then we would have an equation of linear dependence, which would express an element of V_p as an element of $W_k + W_m$, and this is precisely what we ruled out.

Exchanging the role of the V's and W's, we get $p_W \leq p_V$, so they are equal. This concludes the proof that (p, k, m) are congruence class invariants.

The converse follows easily: using the notation above, construct linear maps between V_p and W_p, between V_k and W_k, and between V_m and W_m sending basis elements to basis elements. This is possible since there are the same number of basis elements in all three cases. This gives the desired change of basis. The theorem is proved. □

The Law of Inertia allows us to talk about the signature of q: it is the signature of any diagonal matrix representing q.

8.5.6 Corollary. *A quadratic form q in $\mathbb{R}^n$ is:*

Positive definite, if its signature is n, which forces the rank to be n;
Positive semidefinite, if its signature is m, $m \leq n$, and its rank m;
Negative definite, if its signature is $-n$, which forces the rank to be n;
Negative semidefinite, if its signature is $-m$, $m \leq n$, and its rank m;
Indefinite, if its signature is less than the rank, so both p and m are positive.

Proof. Call the signature s and the rank r. Then $s = p - m$, $r = p + m$. Referring back to Definition 8.1.7, the proof is immediate. □

Here is a second example showing what happens when the quadratic form does not have maximum rank.

8.5.7 Example. Using the same notation as in the previous example, assume that q can be written in the $\mathbf{e}$-basis as $q(x_1, x_2) = x_1^2$, so its matrix is

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

The bilinear form associated to q is $b(\mathbf{x}, \mathbf{y}) = x_1 y_1$, as per (8.2.5). Pick any vector $\mathbf{f}^1 = a_1\mathbf{e}^1 + a_2\mathbf{e}^2 \in V$, so that $q(\mathbf{f}^1) \neq 0$. This just says that $a_1 \neq 0$. In this case we divide by a_1, and write $\mathbf{f}^1 = \mathbf{e}^1 + a\mathbf{e}^2$. Then, following the proof of Theorem 8.3.2, we consider the linear form $b(\mathbf{x}, \mathbf{a}) = x_1$ and find a non-zero element $\mathbf{f}^2$ in its nullspace. We must take $\mathbf{f}^2 = c\mathbf{e}^2$, for $c \neq 0$ Let

$$D = \begin{bmatrix} 1 & a \\ 0 & c \end{bmatrix}$$

be the change of basis matrix from the $\mathbf{e}$-basis to the $\mathbf{f}$-basis. D is invertible because its determinant $c \neq 0$ by choice of $\mathbf{f}^1$ and $\mathbf{f}^2$. Then we have

$$\begin{bmatrix} \mathbf{f}^1 \\ \mathbf{f}^2 \end{bmatrix} = D \begin{bmatrix} \mathbf{e}^1 \\ \mathbf{e}^2 \end{bmatrix} \text{ and } \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = D^T \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

Then the matrix of our quadratic form in the $\mathbf{f}$-basis is

$$B = DAD^T = \begin{bmatrix} 1 & a \\ 0 & c \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ a & c \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

so, as predicted, it is diagonal. In this example, because we normalized the length of the first new basis vector $\mathbf{f}^1$, then entries of the new diagonal matrix are the same as the ones we started with.

The form in Example 8.5.3 has signature 1. It is degenerate and indefinite.

8.5.8 Example. The matrix of the quadratic form

$$q(x_1, x_2, x_3) = x_1^2 + x_1x_2 + x_1x_3 + x_2^2 + x_2x_3 + x_3^2 \tag{8.5.9}$$

is

$$A = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix} \tag{8.5.10}$$

since $\mathbf{x}^T A\mathbf{x} = q(\mathbf{x})$. In order to diagonalize q, we compute the eigenvalues and eigenvectors of A. This computation can be done easily, because the matrix A has an unusual amount of symmetry.

We get an eigenvalue λ when the rank of the characteristic matrix

$$\lambda I_3 - A = \begin{bmatrix} \lambda - 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \lambda - 1 & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \lambda - 1 \end{bmatrix} \tag{8.5.11}$$

is less than 3. When $\lambda = 1/2$, all three columns of $\lambda I_3 - A$ are the same, showing that this matrix has rank 1, so $1/2$ is an eigenvalue. Furthermore any vector x with $x_1 + x_2 + x_3 = 0$ is an eigenvector. The eigenvectors associated to this eigenvalue form a vector space V_2 of dimension 2. We say that 1/2 is an eigenvalue of multiplicity 2, or that two eigenvectors have eigenvalue 1/2. Two possible independent eigenvectors are $(1, -1, 0)$ and $(0, 1, -1)$, as you should check.

When $\lambda = 2$ the columns of $\lambda I_3 - A$ add to the zero vector, and thus are linearly dependent, so 2 is an eigenvalue. The associated eigenvector is (up to multiplication by a scalar) the vector $(1, 1, 1)$. It is orthogonal to V_2.

So we have found a basis for V consisting of the three eigenvectors. In this basis, the matrix for q is diagonal with the eigenvalues down the diagonal. This is the key point, that we will develop in Theorem 9.2.5. Since all three eigenvalues are positive, the matrix is positive definite. Thus the signature of the form is 3.

The eigenvalue/eigenvector computation is unusually simple in this example. Usually one has to factor the characteristic polynomial, which can be painful. For $n \geq 5$ one generally needs to use an iterative technique such as Newton's method to diagonalize under eigenvectors. So we use a different approach on the next example.

8.5.12 Example. We compute the signature of the $n \times n$ symmetric matrix M_n with all diagonal terms equal to $n - 1$ and all off diagonal terms equal to -1:

$$M_n = \begin{bmatrix} n-1 & -1 & \dots & -1 \\ -1 & n-1 & \dots & -1 \\ \dots & \dots & \dots & \dots \\ -1 & -1 & \dots & n-1 \end{bmatrix}$$

We will show that the signature and the rank are $n - 1$, so that the form is positive semidefinite. We do this by first computing the signature for $n = 2$ and then setting up a proof by induction. Letting $n = 2$, we get

$$M_2 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

By using symmetric Gaussian elimination (see §8.6) we can transform this to the diagonal matrix $(1, 0)$, so $p = 1$, $k = 1$ and $m = 0$. We are done. Next

$$M_3 = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}$$

By symmetric Gaussian elimination again, this transforms our matrix into the congruent matrix: We get

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{3}{2} & -\frac{3}{2} \\ 0 & -\frac{3}{2} & \frac{3}{2} \end{bmatrix}$$

and the 2×2 matrix in the bottom right is just M_2 multiplied by $\frac{3}{2}$. The 1 in upper left-hand corner just adds 1 to the signature we found in the case $n = 2$, so the signature is $(2, 0)$. This suggests the general strategy: we prove by induction that the signature of M_n is $n - 1$ and the rank $n - 1$. By row reduction, first dividing the top row by $n - 1$, and then clearing the first column, you get

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & \frac{n(n-2)}{n-1} & \dots & -\frac{n}{n-1} \\ \dots & \dots & \dots & \dots \\ 0 & -\frac{n}{n-1} & \dots & \frac{n(n-2)}{n-1} \end{bmatrix}$$

The bottom right matrix of size $(n - 1) \times (n - 1)$ is $\frac{n}{n-1}$ times the matrix M_{n-1}. By induction we know that the signature and the rank of M_{n-1} are both $n - 2$ and we are done. Note that we are using Sylvester's law of inertia 8.5.5 to say that this matrix is congruent to M_n. Compare to Example 9.4.9.

We will develop more tests for positive definiteness, negative definiteness and the like in §9.4 , but first we must prove the most important theorem connected to real symmetric matrices: the Spectral Theorem. We do this in the next lecture.

8.6 Completing the square

The technique described below is known as Lagrange's reduction method[2], published in 1759 in [37], long before matrices were invented. We illustrate it with examples after the proof of the theorem. Some are taken from Lagrange's original paper.

[2] See Steen [62], p.360, and Gantmacher [25] §X.3.

We start with the quadratic form, written as in (8.1.2):

$$q(\mathbf{x}) = \sum_{i=1}^{n} \Bigl(\sum_{j=1}^{n} a_{ij} x_i x_j \Bigr),$$

with $a_{ij} = a_{ji}$ both appearing in the sum, $i \neq j$.

8.6.1 Theorem (Lagrange's Reduction Method for Quadratic Forms)**.** *Given the quadratic form $q(\mathbf{x})$ written as above, it can be diagonalized inductively by one of the two following changes of basis.*

1. *If the diagonal term $a_{kk} \neq 0$, then make the change of variables:*

$$y_k = \sum_{j=1}^{n} a_{kj} x_j, \text{ with } y_j = x_j \text{ for } j \neq k. \tag{8.6.2}$$

   In the new system of coordinates the variable y_k only appears as a square in the quadratic form.

2. *If the two diagonal elements a_{gg} and a_{hh} are both zero, but a_{gh} is not zero, then make the change of variables*

$$\begin{aligned} y_g &= \sum_{j=1}^{n} (a_{gj} + a_{hj}) x_j \\ y_h &= \sum_{j=1}^{n} (a_{gj} - a_{hj}) x_j \\ y_j &= x_j \text{ otherwise.} \end{aligned}$$

   In the new system of coordinates the variables y_g and y_h only appear as squares in the quadratic form.

Proof. We divide the proof in two parts corresponding to the two cases of the theorem.

8.6.1 A diagonal term is non-zero

First assume one of the diagonal terms, say a_{kk}, is non-zero. This case is easy. Define

$$s_k(\mathbf{x}) = \frac{1}{a_{kk}} \Bigl(\sum_{j=1}^{n} a_{kj} x_j \Bigr)^2 \tag{8.6.3}$$

Notice that the coefficients of all the terms $x_k x_j$, $1 \le j \le n$, in $s(\mathbf{x})$ are the same as the corresponding ones in $q(\mathbf{x})$. Indeed, the coefficient of x_k^2 is a_{kk}, while that of $x_k x_j$, $j \neq k$, is $2a_{kj} = a_{kj} + a_{jk}$. Thus the difference $q(\mathbf{x}) - s_k(\mathbf{x})$, which we call $q_1(\mathbf{x})$ does not contain the variable x_k. If we let y_k be as in the statement we can replace x_k by y_k to get a new basis, and by construction y_k only appears as a square in the quadratic form. Thus the variable x_k does not appear in q_1 (nor does y_k, of course), and

$$q = q_1 + \frac{1}{a_{kk}} y_k^2.$$

This allows us to proceed by induction on n, as claimed in the theorem

8.6.2 Two diagonal terms vanish

Assume that the two diagonal terms a_{gg} and a_{hh} vanish, and that a_{gh} does not vanish. In fact, since this method is harder than the previous one, it is better to use it only when necessary, namely when there are no non-zero diagonal terms left. Unless the quadratic form is identically zero, we can find an off-diagonal term, say $a_{gh} \neq 0$. In this case the Lagrange method says to complete two squares at the same time:

$$s_g(\mathbf{x}) = \frac{1}{2a_{gh}} \Big(\sum_{j=1}^{n} (a_{gj} + a_{hj}) x_j \Big)^2 \tag{8.6.4}$$

$$s_h(\mathbf{x}) = \frac{1}{2a_{gh}} \Big(\sum_{j=1}^{n} (a_{gj} - a_{hj}) x_j \Big)^2. \tag{8.6.5}$$

We replace the two variables x_g and x_h by variables y_g and y_h as indicated in the statement of the theorem. We can write the change of basis matrix as a partitioned matrix (see §6.10), listing the g and h coordinates first in both coordinate systems. Thus we get

$$\begin{bmatrix} C & X \\ 0 & I \end{bmatrix}$$

where I is the $n-2$ identity matrix, 0 the zero matrix, X a matrix we do not need to compute, and C the 2×2 matrix

$$\begin{bmatrix} 0 & a_{gh} \\ a_{gh} & 0 \end{bmatrix}$$

Thus the determinant of our big matrix is $-a_{gh}^2$, and thus is non-zero by assumption. So it can be used as a change of basis matrix.

Next we consider the quadratic form:

$$q_1(\mathbf{x}) = q(\mathbf{x}) - \big(s_g(\mathbf{x}) - s_h(\mathbf{x})\big) = q(\mathbf{x}) - \frac{1}{2a_{gh}}\big(y_g^2 - y_h^2\big).$$

Claim: q_1 does not contain the variables x_g or x_h. If so, we have diagonalized two variables in q. We prove this in several steps. Notice the minus sign between s_g and s_h in this expression.

First we consider the x_g^2 term, so we examine the term $j = g$ in the sums (8.6.4) and (8.6.5). Its coefficient in q is $a_{gg} = 0$, which vanishes by hypothesis. In s_g we get

$$\frac{1}{2a_{gh}}(a_{gg} + a_{hg})^2 = a_{gh}/2$$

because $a_{gg} = 0$. In s_h we get

$$\frac{1}{2a_{gh}}(a_{gh} - a_{gg})^2 = a_{gh}/2$$

so in $s_g - s_h$ get 0, as required. The term x_h^2 in treated the same way.

Next the term $x_g x_h$. Its coefficient in s_g is, setting $j = h$:

$$\frac{1}{2a_{gh}} 2(a_{gg} + a_{hg})(a_{gh} + a_{hh}) = a_{hg}$$

and in s_h is $-a_{hg}$, so the total contribution is $2a_{hg}$, as required.

Finally we take a term $x_g x_j$, where j is neither g nor h. The coefficient in s_g is

$$\frac{1}{2a_{gh}} 2a_{hg}(a_{gj} + a_{hj}) = a_{gj} + a_{hj}$$

while that in s_h is

$$-\frac{1}{2a_{gh}} 2a_{hg}(a_{gj} - a_{hj}) = -a_{gj} + a_{hj}$$

so the difference is $2a_{gj}$, so we get agreement again. The terms $x_h x_j$, follow from symmetry so we are done. □

Each step in the Lagrange reduction changes the basis on the vector space in order to simplify the quadratic form. By Proposition 8.3.1, at each step the matrix A of the quadratic form gets replaced by a new matrix of the form $E^T AT$. In case 1, E is a triangular matrix of the type consider in §**??** and in ordinary Gaussian elimination. In case 2, the matrix E derives from the change of coordinates written down in the statement of the theorem,and is easily seen not to be triangular. It is therefore less familiar.

And now for some examples. A simple example will show why this method is called completing the square. Take a form in two variables, say $q(\mathbf{x}) = x_1^2 + 2x_1x_2 + 3x_2^2$. Note that by our convention, this means that $a_{12} = 1$. Then we consider all the terms involving x_1, namely $x_1^2 + 2x_1x_2$, and ask: can we complete this expression so that it is a square? Yes, it is the beginning of the square of x_1+x_2. So $s_1(\mathbf{x}) = (x_1 + x_2)^2$ and so $q(\mathbf{x}) - s_1(\mathbf{x}) = 2x_2^2$, so we have eliminated all the terms involving x_1 as claimed.

8.6.6 Example. In his 1759 paper on finding maxima and minima, Lagrange writes the quadratic form in 2 variables as

$$At^2 + 2Btu + Cu^2$$

using capital letters for the constants and u, v for the variables, and notes that if $A \neq 0$ it can be written as the sum of the squares

$$A\Big(t + \frac{Bu}{A}\Big)^2 + \Big(C - \frac{B^2}{A}\Big)u^2.$$

8.6.7 Example. Then Lagrange writes the quadratic form in 3 variables as

$$At^2 + 2Btu + Cu^2 + 2Dtv + 2Euv + Fv^2,$$

using capital letters for the constants and t, u, v for the variables. He reduces first to

$$A\Big(t + \frac{Bu}{A} + \frac{Dv}{A}\Big)^2 + \Big(C - \frac{B^2}{A}\Big)u^2 + 2\Big(E - \frac{BD}{A}\Big)uv + \Big(F - \frac{D^2}{A}\Big)v^2$$

then, setting

$$a = C - \frac{B^2}{A},\ b = E - \frac{BD}{A},\ c = F - \frac{D^2}{A},$$

his form becomes

$$A\Big(t + \frac{Bu}{A} + \frac{Dv}{A}\Big)^2 + au^2 + 2buv + cv^2$$

So we should replace the variable t by a new variable $w = t + \frac{Bu}{A} + \frac{Dv}{A}$, and this is exactly what out general algorithm above tells us to do. Then we just use the two-variable case, assuming that $a \neq 0$.

Then he says (§8) "One can extend the same theory to functions of four or more variables. Whoever has understood the spirit of the reduction I have used up to now, will be able to discover without difficulty those needed in any special case."

Here now is a case with actual numbers:

8.6.8 Example. We use Example 8.1.4, so $q(x_1, x_2, x_3) = x_1^2 + 2x_1x_2 - x_1x_3 - x_2^2 + x_2x_3 + 4x_3^2$. We look for a variable with a pure square term: x_1 will do. The terms involving x_1 are $x_1^2 + 2x_1x_2 - x_1x_3$. The degree-one polynomial whose square starts out like this is $z_1 = x_1 + x_2 - x_3/2$. Then

$$\begin{aligned} q(x_1, x_2, x_3) &= z_1^2 - x_2^2 + x_2x_3 - x_3^2/4 - x_2^2 + x_2x_3 + 4x_3^2 \\ &= z_1^2 - 2x_2^2 + 2x_2x_3 + \frac{15}{4}x_3^2 \end{aligned}$$

We can now repeat the process on

$$q_1(x_2, x_3) = -2x_2^2 + 2x_2x_3 + \frac{15}{4}x_3^2,$$

using x_2 as our next variable for completing the square. Since $a_{22} = -2$, and $a_{23} = 1$, by (8.6.3) we have $s_2 = (-1/2)(-2x_2 + x_3)^2 = -2x_2^2 + 2x_2x_3 - x_3^2/2$, so when we subtract s_2 from q_1, all the terms involving x_2 vanish. Let $z_2 = -2x_2 + x_3$, so $s_2 = -z_2^2/2$. We get $q_1 + z_2^2/2 = \frac{15}{4}x_3^2 + x_3^2/2 = \frac{17}{4}x_3^2$. Finally let $z_3 = x_3$. Substituting everything in, we have q written in our new coordinates z_i as $z_1^2 - z_2^2/2 + \frac{17}{4}z_3^2$. So the quadratic form is diagonalized, the signature is 1 and the rank 3. You can check the result using Lagrange's formulas from Example 8.6.7.

8.6.9 Example. Let's write a general three-by-three case: So $q(x_1, x_2, x_3) = 2(a_{12}x_1x_2 + a_{13}x_1x_3 + a_{23}x_2x_3$. We assume $a_{12} \neq 0$, so we will eliminate the first two variables simultaneously.

Write linear forms $y_1 = a_{12}x_2 + (a_{13} + a_{23})x_3$ and $y_2 = -a_{12}x_1 + (a_{13} - a_{23})x_3$ to make the change of variable. The change of basis matrix is

$$\begin{bmatrix} 0 & a_{12} & a_{13} + a_{23} \\ -a_{12} & 0 & a_{13} - a_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

which is indeed invertible,

It is now an easy exercise to subtract

$$\frac{1}{2a_{gh}}(y_g^2 - y_h^2)$$

from q, to verify that all the terms in x_1 and x_2 have indeed vanished.

Now a special case:

8.6.10 Example. Let $q = 2x_1x_2 + 4x_1x_3 + 8x_2x_3$, so $a_{12} = 1$, $a_{13} = 2$, and $a_{23} = 4$. We use the first two variables, so $g = 1$ and $h = 2$.

$$s(\mathbf{x}) = \frac{1}{2}(x_1 + x_2 + 6x_3)^2$$
$$t(\mathbf{x}) = \frac{1}{2}(-x_1 + x_2 - 2x_3)^2$$

and then $q - (s - t) = -16x_3^2$, so all the terms involving x_1 and x_2 have been eliminated. If we set $y_1 = x_1 + x_2 + 6x_3$, and $y_2 = -x_1 + x_2 - 2x_3$, $y_3 = 4x_3$, then the form becomes $(y_1^2 - y_2^2)/2 - y_3^2$.

To recapitulate how one uses the method in practice:

8.6.11 Definition. We have two moves:

1. Move I: When a diagonal term is non-zero (say $a_{kk}x_k^2$, with $a_{kk} \neq 0$), complete the square on x_k so as to eliminate all the off-diagonal terms involving x_k. This is illustrated by Example 8.6.8.
2. Move II: When all the remaining diagonal coefficients of A are zero, but the off-diagonal coefficients $a_{gh} = a_{hg}$ are non-zero, use (8.6.4) and (8.6.5) to eliminate the variables x_g and x_h from the quadratic form, and replace them by two new variables that only occur as squares.

When we diagonalize a quadratic form in this way, we do not have to worry about the order of the variables: so for example we can complete the square on x_2 before doing anything to x_1. When we write down a matrix for q, it is usual to do the operations starting with the first row and column, so that we want to deal with x_1 before x_2, etc. This is not a problem, since any order of the variables can be achieved by a permutation. On the matrix level, this is realized by conjugation by the appropriate permutation matrix.

8.6.12 Remark. Move II is only used if the quadratic form q is indefinite. Indeed, in the diagonalization it produces diagonal terms with opposite signs.

8.6.13 Example. We revisit Example 8.5.8, that we solved using eigenvalues and eigenvectors earlier. The matrix of the quadratic form is

$$q(x_1, x_2, x_3) = x_1^2 + x_1x_2 + x_1x_3 + x_2^2 + x_2x_3 + x_3^2 \tag{8.6.14}$$

We want to change variables to new variables z_1, z_2, z_3 by completing the square to remove the cross terms. Starting with x_1, we see we must take

$$z_1 = x_1 + \frac{1}{2}x_2 + \frac{1}{2}x_3.$$

transforming the form to

$$z_1^2 + \frac{3}{4}x_2^2 + \frac{1}{2}x_2x_3 + \frac{3}{4}x_3^2 \tag{8.6.15}$$

So we have eliminated x_1 and replaced it by the new variable z_1. We now repeat the process, eliminating x_2 by completing the square on the terms

$$\frac{3}{4}x_2^2 + \frac{1}{2}x_2x_3 = \frac{3}{4}(x_2^2 + \frac{2}{3}x_2x_3) \tag{8.6.16}$$

We set $z_2 = x_2 + \frac{1}{3}x_3$, square this and substitute $\frac{3}{4}(x_2^2 + \frac{2}{3}x_2x_3)$ out of (8.6.15) as before to get

$$z_1^2 + \frac{3}{4}z_2^2 + \frac{2}{3}x_3^2 \tag{8.6.17}$$

We set $z_3 = x_3$, completing the change of variables:

$$\begin{array}{llll} z_1 = & x_1 & +\frac{1}{2}x_2 & +\frac{1}{2}x_3 \\ z_2 = & & x_2 & +\frac{1}{3}x_3 \\ z_3 = & & & x_3 \end{array}$$

We call the triangular, and invertible, matrix of coefficients D^{-1}, so that we have $\mathbf{z} = D^{-1}\mathbf{x}$. We use D^{-1} and not D to keep the same notation as in (7.8.6): it is D^{-1}, since here we are writing the new basis in terms of the old one. We need to compute D. By back-substitution, solving for the x_i in terms of z_j, we see it is the upper triangular matrix

$$D = \begin{bmatrix} 1 & -1/2 & -1/3 \\ 0 & 1 & -1/3 \\ 0 & 0 & 1 \end{bmatrix}$$

8.6.18 Exercise. Compute $B = D^T A D$:

$$B = \begin{bmatrix} 1 & -0 & 0 \\ -1/2 & 1 & 0 \\ -1/3 & -1/3 & 1 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 1 & -1/2 & -1/3 \\ 0 & 1 & -1/3 \\ 0 & 0 & 1 \end{bmatrix}$$

and show you get the diagonal matrix $D(1, 3/4, 2/3)$, confirming the computation above.

It is worth double checking this (and any other) computation. The determinant of A is $1/2$ by an easy computation. The determinant of B is the product of the diagonal elements, thus also $1/2$. This is as it should be. The determinant of the product $D^T AD$ is the product of the individual determinants, and D and D^T have determinant 1, since they are triangular with 1's along the diagonal.

Lecture 9

The Spectral Theorem

The main result of the lecture, one of the most important results in linear algebra, indeed, all of mathematics, the Spectral Theorem 9.2.1, tells us that the eigenvalues and eigenvectors of a real symmetric matrix are real. We prove it without introducing complex number, using the Rayleigh Quotient instead. An immediate corollary of the Spectral Theorem is Theorem 9.2.5, which shows that we can diagonalize real symmetric matrices using orthogonal matrices, studied in §6.7 precisely for this appearance. This gives another proof of Theorem 8.3.2.

The cost of not using complex numbers is that we have to use the Weierstrass Theorem 16.2.2 that says that a continuous function on a compact set has both a minimum and a maximum.

Then we list various ways of characterizing positive definite (Theorem 9.4.1) and semidefinite (Theorem 9.5.1) forms. You need to learn to recognize when a Hessian is positive semidefinite or not.

9.1 The Rayleigh Quotient

We fix a basis on the vector space$V$, as well as the standard inner product for this basis. Given the quadratic form $\mathbf{x}^T A\mathbf{x}$, we make the following definition which will be used in the proof of the Spectral Theorem 9.2.1 and in Theorem 13.1.3.

9.1.1 Definition. The *Rayleigh quotient* is the real-valued function, defined for all non-zero vectors $\mathbf{x}$, defined as:

$$R(\mathbf{x}) = \frac{\langle \mathbf{x}, A\mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle}.$$

We have already looked at a Rayleigh-like functions in Example 11.1.6 (the second example). Also see Example 16.3.8. Throughout we use Theorem 7.4.5 for symmetric matrices: $\langle \mathbf{x}, A\mathbf{x}\rangle = \mathbf{x}^T A\mathbf{x} = \langle \mathbf{A}x, \mathbf{x}\rangle$.

9.1.2 Theorem. *The Rayleigh quotient is a continuous function on $\mathbb{R}^n \setminus \mathbf{0}$. Let m denote the* inf*, and M the* sup *of its values on $\mathbb{R}^n\setminus\mathbf{0}$. Then both are finite, and $R(\mathbf{x})$ attains both m and M at points $\mathbf{e}_m$ and $\mathbf{e}_M$ on the unit sphere $U = \{\mathbf{x} | \|\mathbf{x}\| = 1\}$. Furthermore*

- *m is the smallest eigenvalue of A and $\mathbf{e}_m$ is any eigenvector of A corresponding to m.*
- *M is the largest eigenvalue of A and $\mathbf{e}_M$ is any eigenvector of A corresponding to M.*

Note that this shows that A has real eigenvalues m and M.

Proof. R is clearly continuous everywhere it is defined, namely everywhere except at the origin.

9.1.3 Lemma. *For any non-zero $t \in \mathbb{R}$, $R(t\mathbf{x}) = R(\mathbf{x})$, so the Rayleigh quotient is homogeneous of degree 0.*

Proof. Both numerator and denominator of the Rayleigh quotient are homogeneous functions of degree 2 (see Definition 12.3.1), so t^2 can be factored out of both, removing all the t from R, showing that it is homogenous of degree 0. □

9.1.4 Definition. A *ray* emanating from a point $\mathbf{c}$ through a point $\mathbf{e}$ different $\mathbf{c}$ is the set of $t(\mathbf{e} - \mathbf{c})$, for $t \in \mathbb{R}_+$. Thus a ray is a half-line ending at $\mathbf{c}$ and passing through $\mathbf{e}$.

The lemma says that the Rayleigh quotient is constant along rays emanating from the origin. Since each ray intersects the unit sphere in a point, all values of R are attained on the unit sphere U.

Since U is closed and bounded, and $R(\mathbf{x})$ is continuous on U, we can apply the maximum theorem: $R(\mathbf{x})$ attains both its minimum and its maximum values on U. As we will see in the easy Theorem 13.1.1, any maximizer or minimizer for R is a critical point for R, namely a point $\mathbf{e}$ where the gradient of R vanishes. Thus the importance of the following proposition:

9.1.5 Proposition. *Let $\mathbf{e} \in U$ be a point where the gradient of R vanishes: $\nabla R(\mathbf{e}) = \mathbf{0}$. Then $\mathbf{e}$ is an eigenvector of A with eigenvalue $a = R(\mathbf{e})$.*

Proof. Let $\mathbf{f}$ be an arbitrary but fixed non-zero vector in $\mathbb{R}^n$, and let t be a real variable. We evaluate the Rayleigh quotient at $\mathbf{e} + t\mathbf{f}$, and write the composite function as

$$g(t) = R(\mathbf{e} + t\mathbf{f}).$$

The numerator of $g(t)$ is

$$p(t) = \langle \mathbf{e} + t\mathbf{f}, A(\mathbf{e} + t\mathbf{f})\rangle = \langle \mathbf{e}, A\mathbf{e}\rangle + 2t\langle \mathbf{e}, A\mathbf{f}\rangle + t^2\langle \mathbf{f}, A\mathbf{f}\rangle \tag{9.1.6}$$

and its denominator is

$$r(t) = \langle \mathbf{e} + t\mathbf{f}, \mathbf{e} + t\mathbf{f}\rangle = \langle \mathbf{e}, \mathbf{e}\rangle + 2t\langle \mathbf{e}, \mathbf{f}\rangle + t^2\langle \mathbf{f}, \mathbf{f}\rangle \tag{9.1.7}$$

Now $g'(t) = \langle \nabla R((\mathbf{e} + t\mathbf{f}), \mathbf{f}\rangle$ by the chain rule. We evaluate $g'(t)$ at $t = 0$. Since $\nabla R(\mathbf{e}) = 0$ by hypothesis, we get $g'(0) = 0$.

On the other hand, since $g(t) = p(t)/r(t)$, by the quotient rule we get

$$g'(0) = \frac{p'(0)r(0) - p(0)r'(0)}{r^2(0)} = 0.$$

Now $r^2(0) = 1$, since $\mathbf{e}$ is on U, and $p(0) = R(\mathbf{e})$, which we denote a. So we get:

$$g'(0) = p'(0) - ar'(0) = 0. \tag{9.1.8}$$

Next we compute the derivatives of $p(t)$ and $r(t)$ at 0, using (9.1.6) and (9.1.7) respectively.

$$p'(0) = 2\langle \mathbf{f}, A\mathbf{e}\rangle$$
$$r'(0) = 2\langle \mathbf{f}, \mathbf{e}\rangle$$

Equation 9.1.8 reads, after substituting in these values:

$$2\langle \mathbf{f}, A\mathbf{e}\rangle - 2a\langle \mathbf{f}, \mathbf{e}\rangle = 0 \quad \text{, or} \quad \langle \mathbf{f}, A\mathbf{e} - a\mathbf{e}\rangle = 0.$$

Since $\mathbf{f}$ is an arbitrary vector in $\mathbb{R}^n$, this means that $A\mathbf{e} - a\mathbf{e}$ is perpendicular to every vector, which can only happen if it is the zero vector: $A\mathbf{e} - a\mathbf{e} = \mathbf{0}$. Thus $\mathbf{e}$ is an eigenvector of A with eigenvalue $a = R(\mathbf{e})$, which concludes the proof of the proposition. □

It is now easy to prove Theorem 9.1.2. The maximum and minimum values of R are attained on U at critical points of R, so that the gradient of R vanishes there. Thus the corresponding vector $\mathbf{e}$ is an eigenvector with eigenvalue a. Thus the minimum is attained at the eigenvector(s) corresponding to the smallest eigenvalue, and the maximum at the eigenvector(s) corresponding to the largest eigenvalue. This concludes the proof of the theorem. □

9.2 The Spectral Theorem for a Real Symmetric Matrix

In §8.3, starting with our symmetric matrix A, we looked for diagonalizations $D = C^T AC$ where D is diagonal and C is a change of basis matrix. In this section we will show that this diagonalization can be achieved with an orthogonal matrix C, namely a matrix such that $C^{-1} = C^T$. This means that D and A are similar, and therefore have the same characteristic polynomial. See §6.7.

9.2.1 Theorem (The Spectral Theorem)**.** *If A is a real symmetric $n \times n$ matrix, then its eigenvalues are real and its eigenvectors can be selected to form an orthogonal basis of the vector space V.*

The spectrum of a matrix is the set of its eigenvalues. This theorem is called the spectral theorem because it describes the eigenvalues of a real symmetric matrix: they are real. The first paragraph of Steen [62] discusses the early history of the spectral theorem, at the time it was called the principal axis theorem. We have already seen the contribution of Sylvester in his law of inertia 8.5.5. We have used the method of Lagrange (§8.6) and that of Jacobi (Theorem 6.9.6) to diagonalize quadratic forms. In §13.7 we will discuss principal axes when considering the level sets of objective functions.

9.2.2 Example. Before starting the proof, let's work out the familiar 2 × 2 case. Let A be an arbitrary 2 × 2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

To compute the eigenvalues of A, we need the roots of the characteristic polynomial of A, namely the determinant

$$\begin{vmatrix} t-a & -b \\ -c & t-d \end{vmatrix} = t^2 - (a+d)t + ad - bc.$$

The quadratic formula tells us that this polynomial has real roots if and only if the discriminant is non-negative. The discriminant is

$$(a+d)^2 - 4(ad-bc) = a^2 + 2ad + d^2 - 4ad + 4bc = (a-d)^2 + 4bc.$$

When the matrix is symmetric, $b = c$, so we get $(a-d)^2 + 4b^2$, a sum of squares, which is always non-negative. So the eigenvalues:

$$\lambda_i = \frac{a + d \pm \sqrt{(a-d)^2 + 4b^2}}{2}$$

are real.

What about the eigenvectors? We could compute them, but we only need to show they are orthogonal. First assume the matrix has a double eigenvalue. This corresponds to the discriminant being 0, which means that $b = 0$ and $a = d$. Because the matrix is diagonal, any non-zero vector in the plane is an eigenvector. There is therefore no difficulty in finding two eigenvectors that are orthogonal. Now assume that the eigenvalues are distinct. Then Proposition 7.4.7 shows they are orthogonal, which settles the theorem in dimension 2.

We now do the case of general n.

Proof of the Spectral Theorem. Let $R(\mathbf{x})$ be the Rayleigh quotient of the symmetric matrix A.

Then Proposition 9.1.2 gives us an eigenvector $\mathbf{e}_m$ of A of length 1, where the Rayleigh quotient achieves its minimum value on the unit sphere, with corresponding **real** eigenvalue a. We now rename these:

$$\mathbf{e}_1 = \mathbf{e}_m \text{ and } a_1 = a$$

We want to apply Proposition 9.1.2 to A restricted to the subspace V_1 of V, the orthogonal complement of $\mathbf{e}_1$ in V. V_1 is a vector space of dimension $n - 1$. We need the following

9.2.3 Lemma. *A maps V_1 to V_1. In other words, if $\mathbf{x} \in V_1$, then $A\mathbf{x}$ is in V_1, and so is perpendicular to $\mathbf{e}_1$.*

Proof. We have

$$\begin{aligned} \langle A\mathbf{x}, \mathbf{e}_1\rangle &= \langle \mathbf{x}, A\mathbf{e}_1\rangle && \text{by self-adjointness 7.4.5,} \\ &= a_1\langle \mathbf{x}, \mathbf{e}_1\rangle && \text{because } A\mathbf{e}_1 = a_1\mathbf{e}_1, \\ &= 0 && \text{because } \mathbf{x} \in V_1. \end{aligned}$$

□

The Rayleigh quotient, when restricted to the unit sphere U_1 in V_1, has a minimum value a_2. Since $U_1 \subset U$, we have $a_2 \geq a_1$. Proposition 9.1.2 shows that a_2 is a real eigenvalue of A with unit length eigenvector $\mathbf{e}_2$. Repeat the process. Continuing in this way, we find an orthonormal basis of eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ with real eigenvalues $a_1 \leq a_2 \leq \dots \leq a_n$. □

We could also have started with an eigenvector corresponding to the largest eigenvalue, which we also know is real from our Rayleigh quotient result.

9.2.4 Definition. Let A be a symmetric $n \times n$ matrix. Let $\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_n$ be the collection of orthonormal eigenvectors found in the Spectral Theorem, and λ_i the corresponding eigenvalues. Let Q be the matrix whose i-th column is the eigenvector $\mathbf{e}_i$. Then Q is called the *matrix of eigenvectors* of A, and $\lambda = (\lambda_1, \dots, \lambda_n)$ the *vector of eigenvalues*.

We write D for $D(\lambda_1, \lambda_2, \dots, \lambda_n)$, the diagonal matrix with diagonal entries the eigenvalues.

9.2.5 Theorem. *Let A be a real symmetric $n \times n$ matrix, Q its matrix of eigenvectors, and λ its vector of eigenvalues. Then Q is an orthogonal matrix and*

$$Q^{-1}AQ = D \quad \textit{or} \quad A = QDQ^T \tag{9.2.6}$$

Proof. That the matrix Q is orthogonal follows immediately from the fact that its columns, the eigenvectors, are orthonormal. We can write all the eigenvector-eigenvalue equations in one matrix equation:

$$AQ = QD, \tag{9.2.7}$$

as a moment's thought will confirm. Multiply on the left by Q^{-1}, to get $Q^{-1}AQ = Q^{-1}QD = D$. □

9.2.8 Exercise. Show that (9.2.7) encodes all the eigenvector-eigenvalues, as claimed.

Review Definition 8.5.2 for the meaning of p, k, and m in the next result.

9.2.9 Corollary. *Start with a symmetric matrix A. Its rank is the number of nonzero eigenvalues. p is the number of positive eigenvalues, k is the number of zero eigenvalues, and m is the number of negative eigenvalues.*

Proof. The matrix $D = D(\lambda_1, \lambda_2, \dots, \lambda_n)$ is congruent to A because Q is an orthogonal matrix, so $Q^{-1} = Q^T$. Now p, k, and m are invariants of the congruence class. They are easy to compute for the matrix D. □

9.2.10 Corollary. *Assume further that A is positive definite, thus invertible. The eigenvalues of A^{-1} are $1/\lambda_i$ with the same eigenvectors $\mathbf{e}_i$, and therefore the same eigenvector matrix Q, so A^{-1} is also positive definite. Then the eigenvalue-eigenvector decomposition of A^{-1} can be written:*

$$Q^{-1}A^{-1}Q = D(1/\lambda_1, 1/\lambda_2, \dots, 1/\lambda_n)$$

Proof. All the matrices in (9.2.6) are invertible, so just compute the inverse using the fact that the inverse of the orthogonal matrix Q is its transpose, that the inverse of the diagonal matrix $D(\lambda_1, \lambda_2, \dots, \lambda_n)$ is $D(1/\lambda_1, 1/\lambda_2, \dots, 1/\lambda_n)$ and that computing the inverse of a product of invertible matrices reverses the factors of the product. □

So Q is a change of basis matrix that diagonalizes the quadratic form, as in Theorem 8.3.2. It is a "better" change of basis because it preserves distance and angle - that is what being orthogonal means. Note finally, that the diagonal matrix obtained by this method is uniquely defined, since it consists in the eigenvalues of A.

Why not always diagonalize by this method? The answer is that it is harder (and more expensive computationally) to compute the eigenvalues than to do Gaussian elimination.

9.2.11 Example (Example 8.5.8 once again)**.** In §8.6 we will compute a diagonal matrix (associated to the quadratic form given by (8.5.10) by change of basis), and obtain $D(1, 3/4, 2/3)$. In (8.5.8) we computed the eigenvalues of the same form q, and obtained $D(1/2, 1/2, 2)$. From the preceding remark see that $D(1/2, 1/2, 2)$ can also be viewed as being obtained from a change of basis. Thus, as we claimed in the remark before Definition 8.5.2, the matrix D itself is not unique. However, in accordance with the Law of Inertia 8.5.5, the numbers p_+, p_0 and p_- are the same: indeed, for both, we get $(3, 0, 0)$. The form q is positive definite.

9.2.12 Example. By Proposition 7.7.12, A and A^σ have the same type: if one is positive definite, the other is; if one is positive semidefinite, the other is, and so on.

Indeed, they have the same characteristic polynomial and therefore the same eigenvalues. Therefore by Corollary 9.2.9 they have the same signature.

9.3 The Symmetric Square Root of a Symmetric Matrix

9.3.1 Definition. Let A be a positive semidefinite symmetric $n \times n$ matrix. Write its orthogonal matrix of eigenvectors Q and its matrix Λ of eigenvalues, so that by Theorem 9.2.5, $A = Q\Lambda Q^T$. Because A is positive semidefinite, its eigenvalues λ_i are nonnegative by the Spectral Theorem 9.2.1. So set $\sigma_i = \sqrt{\lambda_i}$, let Σ the diagonal matrix with the σ_i on the diagonal, so $\Sigma^2 = \Lambda$. Finally let $R = Q\Sigma Q^T$. The symmetric $n \times n$ matrix R is called the *square root* of A, since $A = R^2$. R is written $\sqrt{A}$.

Conversely, if a symmetric matrix A has a square root R, meaning a symmetric matrix R such that $A = R^2$, then for any $\mathbf{x}$,

$$\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T R^2\mathbf{x} = \langle R^T\mathbf{x}, R\mathbf{x}\rangle = \|R\mathbf{x}\| \geq 0,$$

so A is positive semidefinite. We have established:

9.3.2 Proposition. *Every positive semidefinite matrix A has a square root. Conversely if a symmetric matrix has a square root, it is positive semidefinite. Furthermore, A is positive definite if and only if its square root is.*

The last statement is clear, since the eigenvalues of $\sqrt{A}$ are the square roots of those of A.

9.3.3 Proposition. *Let $M = UR$, where R is the symmetric square root of the positive semidefinite matrix A, and U is any orthogonal matrix. Then*

$$M^T M = A$$

.

Proof. Since U is orthogonal, Proposition 6.7.2) tells us that $U^T = U^{-1}$. Then

$$M^T M = R^T U^T U R = R^T R = A.$$

□

So M could also be called a square root of R, but it is not generally symmetric. If we replace the diagonal matrix Σ in Definition 9.3.1 by a diagonal matrix S where we replace any number of diagonal elements by their negative, we get a new symmetric matrices S such that $A = S^2$. We will not consider these.

9.4 Positive Definite Quadratic Forms

We defined positive definite quadratic forms, and positive semidefinite quadratic forms in Definition 8.1.7, and noted their importance in minimization theory in Proposition 8.1.9. We now study them in more detail.

Here we collect different ways of characterizing positive definite quadratic forms, in others words necessary and sufficient conditions for a form to be positive definite. We also collect a number of necessary conditions.

9.4.1 Theorem. *Let $q(\mathbf{x})$ be a quadratic form in n variables, with matrix A. The following conditions are equivalent.*

1. *q is positive definite. So $q(\mathbf{x}) > 0$ for all $\mathbf{x} \neq 0$. In matrix terms, $\mathbf{x}^T A\mathbf{x} > 0$.*

2. *The eigenvalues of A are positive.*

3. *The signature of A is n.*

4. *The leading principal minors of the matrix A of q are positive.*

5. *There is a $n \times n$ symmetric invertible matrix R such that $A = R^T R$.*

Note that the number of positive pivots is the number p in the inertia. So (3) is equivalent to having all the pivots positive.

Proof. Much of the proof follows from earlier work. First, (2) $\Leftrightarrow$ (3) is a special case of Corollary 9.2.9.

Then (3) $\Leftrightarrow$ (4) follows from Theorem 6.9.6. Indeed, since $D_0 = 1$ in that theorem, if all the pivots d_k are positive, then all the leading principal minors D_k are, and conversely, by the formula $d_k = D_k/D_{k-1}$, $1 \leq k \leq n$.

Next we show the equivalence of (1) and (2). First (1) $\Rightarrow$ (2): for every eigenvector $\mathbf{e}$ with eigenvalue a, we have

$$q(\mathbf{e}) = \langle \mathbf{e}, A\mathbf{e}\rangle = \langle \mathbf{e}, a\mathbf{e}\rangle = a\langle \mathbf{e}, \mathbf{e}\rangle$$

Since q is positive definite $q(\mathbf{e}) > 0$, and $\langle \mathbf{e}, \mathbf{e}\rangle$ is the norm of $\mathbf{e}$, and therefore positive. Thus a is positive.

Now (2) $\Rightarrow$ (1). By the Spectral Theorem, we can write any vector $\mathbf{z}$ in $\mathbb{R}^n$ as a linear combination $\sum_{i=1}^n z_i\mathbf{e}_i$ of the eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$. As before, let a_i denote the eigenvalue associated to $\mathbf{e}_i$. Then

$$A\mathbf{z} = \sum_{i=1}^n z_i A\mathbf{e}_i = \sum_{i=1}^n z_i a_i \mathbf{e}_i.$$

We need to show that $q(\mathbf{z})$ is positive for all $\mathbf{z} \neq 0$.

$$q(\mathbf{z}) = \langle \mathbf{z}, A\mathbf{z}\rangle = \sum_{i=1}^n a_i z_i^2$$

because the eigenvectors are orthogonal of length 1. This is positive since all the a_i are positive, so we are done.

Finally we show the equivalence of (1) and (5). To show (5) $\Rightarrow$ (1), just note that if $\mathbf{x} \neq \mathbf{0}$, then $R\mathbf{x} \neq \mathbf{0}$, since R is invertible. So

$$\langle R\mathbf{x}, R\mathbf{x}\rangle = \langle \mathbf{x}, R^T R\mathbf{x}\rangle = \langle \mathbf{x}, A\mathbf{x}\rangle > 0.$$

To show $(1) \Rightarrow (5)$ just use the symmetric square root $\sqrt{A}$ for R. It is invertible when A is positive definite by Proposition 9.3.2.

□

9.4.2 Corollary. *If A is positive definite, all the diagonal elements a_{ii} are positive. All principal minors are positive too.*

9.4.3 Exercise. Prove the corollary using Proposition 7.7.12: if A is positive definite, then so is A^σ for any permutation σ. Exercise 7.7.13 then shows that any principal minor of A can be transformed into a leading principal minor of A^σ for a suitable σ.

This corollary is a useful tool for disproving that a quadratic form is positive definite. All you need to do is find a diagonal entry or a principal minor that is not positive, and you know the form is not positive definite.

9.4.4 Example. The matrix associated to the quadratic form considered in Example 8.6.9 has zeroes along the diagonal, so it is not positive definite. Indeed, we noted that it is indefinite.

9.4.5 Proposition. *Write the characteristic polynomial $P(t) = \det(tI - A)$ of A as (see* (7.3.2)*)*

$$p(t) = t^n - p_1 t^{n-1} + p_2 t^{n-2} + \cdots + (-1)^n p_n \tag{9.4.6}$$

Then p_i is the sum of the $\binom{n}{i}$ principal minors of degree i. In particular p_1 is the trace, which is the sum of the n principal minors a_{ii} of degree 1, and p_n is the determinant of A, namely the unique principal minor of degree n.

Work out what happens when $n = 3$. The general case is proved using permutations and symmetric functions. We will not provide a proof, but one can be found in Horn and Johnson [32], p. 42. This result is used in Theorem 9.5.1.

9.4.7 Corollary. *A is positive definite if and only if $p_i > 0$ for all i, $1 \le i \le n$.*

Proof. First assume all the p_i are positive. In particular $p_n \neq 0$, so that 0 is not a root of $p(t)$. A eigenvalue of A is by definition a root of $p(t)$: the alternation of the signs of the coefficients means that for negative t all the terms in the sum (9.4.6) have the same sign, so that characteristic polynomial cannot vanish. Therefore all the eigenvalues are positive and we are done.

Now assume that A is positive definite. Then all the principal minors of A are positive (see Corollary 9.4.2), so their sum is also, by Proposition 9.4.5. □

This is very convenient for proving or disproving that a symmetric matrix gives a positive definite form: just compute the characteristic polynomial and check the signs of the p_i.

9.4.8 Example. Compute the characteristic polynomial of the matrix (8.5.10). You will get

$$t^3 - 3t^2 + \frac{9}{8}t - \frac{1}{2}$$

so the p_i as defined by (9.4.6) are all positive, and the matrix is positive definite. This confirms something we already knew for this example.

9.4.9 Example. Let Q_n is the $n \times n$ matrix

$$\begin{bmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & -1 & 2 & -1 \\ 0 & \dots & 0 & 0 & -1 & 2 \end{bmatrix}$$

Let's show Q_n is positive definite. Note that $\det Q_1 = 2$, $\det Q_2 = 3$, and $\det Q_n = 2\det Q_{n-1} - \det Q_{n-2}$, so $\det Q_n = n+1$. This allows us to compute all the leading principal subminors: they are all positive, so Q_n is positive definite. Amusingly one can find a formula for the inverse of Q_n. First consider the $n \times n$ matrix R_n, where

$$R_2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, R_3 = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix}, R_4 = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 6 & 4 & 2 \\ 2 & 4 & 6 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix},$$

so that R_n is the symmetric matrix with entries

$$r_{ij} = (n+1-j)i\,, i \le j. \tag{9.4.10}$$

Note that if E_n denotes the $n \times n$ matrix with all entries equal to 1, then the upper left hand $n \times n$ block of $R_{n+1} = R_n + E_n$. Then by induction and using the appropriate block decomposition (see §6.10) one can show that: $Q_n R_n = (n+1)I_n$, so the inverse of Q_n is $\frac{1}{n+1}R_n$. We use this matrix in an optimization problem in Example 13.2.3. Compare this example to Example 8.5.12, which is only positive semidefinite.

9.4.11 Example. We work on a $(n+1)$-dimensional vector space with coordinates $x_0, x_1, \dots, x_n$. We let the quadratic form be $\mathbf{x}^T A\mathbf{x}$, where the entries of A are $a_{ij} = \frac{1}{i+j-1}$, $0 \le i, j \le n$. So

$$A = \begin{bmatrix} 1 & 1/2 & 1/3 & \dots & 1/(n+1) \\ 1/2 & 1/3 & 1/4 & \dots & 1/(n+2) \\ \dots & \dots & \dots & \dots & \dots \\ 1/(n+1) & 1/(n+2) & 1/(n+3) & \dots & 1/(2n+1) \end{bmatrix}$$

Thus A is a Hankel form, as defined in Example 8.2.7. Our goal is to show that this quadratic form is positive definite. Instead of using one of the criteria we have developed, recall from integral calculus that

$$\int_0^1 y^k dy = \frac{1}{k+1}\ , k \ge 0.$$

The quadratic form can be written

$$\begin{aligned} \mathbf{x}^T A\mathbf{x} &= \sum_{i=0}^{n}\sum_{j=0}^{n}\Big(\int_0^1 y^{i+j}dy\Big)x_i x_j \\ &= \int_0^1 \Big(\sum_{i=0}^{n}\sum_{j=0}^{n} y^{i+j}x_i x_j\Big)dy = \int_0^1 \Big(\sum_{i=0}^{n} y^i x_i\Big)^2 dy \end{aligned}$$

Notice what we did: we interchanged the integral with the two finite summations, and then the key step was to recognize the integrand on the second line as a square.

Then, to show that this form is positive definite, we need to show that for any choice of constants $x_0, x_1, \dots, x_n$, not all zero, this integral is non-zero. It is clearly non-negative, so the only issue is showing that it is not zero. This is the integral of the square of a non-zero polynomial, thus a continuous function. In a neighborhood of a point where the polynomial does not vanish, we get a positive contribution to the integral, so we are done.

One can produce many more examples of this type by using the algebraic moments

$$s_k = \int_0^1 y^k f(y)dy\ , k \ge 0.$$

of an integrable function $f(y)$. See [32] p. 393.

In the same vein, consider a polynomial $f(x) = \sum_{i=0}^m a_i x^i$ of degree m in one variable x that never takes on negative values. Then m must be even: just consider the behavior at ±∞. So write $m = 2n + 2$. It is possible to show that $f(x)$ is the sum of the squares of two polynomials h and g of degree at most $n+1$:

$f(x) = g(x)^2 + h(x)^2$: see [51], Chapter 7, for a proof. From the coefficients of f form the $(n+1) \times (n+1)$ Hankel matrix where $s_i = a_i$. Then one can prove:

9.4.12 Proposition. *A Hankel form is positive semidefinite if and only if the sequence s_i are the coefficients of a polynomial that never takes on negative values.*

See [10], Exercise 2.37, for the sketch of a proof.

9.5 Positive Semidefinite Quadratic Forms

Here is the theorem, parallel to Theorem 9.4.1, characterizing positive semidefinite quadratic forms. The parallelism fails in one respect: one needs to check that all the principal minors, not just the leading principal minors, are nonnegative.

9.5.1 Theorem. *Let $q(\mathbf{x})$ be a quadratic form with matrix A. The following conditions are equivalent.*

1. *q is positive semidefinite. So $q(\mathbf{x}) \geq 0$ for all $\mathbf{x}$. In matrix terms, $\mathbf{x}^T A\mathbf{x} \geq 0$.*
2. *The eigenvalues of A are nonnegative.*
3. *The signature of A is its rank, so the pivots are all nonnegative.*
4. *The principal minors of the matrix A are all nonnegative.*
5. *There is a $n \times n$ symmetric matrix R such that $A = R^T R$.*

Proof. The proofs of (1) ⇔ (2), (2) ⇔ (3), (1) ⇔ (5) are similar to those in Theorem 9.4.1.

It remains to show (1) ⇔ (4). First $(1) \Rightarrow (4)$: replace the form with matrix A by one with matrix $A_\epsilon = A + \epsilon I$, where ϵ is any positive number and I is the $n \times n$ identity matrix. Since A is positive semidefinite, for any $\epsilon > 0$, A_ϵ is positive definite. Thus by Corollary 9.4.2 all its principal minors are positive. Take the limit as $\epsilon \to 0$, we see that the principal minors of A must be nonnegative.

Finally, $(4) \Rightarrow (1)$. So we assume that all the principal minors of A are nonnegative. Replace A with A_ϵ as before. Notice that every principal minor of A_ϵ is of the form $\det(M + \epsilon I_m)$, where M is a principal submatrix of size m of A, and I_m is the identity matrix of size m. Thus, other than a missing negative sign, this is the characteristic polynomial of the matrix M, using the variable ϵ. By Proposition 9.4.5, this shows that this minor can be written as a polynomial in ϵ of degree m: $\epsilon^m + \dots$ where all the lower degree terms have nonnegative coefficients. So it is positive, and now Corollary 9.4.7 implies that A_ϵ is positive definite: $\mathbf{x}^T A_\epsilon \mathbf{x} > 0$ when $\mathbf{x} \neq 0$. So let ϵ go to zero to get $\mathbf{x}^T A\mathbf{x} \geq 0$ as required. □

9.5.2 Example. Consider the matrix

$$A = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}.$$

Its leading principal minors are both zero, and yet it is clearly not positive semidefinite. The problem is that the 1×1 principal minor $a_{22} = -1$ is negative, so the hypotheses of the theorem are not satisfied. Indeed, this form is negative semidefinite and the associated function has a weak maximum at the origin.

More generally, a theorem of Jacobi ([25], theorem 2 in §X.3) shows how to compute the signature of a quadratic form from the signs of the leading principal minors.

Part IV

Multivariable Calculus and Unconstrained Optimization

Lecture 10

Sequences

We review sequences infinite sequences in 10.1, then prove the standard results about convergence of sequences in §10.2. All this material is covered in multivariable calculus, and is provided for the reader's convenience.

10.1 Sequences

An *infinite sequence* in $\mathbb{R}^n$ is an infinite list of points in $\mathbb{R}^n$:

$$\mathbf{a}_1, \quad \mathbf{a}_2, \quad \mathbf{a}_3, \dots, \mathbf{a}_n, \dots,$$

usually, as here, indexed by the *natural numbers* $\mathbb{N}$.

Here is the more formal definition[1] with values in $\mathbb{R}^n$:

10.1.1 Definition. A *sequence* is a function from the natural numbers $\mathbb{N}$ to $\mathbb{R}^n$.

As is customary, instead of using the usual functional notation $a(i)$ for sequences, we write the variable i of the sequence function as a subscript, so we write $\mathbf{a}_i$, and when $\mathbf{a}_i$ is a vector, we write it in boldface. It is not important that the first index in the sequence be 1: it could be -1, 0, 100, or, in fact, any integer whatsoever. It is convenience and the context that suggests the start index.

The value of the sequence at any given natural number i is called the i-th *term* of the sequence. We typically write sequences within curly brackets: $\{a_i\}$, or by listing a few terms in the sequence, say $a_1, a_2, a_3, a_4, \dots$ enough so that the pattern of subsequent terms is clear. As already mentioned, if the sequence takes values in $\mathbb{R}^n$, $n > 1$, we write $\{\mathbf{a}_i\}$.

[1] Stewart [63] has a discussion of sequences in §11.1.

10.1.2 Definition. A sequence $\{\mathbf{a}_i\}$ *converges* if there is a point $\mathbf{a} \in \mathbb{R}^n$ such that: For every $\epsilon > 0$ there is a sufficient large integer N such that $i \geq N$ implies that $d(\mathbf{a}_i, \mathbf{a}) < \epsilon$. When the sequence converges the point $\mathbf{a}$ is called the *limit* of the sequence, and is written $\lim_{i \to \infty} \mathbf{a}_i$. When a sequence does not converge, it is said to *diverge*.

10.1.3 Example. $\{1/i\}$ means the sequence $1, 1/2, 1/3, \dots$. It is easy to see that this sequence converges to 0.

10.1.4 Example. The sequence $\{\cos \pi i\}$ can be evaluated to give the sequence $-1, 1, -1, 1, -1, \dots$. Note, as in this example, that the terms of a sequence do not have to be distinct. The sequence $\{\cos \pi i\}$ does not converge: instead it bounces back and forth between -1 and 1.

10.1.5 Example. The sequence $\{i^2\}$, or $1, 4, 9, 16, \dots$ does not converge, because the terms get arbitrarily large. We say it diverges to infinity.

Here are three simple but important theorems about convergent sequences.

10.1.6 Theorem. *The limit of a convergent sequence is unique.*

10.1.7 Exercise. Prove Theorem 10.1.6. Hint: do this by contradiction: assume the sequence converges to two different values. See the proof of Theorem 3.1.3 for a model.

10.1.8 Theorem. *Suppose a_i and b_i are two convergent sequences in $\mathbb{R}$, converging to a and b respectively. Then*

1. $\lim_{i\to\infty}(a_i + b_i) = \lim_{i\to\infty} a_i + \lim_{i\to\infty} b_i = a + b$ *;*
2. $\lim_{i\to\infty} ca_i = c \lim_{i\to\infty} a_i = ca$*, for all* $c \in \mathbb{R}$*;*
3. $\lim_{i\to\infty}(a_i b_i) = ab$*;*
4. $\lim_{i\to\infty} 1/a_i = 1/a$ *as long as all the terms of the sequence* $\{a_i\}$ *and its limit* a *are non-zero.*

Proof. We sketch a proof of the first statement. We must show that for any $\epsilon > 0$, there is an integer N such that for all $i \geq N$,

$$|a_i + b_i - a - b| < \epsilon \tag{10.1.9}$$

Since $\{a_i\}$ converges to a, we can find a N_1 such that for $i \geq N_1$,

$$|a_i - a| < \epsilon/2. \tag{10.1.10}$$

Similarly we can find a N_2 such that for $i \geq N_2$,

$$|b_i - b| < \epsilon/2. \tag{10.1.11}$$

By the triangle inequality,

$$|a_i + b_i - a - b| < |a_i - a| + |b_i - b| \tag{10.1.12}$$

so that adding (10.1.10) and (10.1.11) and applying (10.1.12) gives the desired result 10.1.9. □

If $\{\mathbf{a}_i\}$ is a sequence in $\mathbb{R}^n$, we write the coordinates of the point $\mathbf{a}_i$ as $a_{i,j}$, $1 \leq j \leq n$.

10.1.13 Theorem. *A sequence $\{\mathbf{a}_i\}$ in $\mathbb{R}^n$ converges to the point $\mathbf{a} = (a_1, a_2, \dots a_n)$ if and only if the sequence of j-th coordinates, $\{a_{i,j}\}$, $1 \leq j \leq n$, converges to a_j for all j, $1 \leq j \leq n$.*

Proof. We only prove one of the two implications, leaving the other one to the reader. We want to show that the sequence $\{\mathbf{a}_i\}$ converges to $\mathbf{a}$, assuming that the sequences of coordinates all converge to the appropriate coordinate of $\mathbf{a}$. Thus for any positive ϵ we need to find an appropriate N. Write $\delta = \epsilon/\sqrt{n}$. For this δ we can find a N_j such that when $i \geq N_j$, $|a_{i,j} - a_j| < \delta$.

Let N be the maximum of the N_j. Note that when $i \geq N$, we have

$$\|\mathbf{a}_i - \mathbf{a}\| = \sqrt{\sum_{j=1}^{n} (a_{i,j} - a_j)^2} \leq \delta\sqrt{n} \tag{10.1.14}$$

So we get the convergence condition for the sequence $\{\mathbf{a}_i\}$ using N.

The other implication is proved in a similar way. □

Thus to check the convergence of a sequence of vectors, it is enough to check the convergence of the coordinates.

10.1.15 Definition. A sequence $\{\mathbf{a}_i\}$ in $\mathbb{R}^n$ is *bounded* if there exists a real number D such that $\|\mathbf{a}_i\| \leq D$ for all i. Thus $\{\mathbf{a}_i\}$ is *unbounded* if for any real number D there exists an index i such that $\|\mathbf{a}_i\| > D$.

Note that the second statement in this definition is just the contrapositive of the first. The sequence in Example 10.1.5 is not bounded, while the sequences in Examples 10.1.3 and 10.1.4 are.

10.2 Convergence Tests for Sequences

10.2.1 Theorem. *If all the terms of a convergent sequence $\{\mathbf{a}_i\}$ lie in a closed set S, then the limit $\mathbf{a}$ is also in S.*

10.2.2 Exercise. Prove Theorem 10.2.1. First treat the easy case where the sequence takes on only a finite number of values. Otherwise, establish that the limit of the sequence is a limit point of S, and use the definition of a closed set.

10.2.3 Definition. Let $\{s_i\}$ be a sequence in $\mathbb{R}$. Then $\{s_i\}$ is *increasing* if $s_{i+1} \geq s_i$ for all i, and $\{s_i\}$ is *decreasing* if $s_{i+1} \leq s_i$ for all i. We simply say that $\{s_i\}$ is *monotonic* if it is either increasing or decreasing.

10.2.4 Theorem. *If the sequence $\{s_i\}$ is monotonic and bounded, then it converges. to the least upper bound of the sequence in the case of an increasing sequence, and to the greatest lower bound for a decreasing sequence.*

Proof. Assume the sequence $\{s_i\}$ is increasing and bounded above by m. Then by Theorem 14.2.5 the set of terms in the sequence has a least upper bound we call s. We need to show that $\{s_i\}$ converges to s. So for every $\epsilon > 0$, we need to find a N such that when $i \geq N$ $s - s_i < \epsilon$. Because s is the least upper bound, for every ϵ there is a term that is within ϵ of s. So there is an s_N satisfying $s - s_N < \epsilon$. Because the sequence is increasing, for any $i > N$ we have $s - s_i \leq s - s_N < \epsilon$ which is exactly what we needed. □

We also have the *squeezing principle* for sequences, which is only stated here in $\mathbb{R}$, but which could be easily adapted to $\mathbb{R}^n$.

10.2.5 Theorem. *Consider three sequences of real numbers $\{a_i\}$, $\{b_i\}$, and $\{c_i\}$, with $a_i \leq b_i \leq c_i$ for all $i \geq i_0$. If $\{a_i\}$ and $\{c_i\}$ converge to the same limit L, then $\{b_i\}$ also converges, and to the same value L.*

Proof. Because $\{a_i\}$ converges to L, for any $\epsilon > 0$ there is an N_1 such that if $i \geq N_1$, $|a_i - L| < \epsilon$. Because $\{c_i\}$ converges to L, for the same ϵ, there is an N_2 such that if $i \geq N_2$, $|c_i - L| < \epsilon$. Let N be greater than i_0, N_1 and N_2. Because $a_i \leq b_i \leq c_i$, we have $|b_i - L| < \epsilon$, which shows that $\{b_i\}$ converges to L. □

Lecture 11

Continuous Functions in Several Variables

This chapter reviews the notion of continuity of a real-valued function at a point, phrasing it using the language of distance functions.[1]

11.1 Definitions

As in the one-variable case (see §3.1) we first need to define the limit of a function f at a point $\mathbf{a}$. Compare this to Definition 3.1.1, which is now the special case $n = 1$. Note in particular that f need not be defined at the point $\mathbf{a}$.

11.1.1 Definition. Let D be a set in $\mathbb{R}^n$, let $\mathbf{a}$ be a limit point of D , and assume that the real-valued function f is defined on D. Then we say that $f(\mathbf{x})$ approaches ℓ as $\mathbf{x}$ approaches $\mathbf{a}$ if for all $\epsilon > 0$ there exists a $\delta > 0$ such that when $d(\mathbf{x}, \mathbf{a}) < \delta$, $x \in D$ and $\mathbf{x} \neq \mathbf{a}$, then $|f(\mathbf{x}) - \ell| < \epsilon$. We write $\lim_{\mathbf{x} \to \mathbf{a}} f(\mathbf{x}) = \ell$, and call ℓ the limit of the function $f(\mathbf{x})$ at $\mathbf{p}$. If there is no value ℓ that works, we say the limit does not exist.

11.1.2 Definition. A function f defined on a set $D \subset \mathbb{R}^n$ to $\mathbb{R}$ is *continuous* at a point $\mathbf{a} \in D$ if $\lim_{\mathbf{x} \to \mathbf{a}} f(\mathbf{x}) = f(\mathbf{a})$.

In other words, for all $\epsilon > 0$ there is a $\delta > 0$ such that all $\mathbf{x} \in D$ such that $d(\mathbf{x}, \mathbf{a}) < \delta$ get mapped by f into the interval of radius ϵ around $f(\mathbf{a})$.

Here is the prototypical example:

[1] Stewart [63] gives a good discussion of both the one-variable case (§2.5) and the multi-variable case (§14.2).

11.1.3 Example. Fix a point $\mathbf{p}$ in $\mathbb{R}^n$, and let $f(\mathbf{x}) = \|\mathbf{x} - \mathbf{p}\|$. Then $f(\mathbf{x})$ is continuous.

Proof. Note that $f(\mathbf{x}) = d(\mathbf{x}, \mathbf{p})$. We fix a $\mathbf{x}$ and show that f is continuous at $\mathbf{x}$. This means that for all $\epsilon > 0$ we need to find a $\delta > 0$ such that for all $\mathbf{y}$ such that $\|\mathbf{x} - \mathbf{y}\| < \delta$, then $|f(\mathbf{x}) - f(\mathbf{y})| < \epsilon$. We can take $\delta = \epsilon$. Indeed, by the triangle inequality,

$$\|\mathbf{x} - \mathbf{p}\| \leq \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{p}\| \quad \text{or} \quad f(\mathbf{x}) \leq \|\mathbf{x} - \mathbf{y}\| + f(\mathbf{y})$$

so just move the right-most term to the left side, and then repeat, interchanging the role of $\mathbf{x}$ and $\mathbf{y}$. □

11.1.4 Exercise. Now consider the distance function $d(\mathbf{x}, \mathbf{y})$ in $\mathbf{R}^n$ as a function of the $2n$ coordinates of $\mathbf{x}$ and $\mathbf{y}$. Show that d is a continuous function at $(\mathbf{x}^*, \mathbf{y}^*)$ by taking a point $\mathbf{x}$ close to $\mathbf{x}^*$ and a point $\mathbf{y}$ close to $\mathbf{y}^*$, and writing

$$\|\mathbf{x} - \mathbf{y}\| \leq \|\mathbf{x} - \mathbf{x}^*\| + \|\mathbf{x}^* - \mathbf{y}^*\| + \|\mathbf{y}^* - \mathbf{y}\|$$

by a generalization of the triangle inequality. Write down all the details of an $\epsilon - \delta$ proof, and illustrate with a picture in $\mathbb{R}^2$. In Example 22.3.4 we show that this function is convex.

11.1.5 Exercise. In the same vein, now consider a non-empty set S in $\mathbf{R}^n$, and define the distance of $\mathbf{x}$ to S to be

$$D_S(\mathbf{x}) = \inf_{\mathbf{s} \in S} d(\mathbf{x}, \mathbf{s}).$$

Show that $D_S(\mathbf{x})$ is a continuous function. For which points $\mathbf{x}$ does one have $D_S(\mathbf{x}) = 0$?

Hint: To say that $D_S(\mathbf{x}) = d$ means that there is an infinite sequence of points $\mathbf{s}_n \in S$ such that $\|\mathbf{x} - \mathbf{s}_n\| \leq d + 1/n$. This is just a rephrasing of the definition of the least upper bound. You need to show that for any $\epsilon > 0$ there is a $\delta > 0$ such that for all $\mathbf{y}$ with $\|\mathbf{x} - \mathbf{y}\| < \delta$, then $|D_S(\mathbf{x}) - D_S(\mathbf{y})| < \epsilon$. Use the triangle inequality $\|\mathbf{y} - \mathbf{s}_n\| \leq \|\mathbf{y} - \mathbf{x}\| + \|\mathbf{x} - \mathbf{s}_n\|$.

To understand what it means for a real-valued function to have a limit in several variables, we look at some examples in $\mathbb{R}^2$.

11.1.6 Example. Consider the function in the plane

$$f(x, y) = \frac{x^2 - y^2}{\sqrt{x^2 + y^2}}. \tag{11.1.7}$$

It is defined everywhere in the plane except the origin.

Check that $\lim_{(x,y)\to(0,0)} f(x,y) = 0$. Thus we can extend f to the plane by giving it the value 0 at the origin, and the extended function is continuous.

Now consider

$$g(x,y) = \frac{xy}{x^2+y^2}, \tag{11.1.8}$$

which is also defined everywhere but the origin. Switching to polar coordinates, we let $x = r\cos\theta$ and $y = r\sin\theta$. Then after simplification, we get $g(x,y) = \cos\theta\sin\theta = \sin 2\theta/2$. Now a δ-neighborhood of $\mathbf{0}$ is given by $r < \delta$. In other words, to let (x,y) go to $\mathbf{0}$, we let $r \to 0$. We see that no matter how small δ is, $g(x,y)$ takes on all values between $-1/2$ and $1/2$ in the δ-neighborhood, depending on the angle θ at which we approach the origin. Therefore $g(x,y)$ does not have a limit at $\mathbf{0}$. We continue studying this function in Example 12.1.15.

You should compare these two examples to Example 16.3.8 and Example 12.1.29.

11.2 A Theorem

Next we make the connection with limits of sequences.

11.2.1 Theorem. *Let f be a continuous function defined on a set $D \subset \mathbb{R}^n$, and $\{\mathbf{a}_i\}$ a convergent sequence in D such that its limit $\mathbf{a}$ still lies in D. Then*

$$\lim_{i\to\infty} f(\mathbf{a}_i) = f(\mathbf{a}). \tag{11.2.2}$$

Conversely, if for any sequence $\{\mathbf{a}_i\}$ in D converging to $\mathbf{a}$, (11.2.2) is satisfied, then the function f is continuous at $\mathbf{a}$.

Proof. We just provide a sketch of a proof, and leave the details to you. First assume f is continuous at $\mathbf{a}$, and pick an $\epsilon > 0$. Thus we can find a $\delta > 0$ such that all points in $N_\delta(\mathbf{a})$ satisfy $|f(\mathbf{a}) - f(\mathbf{a}_i)| < \epsilon$. We now use this δ as the 'ϵ' for the sequence function $\{\mathbf{a}_i\}$: since it converges to $\mathbf{a}$, there is a N for which all the terms $\mathbf{a}_i$, $i \geq N$ satisfy $d(\mathbf{a}, \mathbf{a}_i) < \delta$. So $|f(\mathbf{a}) - f(\mathbf{a}_i)| < \epsilon$, and we are done.

Note that this is just a special case of the general result saying that the composition of two continuous functions is continuous.

For the converse, do a proof by contradiction: assume f is not continuous, and construct a sequence $\{\mathbf{a}_i\}$ converging to $\mathbf{a}_i$ such that $f(\mathbf{a}_i)$ does not converge to $f(\mathbf{a})$. □

11.2.3 Example. The coordinate functions $f(\mathbf{x}) = x_1$, $f(\mathbf{x}) = x_2$, etc. are continuous, so any polynomial in the x_i is continuous. Any *rational function*, meaning a quotient of polynomials, is therefore continuous except where the denominator is equal to zero. The function $\|\mathbf{x}\|$ is continuous, as a special case of Example 11.1.3.

11.2.4 Exercise. Now consider the real-valued function $f(x, y)$ on the open disk $x^2 + y^2 < 1$ in $\mathbb{R}^2$ such that

$$f(x,y) = \begin{cases} \|(x,y)\| & \text{if } (x,y) \neq \mathbf{0}; \\ 1 & \text{if } (x,y) = \mathbf{0}. \end{cases} \tag{11.2.5}$$

Use Definition 11.1.2 to prove that f defined by (11.2.5) is not continuous at $\mathbf{0}$.

Lecture 12

Differentiation and Taylor's Theorem

We complete the review of multivariable calculus by showing how to approximate real-valued C^2-functions of several variables in a neighborhood of a point of interest. This method is important in optimization. The main theorem is Taylor's theorem, with versions of the mean value theorem and the chain rule proved along the way.

Here is a more detailed summary.

After a review of differentiation in several variables in §12.1, we write the chain rule, both for the first derivative and the second derivative, using linear algebra notation. In particular we introduce the Hessian of a function.

We conclude the lecture with Taylor's theorem in several variables. We will only need it in degrees 1 and 2, so those are the only cases presented. There are many ways to approach Taylor's theorem, in one and several variables, especially in terms of getting expressions for the remainder. For our needs, the Lagrange form of the remainder is enough: see the comment following Theorem 4.2.2.[1]

12.1 Differentiation in Several Variables

In this section we give some of the basic definitions and results for differentiation in several variables, showing how similar they are to the definitions in one variable given in Lecture 3. Even though we are mainly interested in functions $f\colon \mathbb{R}^n \to \mathbb{R}$, we start with functions $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$. We use the notation and definitions of §5.1 throughout.

[1] See [74], §8.4.4 for more details on Taylor's theorem in several variables.

Then we define the Hessian of a function in several variables, and finally we prove Clairault's Theorem 12.1.28, which gives hypothesis under which the mixed partials of a function commute. We will use this theorem throughout the course, applying it to the objective function. It tells us that the Hessian of our objective function is a symmetric matrix when f is sufficiently differentiable, which it usually is.

12.1.1 Differentiability

Referring back to the Definition 11.1.1 of a limit, we can define differentiability by extension of the single-variable case:

12.1.1 Definition. Let $\mathbf{f}$ be a function from an open set $U \subset \mathbb{R}^n$ to $\mathbb{R}^m$. Let $\mathbf{x}$ be a point in U, and $\mathbf{h}$ an arbitrary vector such that $\mathbf{x} + \mathbf{h}$ is in the domain of $\mathbf{f}$. If there exists a $m \times n$ matrix A

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|}{\|\mathbf{h}\|} = 0, \tag{12.1.2}$$

then $\mathbf{f}$ is *differentiable* at $\mathbf{x}$ and A is the *derivative* $\mathbf{f}'(\mathbf{x})$ of $\mathbf{f}$ at $\mathbf{x}$.

A necessary condition for the limit (12.1.2) to exist is that $\lim_{\mathbf{h}\to\mathbf{0}} \|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})\| = \mathbf{0}$, in other words that $\mathbf{f}$ be continuous.

Note that in (12.1.2) the norm in the numerator is that on $\mathbb{R}^m$, while that on the denominator is on $\mathbb{R}^n$. Because U is open, for any sufficiently small $\mathbf{h}$, if $\mathbf{x}$ is in U, then so is $\mathbf{x} + \mathbf{h}$, so we can evaluate f at $\mathbf{x} + \mathbf{h}$.

12.1.3 Theorem. *The matrix* $\mathbf{f}'(\mathbf{x})$ *is uniquely defined.*

The proof is the same as that of Theorem 3.1.3, so we omit it. Limits and continuity of functions of several variables are discussed in §16.1.

As in the one-variable case, write

$$\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) = \mathbf{f}'(\mathbf{x})\mathbf{h} + \mathbf{r}_1(\mathbf{h}) \tag{12.1.4}$$

Then by definition of $\mathbf{f}'(\mathbf{x})$, we have

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{r}_1(\mathbf{h})\|}{\|\mathbf{h}\|} = 0 \tag{12.1.5}$$

Thus the derivative of $\mathbf{f}$ at $\mathbf{x}$ is a $n \times m$ matrix $\mathbf{f}'(\mathbf{x})$. As $\mathbf{x}$ varies, so does the matrix $\mathbf{f}'(\mathbf{x})$.

Therefore the derivative function $\mathbf{f}'(\mathbf{x})$ is a function from $\mathbb{R}^n$ to $\mathbb{R}^{nm}$ – since the set of $n \times m$ matrices is a vector space of dimension nm.[2] What kind of function is $\mathbf{f}'(\mathbf{x})$? Just as in the one-variable case, we say:

[2] We study this space in §**??**.

12.1.6 Definition. The function $\mathbf{f}(\mathbf{x})$ is *continuously differentiable*, written $\mathcal{C}^1$, if $\mathbf{f}'(\mathbf{x})$ is a continuous function

If $\mathbf{f}'(\mathbf{x})$ is differentiable, using Definition 12.1.1 with m replaced by nm, then we write its derivative $\mathbf{f}''(\mathbf{x})$. Then $\mathbf{f}''(\mathbf{x})$ is a function from $\mathbb{R}^n$ to $\mathbb{R}^{n^2m}$.

12.1.7 Definition. If $\mathbf{f}''(\mathbf{x})$ is a continuous function, then $\mathbf{f}(\mathbf{x})$ is *twice continuously differentiable*, written $\mathcal{C}^2$.

We now turn to partial derivatives. To keep the notation manageable, we assume that f maps to $\mathbb{R}$ rather than to $\mathbb{R}^m$. Thus A is just a row vector $(a_1, \dots, a_n)$. Let $\mathbf{e}_i$ be the i-th coordinate vector on $\mathbb{R}^n$, Then fixing the point $\mathbf{x}^* \in \mathbb{R}^n$, let the partial derivative of f with respect to x_i be

$$\frac{\partial f}{\partial x_i}(\mathbf{x}^*) = \lim_{t \to 0} \frac{f(\mathbf{x}^* + t\mathbf{e}_i) - f(\mathbf{x}^*)}{t},$$

if this limit exists. In other words, the i-th partial derivative of f at $\mathbf{x}^*$ is the derivative of the function of one variable

$$g_i(t) = f(\mathbf{x}^* + t\mathbf{e}_i). \tag{12.1.8}$$

12.1.9 Definition (The gradient)**.** Let f be a real-valued function defined on an open set U in $\mathbb{R}^n$ containing point $\mathbf{x}^*$, and assume that all the partial derivatives of f exist at $\mathbf{x}^*$. The *gradient* of $f(\mathbf{x})$ at $\mathbf{x}^*$, written $\nabla f(\mathbf{x}^*)$, is the n- vector

$$\nabla f(\mathbf{x}^*) = \Big(\frac{\partial f}{\partial x_1}(\mathbf{x}^*), \frac{\partial f}{\partial x_2}(\mathbf{x}^*), \dots, \frac{\partial f}{\partial x_n}(\mathbf{x}^*)\Big).$$

The gradient is always a row vector.

12.1.10 Example. The gradient of the function $f(x, y, z) = x \sin y - x^2 z + y/z$ is the row vector

$$(\sin y - 2xz, x \cos y + 1/z, -x^2 - y/z^2).$$

12.1.11 Theorem. *If the real-valued function f is differentiable at $\mathbf{x}^*$, all of its partial derivatives exist at $\mathbf{x}^*$, and*

$$f'(\mathbf{x}^*) = \nabla f(\mathbf{x}^*). \tag{12.1.12}$$

Proof. If $f(\mathbf{x})$ is differentiable, so is the function $g_i(t)$ from (12.1.8), and an easy chain rule computation show that

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \frac{dg_i}{dt}(0) = A\mathbf{e}_i = a_i \tag{12.1.13}$$

so $A = \nabla f(\mathbf{x}^*)$. □

12.1.14 Remark. It is tempting to guess that if all the partials of f exists, then f is differentiable.[3] Unfortunately, if the partial derivatives of f are not continuous, this may fail, as the next example shows.

12.1.15 Example. In (11.1.8) we considered the function[4]

$$f(x,y) = \begin{cases} \frac{xy}{x^2+y^2} & \text{if } (x,y) \neq \mathbf{0}; \\ 0 & \text{if } (x,y) = \mathbf{0}. \end{cases} \tag{12.1.16}$$

and showed that it is not continuous at **0**, and therefore not differentiable at **0**. On the other hand, both its partials exist at all points. Indeed, when $xy \neq 0$,

$$\frac{\partial f}{\partial x}(x,y) = \frac{y(y^2-x^2)}{(x^2+y^2)^2}, \quad \frac{\partial f}{\partial y}(x,y) = \frac{x(x^2-y^2)}{(x^2+y^2)^2},$$

while both its partial derivatives at **0** exist, and are equal to zero, as a direct limit computation shows. For example

$$\frac{\partial f}{\partial x}(0,0) = \lim_{x\to 0} \frac{f(x,0)-f(0,0)}{x} = \lim_{x\to 0} \frac{0}{x} = 0$$

Thus the functions $\frac{\partial f}{\partial x}(x,y)$ and $\frac{\partial f}{\partial y}(x,y)$ are defined everywhere, but they are not continuous at **0**. For instance

$$\frac{\partial f}{\partial x}(0,y) = 1/y,$$

which does not approach 0 as $y \to 0$.

12.1.17 Theorem. *f is $\mathcal{C}^1$ on a open set U in $\mathbb{R}^n$ if and only if all its partial derivatives exist and are continuous on U.*

Proof. The ⇒ implication follows from Theorem 12.1.11. For the ⇐ implication, we show that if all the partial derivatives exist and are continuous on U, then f is $\mathcal{C}^1$. For simplicity we only write the proof when $n = 2$. It is a good exercise in summation notation to write down the general case. Fix a point $\mathbf{x}^* = (a,b)$ of the domain of f and small numbers h and k so that the rectangle bounded by $(a \pm h, b \pm k)$ is entirely in the domain U of f. This is possible since U is open. Write $\mathbf{h} = (h,k)$.

We first show f is differentiable, with derivative given by the partials. In other words, we show that the limit of

$$\frac{f(\mathbf{x}^* + \mathbf{h}) - f(\mathbf{x}^*) - h\frac{\partial f}{\partial x}(\mathbf{x}^*) - k\frac{\partial f}{\partial y}(\mathbf{x}^*)}{\|\mathbf{h}\|} \tag{12.1.18}$$

[3] As Cauchy did in [16], §33.

[4] See [48] §I.3, for an early reference.

as $\mathbf{h} \to \mathbf{0}$ is 0.

Write

$$f(a+h,b+k)-f(a,b)=f(a+h,b+k)-f(a,b+k)+f(a,b+k)-f(a,b)$$

and apply the mean value theorem twice to get

$$\begin{aligned} f(\mathbf{x}^*+\mathbf{h})-f(\mathbf{x}^*) &= f(a+h,b+k)-f(a,b) \\ &= h\frac{\partial f}{\partial x}(a+\lambda h,b+k)+k\frac{\partial f}{\partial y}(a,b+\mu k), \end{aligned}$$

where $0<\lambda<1$, and $0<\mu<1$.

Now substitute this into (12.1.18) to get

$$\frac{h\frac{\partial f}{\partial x}(a+\lambda h,b+k)+k\frac{\partial f}{\partial y}(a,b+\mu k)-h\frac{\partial f}{\partial x}(\mathbf{x}^*)-k\frac{\partial f}{\partial y}(\mathbf{x}^*)}{\|\mathbf{h}\|}$$

or

$$\frac{h\Big(\frac{\partial f}{\partial x}(a+\lambda h,b+k)-\frac{\partial f}{\partial x}(\mathbf{x}^*)\Big)+k\Big(\frac{\partial f}{\partial y}(a,b+\mu k)-\frac{\partial f}{\partial y}(\mathbf{x}^*)\Big)}{\|\mathbf{h}\|} \tag{12.1.19}$$

Now use the fact that the partials are continuous, which implies that as $\mathbf{h} \to 0$,

$$\begin{aligned} \frac{\partial f}{\partial x}(a+\lambda h,b+k) &\to \frac{\partial f}{\partial x}(\mathbf{x}^*) \\ \frac{\partial f}{\partial y}(a,b+\mu k) &\to \frac{\partial f}{\partial y}(\mathbf{x}^*) \end{aligned}$$

Then as $\mathbf{h} \to 0$, the limit of the terms in the big parentheses in (12.1.19) is 0, which shows that f is differentiable. The derivative is given by the gradient according to Theorem 12.1.11. Since the gradient is continuous by assumption, f is $\mathcal{C}^1$, and we are done. □

12.1.2 The Hessian

Next we define the Hessian of a real-valued function f of n variables. We denote the Hessian of a function by the corresponding upper case letter. So the Hessian of f is F.

Recall that

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}) \text{ means } \frac{\partial}{\partial x_j}\Big(\frac{\partial f(\mathbf{x})}{\partial x_i}\Big).$$

12.1.20 Definition. Assume that the function $f(\mathbf{x})$ of n variables is $\mathcal{C}^1$ and that all its second partials exist. Let

$$F_{ij}(\mathbf{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}) \quad \text{for} \quad 1 \le i, j \le n.$$

The $n \times n$ matrix $F(\mathbf{x}) = [F_{ij}(\mathbf{x})]$ is called the *Hessian* matrix of f at $\mathbf{x}$.

In Corollary 12.1.28 we give a simple hypothesis under which the Hessian is a symmetric matrix. Thus for $i \neq j$, $F_{ij}(\mathbf{x}) = F_{ji}(\mathbf{x})$. We say that the mixed partials commute, meaning that the order in which the derivatives are taken is irrelevant. For most of the functions f we use as objective function in these notes, this hypothesis will be satisfied.

12.1.21 Example. Let $f(x, y) = x^3 + 2x^2y - xy^2 - y$, so

$$\frac{\partial f}{\partial x} = 3x^2 + 4xy - y^2, \text{ and } \frac{\partial f}{\partial y} = 2x^2 + -2xy - 1.$$

Computing the y-partial of the first term, and the x-partial of the second, we get the same result:

$$\frac{\partial^2 f}{\partial x \partial y} = 4x - 2y = \frac{\partial^2 f}{\partial y \partial x},$$

so the mixed partials are the same.

12.1.3 Partial Derivatives Commute

The last goal of this section is to prove Clairault's Theorem 12.1.28 that "mixed partials commute".[5]

To prove it, we need another generalization of the mean value theorem 3.2.1.

12.1.22 Theorem. *The function $f(x, y)$ is defined on the open ball $B = B_r(\mathbf{p})$ of radius r around the point $\mathbf{p} = (a, b) \in \mathbb{R}^2$. Pick non-zero reals h and k so that the point $(a + h, b + k)$ is in B. Assume that the partial derivatives $\partial f/\partial x$ and $\partial^2 f/\partial y \partial x$ exist in B. Consider*

$$\Delta(h, k) = f(a + h, b + k) - f(a + h, b) - f(a, b + k) + f(a, b) \qquad (12.1.23)$$

Then there are numbers λ and μ, with $0 < \lambda < 1$ and $0 < \mu < 1$ such that

$$\Delta(h, k) = kh \frac{\partial^2 f}{\partial y \partial x}(a + \lambda h, b + \mu k) \qquad (12.1.24)$$

[5] See [63], p.916 for a multivariable calculus reference, and [55], p. 235 for a statement with fewer hypotheses.

Proof. We apply the mean value theorem twice. Let $u(t) = f(t, b + k) - f(t, b)$. Then $u(t)$ is differentiable by hypothesis, and its derivative is

$$u'(t) = \frac{\partial f}{\partial x}(t, b + k) - \frac{\partial f}{\partial x}(t, b).$$

We define $\Delta(h, k) = u(a + h) - u(a)$. By the mean value theorem we get:

$$\Delta(h, k) = hu'(a + \lambda h),$$

for some λ with $0 < \lambda < 1$. Using the definition of $u(t)$ we get

$$\Delta(h, k) = h\big(\frac{\partial f}{\partial x}(a + \lambda h, b + k) - \frac{\partial f}{\partial x}(a + \lambda h, b)\big) \tag{12.1.25}$$

The function $\partial f/\partial x$ is differentiable as a function of the y variable. We apply the mean value theorem a second time, this time to

$$w(t) = \frac{\partial f}{\partial x}(a + \lambda h, t),$$

on the interval $[b, b + k]$. This gives a μ, $0 < \mu < 1$, so that

$$kw'(b + \mu k) = w(b + k) - w(b).$$

We recognize the right-hand side of (12.1.25) as $h(w(b + k) - w(b)$, so (12.1.24) follows by substituting out this term using our last chain rule computation. □

12.1.26 Theorem (Clairault's Theorem). *Suppose $f(x, y)$ is defined on an open set $U \subset \mathbb{R}^2$, and that the partial derivatives $\partial f/\partial x$, $\partial f/\partial y$, $\partial^2 f/\partial y\partial x$, $\partial^2 f/\partial x\partial y$ exist on U, and are continuous at the point $\mathbf{p} = (a, b) \in U$. Then*

$$\frac{\partial^2 f}{\partial x\partial y}(\mathbf{p}) = \frac{\partial^2 f}{\partial y\partial x}(\mathbf{p})$$

Proof. We apply Theorem 12.1.22 twice, first as written, and then with the roles of x and y interchanged, so that the order of the partials in (12.1.24) is reversed. Let

$$c = \frac{\partial^2 f}{\partial y\partial x}(\mathbf{p}) \text{ and } d = \frac{\partial^2 f}{\partial x\partial y}(\mathbf{p})$$

Then by continuity of $\partial^2 f/\partial y\partial x$ and $\partial^2 f/\partial x\partial y$ at $\mathbf{p}$, for any $\epsilon > 0$ there is a ball $B_r(\mathbf{p})$ such that for any $(a + h, y + k) \in B_r(\mathbf{p})$,

$$\Big|c - \frac{\Delta(h, k)}{hk}\Big| < \epsilon, \text{ and } \Big|d - \frac{\Delta(h, k)}{hk}\Big| < \epsilon. \tag{12.1.27}$$

This implies $c = d$ as required. □

12.1.28 Corollary. *Let $f(\mathbf{x})$ be a $\mathcal{C}^2$-function on an open set $U \in \mathbb{R}^n$, and let $F(\mathbf{x})$ be the Hessian of f at $\mathbf{x}$. Then $F(\mathbf{x})$ is a symmetric matrix for all $\mathbf{x} \in U$.*

We use the function in the next exercise to construct a function where the mixed partials fail to commute.

12.1.29 Exercise. Consider the function in the plane minus the origin

$$g(x, y) = \frac{x^2 - y^2}{x^2 + y^2}.$$

Compute the limit of $g(x, y)$ as you approach the origin along the line $y = tx$, for a given slope t, namely compute

$$\lim_{t \to 0} g(x, tx).$$

Also compute $\lim_{t \to 0} g(ty, y)$. Your computation will show that you get different results: thus $g(x, y)$ does not have a limit as you approach $\mathbf{0}$, so $g(x, y)$ cannot be extended to a continuous function in the plane.

12.1.30 Example. This example, given by Peano in [26], Annotazione N. 103, is one of the earliest examples of a function where both mixed partials exist, and yet are not equal. Let

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

Note that when $(x, y) \neq (0, 0)$, this is $xyg(x, y)$ for the function $g(x, y)$ of Exercise 12.1.29. Show that f is $\mathcal{C}^1$, so that the second partials exist at $\mathbf{0}$. Then compute the two mixed partials of f, notice that they are both defined at $\mathbf{0}$, and evaluate:

$$\frac{\partial^2 f}{\partial x \partial y}(0, 0) = 1 \quad \text{and} \quad \frac{\partial^2 f}{\partial y \partial x}(0, 0) = -1,$$

so Theorem 12.1.26 fails[6]. This is because the second partials are not continuous at $\mathbf{0}$, as you should check.

There is an analogous result for higher order partials, but it will not concern us.

[6] See Rudin [55], Exercise 27 p.242, Zorich [74], p.474, [28], Counterexample IV.4.2 p. 316.

12.2 The Chain Rule and Linear Algebra

As usual let $f(x_1, \dots, x_n)$ be a $\mathcal{C}^1$ real-valued function from an open neighborhood $U \in \mathbb{R}^n$ of a point $\mathbf{x}^*$.

Let $\mathbf{g}(\mathbf{y})$ be a vector valued function: $\mathbb{R}^p \to \mathbb{R}^n$, with coordinate functions $g_j(y_1, \dots, y_p)$, $1 \leq j \leq n$. Assume $\mathbf{g}$ is defined in a neighborhood V of a point $\mathbf{y}^*$ in $\mathbb{R}^p$ that maps to the point $\mathbf{x}^*$, and that $\mathbf{g}(V) \subset U$.[7] Then we can define the *composite* $\varphi = f(\mathbf{g})$ of the two functions near $\mathbf{y}^*$.

$$\varphi(y_1, \dots, y_p) = f(g_1(y_1, \dots, y_p), \dots, g_n(y_1, \dots, y_p)), \tag{12.2.1}$$

or more compactly $\varphi(\mathbf{y}) = f(\mathbf{g}(\mathbf{y}))$. Thus $\varphi(\mathbf{y})\colon \mathbb{R}^p \to \mathbb{R}$.

We assume that both f and $\mathbf{g}$ are $\mathcal{C}^1$. The chain rule shows that the composite φ is $\mathcal{C}^1$, too, and it computes its partial derivatives. We will record this information in the gradient vectors of the functions. Recall that gradients of functions are written as row vectors. So $\nabla\varphi$ is a row vector of length p, ∇f a row vector of length n and $\nabla\mathbf{g}$ a $n \times p$ matrix whose (j, k) entry is $\partial g_j / \partial y_k$.

We now use the chain rule to compute the gradient $\nabla\varphi$. Before doing this, we recall the familiar result from single-variable calculus, where $n = p = 1$. Then the ordinary chain rule says:

$$\varphi'(y) = f'(g(y))g'(y), \tag{12.2.2}$$

or, in words, the derivative of the composite function is the derivative of the outside function evaluated at the value of the inside function, time the derivative of the inside function.

12.2.3 Theorem. *The functions $f(\mathbf{x})$, $\mathbf{g}(\mathbf{y})$ and their composite $\varphi(\mathbf{y})$ are as above: in particular f and $\mathbf{g}$ are $\mathcal{C}^1$. Then φ is also $\mathcal{C}^1$, and*

$$\nabla\varphi(\mathbf{y}) = \nabla f(\mathbf{g}(\mathbf{y}))\nabla\mathbf{g}(\mathbf{y}),$$

so evaluating at $\mathbf{y}^*$, remembering that $\mathbf{x}^* = \mathbf{g}(\mathbf{y}^*)$:

$$\nabla\varphi(\mathbf{y}^*) = \nabla f(\mathbf{x}^*)\nabla\mathbf{g}(\mathbf{y}^*).$$

Note how this generalizes the ordinary chain rule, with derivatives being replaced by gradients. What is the product in the new chain rule? ∇f is a row vector of length n, and $\nabla\mathbf{g}(\mathbf{y})$ is a $n \times p$ matrix, so the product is matrix multiplication, and the result is a row vector of length p, as required.

[7] An elementary approach to this material is given in Stewart [63], §14.5. A more advanced reference is Strichartz [70] §10.1.3.

Proof. Carry out the computation of each partial derivative of f, and then fit them together in a matrix.

$$\frac{\partial \varphi}{\partial y_k}(\mathbf{y}) = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\mathbf{g}(\mathbf{y}))\frac{\partial g_j}{\partial y_k}(\mathbf{y}) = \nabla f(\mathbf{g}(\mathbf{y}))\frac{\partial \mathbf{g}}{\partial y_k}(\mathbf{y}) \tag{12.2.4}$$

where the product on the right is that of the row vector $\nabla f(\mathbf{g}(\mathbf{y}))$ by the k-th column of the $n \times p$ matrix $\nabla \mathbf{g}$. For more detains see for example Stewart [63], §14.5. □

Next we compute the Hessian of the composite function φ (12.2.1). We assume that the functions f and $\mathbf{g}$ are $\mathcal{C}^2$, so that the composite φ is also, by the chain rule applied to the first derivatives. We write F for the $n \times n$ Hessian matrix of f, Φ for the $p \times p$ Hessian matrix of φ and G_j for the $p \times p$ Hessian of g_j.

To see what to expect, we first work out the answer when $n = p = 1$, as in the ordinary chain rule. So we differentiate with respect to y the expression in (12.2.2), using the chain rule and the product rule, getting

$$\varphi''(y) = f''(g(y))(g'(y))^2 + f'(g(y))g''(y). \tag{12.2.5}$$

12.2.6 Example. An important special case of the general formula is the case where $\mathbf{g}$ is a function from $\mathbb{R}$ to $\mathbb{R}^n$. We say that $\mathbf{g}$ is a *parametrized curve*. So we are treating the case $p = 1$, n arbitrary. We use the name t for the variable of $\mathbf{g}$, which has n coordinate functions $g_1(t)$, ..., $g_n(t)$. So the composite $\varphi(t) = f(\mathbf{g}(t))$ is a scalar function of one variable. Let us compute its first and second derivatives. Let $\mathbf{g}'(t)$ be the column vector $(g_1'(t), \dots, g_n'(t))$, and let $\mathbf{g}''(t)$ be the column vector $(g_1''(t), \dots, g_n''(t))$. Then by specializing Theorem 12.2.3 to this case we get:

$$\varphi'(t) = \nabla f(\mathbf{g}(t))\mathbf{g}'(t), \tag{12.2.7}$$

the product of a row n-vector by a column n-vector, giving a number as required. Differentiating again, writing out the dot product on the right-hand side, we can organize the result as

$$\varphi''(t) = \mathbf{g}'(t)^T F(\mathbf{g}(t))\mathbf{g}'(t) + \nabla f(\mathbf{g}(t))\mathbf{g}''(t). \tag{12.2.8}$$

Notice how the two terms on the right-hand side match the two terms in (12.2.5). If the parametrized curve is a line, meaning that the functions $g_i(t)$ are linear in t, then $\mathbf{g}''(t) = \mathbf{0}$, so the second derivative is

$$\varphi''(t) = \mathbf{g}'(t)^T F(\mathbf{g}(t))\mathbf{g}'(t).$$

Finally, the general case.

12.2.9 Theorem. *The chain rule for Hessians gives:*

$$\Phi(\mathbf{y}) = \nabla\mathbf{g}(\mathbf{y})^T F(\mathbf{g}(\mathbf{y}))\nabla\mathbf{g}(\mathbf{y}) + \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\mathbf{g}(\mathbf{y}))G_j(\mathbf{y}) \tag{12.2.10}$$

so evaluating at $\mathbf{y}^*$ *we get*

$$\Phi(\mathbf{y}^*) = \nabla\mathbf{g}(\mathbf{y}^*)^T F(\mathbf{x}^*)\nabla\mathbf{g}(\mathbf{y}^*) + \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\mathbf{x}^*)G_j(\mathbf{y}^*)$$

Proof. This is an easy, if complicated, exercise using the product rule for differentiation. You should first check that the sizes of the matrices are correct. The dimension of most of the matrices is given above: you also need that $\nabla\mathbf{g}(\mathbf{y})$ is a $n \times p$ matrix, whose i-th row is the gradient of g_i. Finally $\frac{\partial f}{\partial x_j}$ is a scalar. Just use the product rule combined with the ordinary chain rule, computing one entry of the symmetric matrix Φ at a time, namely

$$\Phi_{i,k} = \frac{\partial^2 \varphi}{\partial y_i \partial y_k}$$

taking $\frac{\partial}{\partial y_i}$ of (12.2.4). You get

$$\begin{aligned}\Phi_{i,k} &= \sum_{j=1}^{n} \Big(\frac{\partial}{\partial y_i}\Big(\frac{\partial f}{\partial x_j}(\mathbf{g}(\mathbf{y}))\Big)\frac{\partial g_j}{\partial y_k}(\mathbf{y}) + \frac{\partial f}{\partial x_j}(\mathbf{g}(\mathbf{y}))\frac{\partial}{\partial y_i}\Big(\frac{\partial g_j}{\partial y_k}(\mathbf{y})\Big)\Big) \\ &= \sum_{j=1}^{n} \Big(\sum_{l=1}^{n} \Big(\frac{\partial^2 f}{\partial x_l \partial x_j}(\mathbf{g}(\mathbf{y}))\frac{\partial g_l}{\partial y_i}(\mathbf{y})\Big)\frac{\partial g_j}{\partial y_k}(\mathbf{y}) + \frac{\partial f}{\partial x_j}(\mathbf{g}(\mathbf{y}))\frac{\partial^2 g_j}{\partial y_i \partial y_k}(\mathbf{y})\Big)\end{aligned}$$

and you conclude by recognizing this as the appropriate term of (12.2.10). □

12.2.11 Exercise. Verify that the terms on the right-hand side of (12.2.8) have the right dimensions to yield a number.

12.2.12 Exercise. Let $f(x, y, z) = x^2 + y^2 + z^2$ and let $\mathbf{g}(t) = (t, t^2, t^3)$. Compute $\mathbf{g}'(t)$ and the gradient $\nabla f(x, y, z)$. Compute the matrix product in (12.2.7). On the other hand, substitute $x = t$, $y = t^2$ and $z = t^3$ into f, and compute $\varphi'(t)$ directly, confirming your first answer. Next, repeat with the second derivative, so compute the Hessian F of f, and compute the matrix product in (12.2.8). Then, just as you did for the first derivative, compute $\varphi''(t)$ directly, confirming your answer.

12.2.13 Exercise. Let $f(x,y,z) = x^2+y^2+z^2$ again, and let $\mathbf{g}(r,\theta) = (r\cos\theta, r\sin\theta, r)$. The composite $\varphi(r,\theta)$ is a function from $\mathbb{R}^2$ to $\mathbb{R}$. Compute its gradient and its Hessian directly, and then confirm your answer using the chain rule.

We will use Theorem 12.2.9 in (17.7.2) and in §29.1.

12.3 Homogeneous Functions

As a simple example of the chain rule, we look at homogeneous functions.

12.3.1 Definition. A real-valued function $f(x_1,\dots,x_n)$ is *homogeneous* of degree $d \in \mathbb{Z}$, if for any positive real number t,

$$f(t\mathbf{x}) = f(tx_1, tx_2, \dots, tx_n) = t^d f(\mathbf{x}). \tag{12.3.2}$$

Homogeneous functions arise in many branches of mathematics and economics. They are functions that are “independent of scale”, as a moment’s thought should tell you. The central theorem concerning homogeneous functions is due to Euler:

12.3.3 Theorem (Euler’s Theorem)**.** *If $f(\mathbf{x})$ is $\mathcal{C}^1$ and homogeneous of degree d, then its partial derivatives $\partial f/\partial x_i$ are homogenous of degree $d-1$ and we have Euler’s formula:*

$$\sum_{j=1}^{n} x_j \frac{\partial f}{\partial x_j} = df(\mathbf{x}). \tag{12.3.4}$$

Proof. For any fixed $\mathbf{x}$ we let $\mathbf{g}(t) = t\mathbf{x}$, so $\mathbf{g}(t)$ is a function from $\mathbb{R}$ to $\mathbb{R}^n$, and $f(t\mathbf{x})$ is the composite $f(\mathbf{g}(t)) = f(tx_1, tx_2, \dots, tx_n)$. Differentiate both sides of (12.3.2) with respect to t. The right-hand side gives $dt^{d-1}f(\mathbf{x})$. The left-hand side of (12.3.2) is the composite $f(\mathbf{g}(t))$, so we compute the derivative with respect to t using the chain rule. Clearly $g'(t) = \mathbf{x}$, so we get

$$\frac{d}{dt}(f(\mathbf{g}(t)) = \langle \nabla_{t\mathbf{x}} f, \mathbf{x} \rangle$$

so equating the derivatives, we get

$$\sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(t\mathbf{x}) x_j = dt^{d-1} f(\mathbf{x}).$$

Set $t = 1$ to get Euler’s formula.

To show that all the partials of f are homogeneous of degree $d-1$, take the partial of (12.3.2) with respect to x_j. The right-hand side gives $t^d \partial f(\mathbf{x})/\partial x_j$ again. The left-hand side, by the chain rule again, gives $t\partial f(t\mathbf{x})/\partial x_j$. Equating the two and dividing by t, we get the result. □

12.3.5 Example. The prototypical example of a function $f(x_1, \dots, x_n)$ that is homogenous of degree d, is a homogeneous polynomial of degree d. First a concrete example of degree 4 in 3 variables:

$$x^4 + x^2z^2 - 3xy^2z + yz^3$$

You should verify homogeneity and Euler's formula for this example.

More generally, a polynomial in n-variables is homogenenous of degree d if each monomial is of the form: $x_1^{a_1}x_2^{a_2}\dots x_n^{a_n}$, for non-negative integers a_j, where $\sum_{j=1}^n a_j = d$.

Finally the ratio f/g of a homogeneous polynomial of degree d_1 by a homogeneous polynomial of degree d_2 is homogeneous of degree $d = d_1 - d_2$. Thus the function in Example 12.1.29 is homogeneous of degree 0.

12.4 Taylor Polynomials in Several Variables

This section extends our investigation of Taylor polynomials to the case of several variables.

Let $f(\mathbf{x})$ be a function from $\mathbb{R}^n$ to $\mathbb{R}$ defined on S. Fix a point $\mathbf{a} = (a_1, \dots, a_n)$ in the interior of domain S. For the purposes of this section, we assume that f is C^2 (see Definition 12.1.7) in a neighborhood of the point $\mathbf{a}$. As in the case of one variable, we wish to approximate f by a polynomial in a neighborhood of a. This will be a polynomial in the n variables $x_1, x_2, \dots, x_n$.

In this course, we only need the Taylor polynomials of degrees 1 and 2. The Taylor polynomial of degree 1 is

$$P_1(\mathbf{x}) = f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot (\mathbf{x} - \mathbf{a}).$$

so its vanishing gives the tangent space to the graph of f at $\mathbf{a}$.

12.4.1 Definition. The *second-degree Taylor polynomial* of f at the point $\mathbf{a}$ is

$$P_2(\mathbf{a}, \mathbf{x}) = f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot (\mathbf{x} - \mathbf{a}) + \frac{(\mathbf{x} - \mathbf{a})^T F(\mathbf{a})(\mathbf{x} - \mathbf{a})}{2} \tag{12.4.2}$$

12.4.3 Exercise. Verify that when $n = 1$, , the expression in (12.4.2) produces the one-variable Taylor polynomial.

12.4.4 Exercise. How many quadratic terms are there in n variables x_1,x_2 ... x_n? In two variables x_1 and x_2, we have the constant term 1, the two linear terms x_1 and x_2, and the three quadratic terms x_1^2, x_1x_2, and x_2^2. In three variables x_1, x_2 and x_3, we have the constant term 1, the three linear terms x_1, x_2, and x_3, and the six quadratic terms x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, and x_2x_3. What is the general formula?

The first step in establishing Taylor's theorem is to work one direction at a time and apply Taylor's theorem in one variable. So we fix a vector $\mathbf{h} = (h_1, h_2, \dots, h_n)$ and define a new function

$$g(t) = f(\mathbf{a} + t\mathbf{h})$$

of the single variable t, assuming t is sufficiently small that $\mathbf{a} + t\mathbf{h}$ is in the domain of f, and f is C^2 there. Obviously $g(0) = f(\mathbf{a})$. Note that $g(t)$ is the composite of the function

$$t \mapsto (a_1 + h_1 t, a_2 + h_2 t, \dots, a_n + h_n t) = \mathbf{a} + t\mathbf{h}$$

and the function $f(\mathbf{x})$, by setting, for each index i, $x_i = a_i + h_1 t$. We differentiate $g(t)$ using the chain rule.

$$\frac{dg}{dt}(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{a} + t\mathbf{h}) \frac{dx_i}{dt} = \nabla f(\mathbf{a} + t\mathbf{h}) \cdot \mathbf{h} \tag{12.4.5}$$

We differentiate a second time using the chain rule. Since the a_i and h_i are constants, we get

$$\frac{d^2 g}{dt^2}(t) = \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a} + t\mathbf{h}) h_i h_j \tag{12.4.6}$$

We can rewrite the right-hand side of (12.4.6) as the product of three matrices:

$$\sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a} + t\mathbf{h}) h_i h_j = \mathbf{h}^T F(\mathbf{a} + t\mathbf{h})\mathbf{h} \tag{12.4.7}$$

where $\mathbf{h}$ is a column vector, and $\mathbf{h}^T$ is its transpose. The middle $n \times n$ matrix $F(\mathbf{x} + t\mathbf{h})$ is the Hessian matrix of f evaluated at $\mathbf{x} + t\mathbf{h}$. The product of these three matrices—the size of which, going from left to right, being $1 \times n$, $n \times n$, and $n \times 1$—makes sense and is a scalar, as required.

Now we write the second-degree Taylor polynomial of $g(t)$ in terms of the Hessian of f. First we adapt the Generalized Mean Value Theorem 4.2.2 to this situation.

12.4.8 Theorem (GMVT in Several Variables, Degree 2 Case)**.** *Assume the function f is $\mathcal{C}^2$ on an open region containing the line segment $[\mathbf{a}, \mathbf{a} + \mathbf{h}]$. Then*

$$f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot \mathbf{h} + \frac{\mathbf{h}^T F(\mathbf{a} + t^* \mathbf{h})\mathbf{h}}{2} \tag{12.4.9}$$

for some t^*, $0 \le t^* < 1$.

Proof. According to Theorem 4.2.2 in degree 2 applied to g, we have

$$g(t) = g(0) + t\frac{dg}{dt}(0) + \frac{1}{2}\frac{d^2g}{dt^2}(t^*)t^2 \tag{12.4.10}$$

where t^* is a value between 0 and t.

Substituting the values we have computed for the first two derivatives of g in (12.4.5) and (12.4.6), and substituting in (12.4.7), produces

$$f(\mathbf{a}+t\mathbf{h}) = f(\mathbf{a}) + t\nabla f(\mathbf{a})\cdot\mathbf{h} + \frac{\mathbf{h}^T F(\mathbf{a}+t^*\mathbf{h})\mathbf{h}}{2}t^2 \tag{12.4.11}$$

To get (12.4.9) just set $t = 1$ in (12.4.11). □

12.4.12 Theorem (Taylor's Theorem in Several Variables, Degree 2 Case)**.** *Assume the function f is $\mathcal{C}^2$ on an open region containing the line segment $[\mathbf{a}, \mathbf{a}+\mathbf{h}]$. Then*

$$f(\mathbf{a}+\mathbf{h}) = f(\mathbf{a}) + \nabla f(\mathbf{a})\cdot\mathbf{h} + \frac{\mathbf{h}^T F(\mathbf{a})\mathbf{h}}{2} + r_2(\mathbf{h}) \tag{12.4.13}$$

where $r_2(\mathbf{h})$ is a function of $\mathbf{h}$ such that $\lim_{\mathbf{h}\to\mathbf{0}} \frac{r_2(\mathbf{h})}{\|\mathbf{h}\|^2} = 0$.

Proof. We argue as in the corresponding proof in the one variable case: Theorem 4.3.2. We subtract (12.4.13) from (12.4.9) to get

$$\begin{aligned} r_2(\mathbf{h}) &= \frac{\mathbf{h}^T F(\mathbf{a}+t^*\mathbf{h})\mathbf{h}}{2} - \frac{\mathbf{h}^T F(\mathbf{a})\mathbf{h}}{2} \\ &= \frac{\mathbf{h}^T \big(F(\mathbf{a}+t^*\mathbf{h}) - F(\mathbf{a})\big)\mathbf{h}}{2} \end{aligned} \tag{12.4.14}$$

Now divide by $\|\ \mathbf{h}\ \|^2$. The right-hand side of (12.4.14) is a sum of terms each having a product $h_i h_j$ in factor, one h_i from the $\mathbf{h}$ on the left, and the other h_j from the $\mathbf{h}$ on the right. Clearly $|h_i h_j| \le\|\ \mathbf{h}\ \|^2$. What is left in the sum are terms of the form

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{a}+t^*\mathbf{h}) - \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{a}) \tag{12.4.15}$$

and because f is $\mathcal{C}^2$, these go to 0 as $\mathbf{h}\to\mathbf{0}$, proving the result.[8]

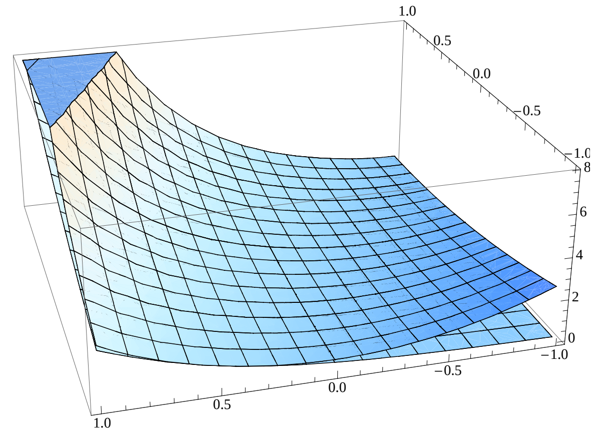
□

[8] For a different proof, see [70], §10.2.3.

12.4.16 Example. Consider $f(x,y) = e^{x+2y}$ at $\mathbf{a} = (0,0)$. Then $f(0,0) = 1$ and $\nabla f(0,0) = (1,2)$. The Hessian at the origin is

$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \tag{12.4.17}$$

and is symmetric . Then the approximation for e^{x+2y} at the origin is $1 + x + 2y + \frac{1}{2}(x^2 + 4xy + 4y^2)$. Below is a graph of $f(x,y)$ together with the approximation (below it):



12.4.18 Question. Can one determine if the function f has a local minimum or maximum at an interior point $\mathbf{a}$, just from its quadratic approximation given by Theorem 12.4.12? The answer is often yes.

The tools for answering this question will be developed in the next lecture on linear algebra, and the question is answered in §13.1. In some cases, an answer cannot be given just from the quadratic information: see §13.5.

Lecture 13

Unconstrained Optimization

We now have all the tools needed to do multivariable optimization. The key results in this lecture are in §13.1: they are Theorem 13.1.2, and especially Theorem 13.1.3. To prove them, we need the whole repertory of mathematics we have developed: Taylor's Theorem and the Rayleigh quotient, as well as our detailed study of the Hessian of functions.

The two theorems correspond to sufficient and necessary conditions for a minimum. In §13.5 we discuss the ambiguous case, meaning the case where we are at a critical point where the Hessian is positive semidefinite but not positive definite, so that the necessary condition is satisfied, but the sufficent condition fails. We will see that even in two dimensions the situation is complicated: we give the classic example of Peano 13.5.3.

In §13.3 we treat the method of Least Squares, sometimes described as the most important optimization result, because of its innumerable applications. There are another long optimization example worth mentioning: the maximization of the Cobb-Douglas utility function 13.4.1, used extensively in economics.

Then in §13.7 we turn to the description of the level sets of a positive definite Hessian: they are ellipsoids. We analyze them by finding their principal axes, which means finding their eigenvectors. This is a geometric interpretation and refinement of what was done in Lecture 8, and is not needed in the rest of the lectures.

13.1 The Theorems

We state the main results of unconstrained optimization. The main goal of the rest of this course is to generalize these results to the case where the set over which optimization is considered is a subset of $\mathbb{R}^n$ defined by equalities (see Lecture 28 and 29) or inequalities (see Lecture 31 and 23).

The outline will be the same in all cases: first we prove a necessary condition for an extremum, assuming only that the objective function is $\mathcal{C}^1$. Because only first derivatives are involved, this is called a first-order condition. Then we turn to second-order conditions, meaning conditions involving the second derivative: first a necessary condition and then a sufficient condition.

Here are the three results in the current situation.

13.1.1 Theorem. *Let $f(\mathbf{x})$ be a real-valued $\mathcal{C}^1$ function defined on an open set U in $\mathbb{R}^n$. Assume that $\mathbf{a} \in U$ is a local minimizer or maximizer for f. Then a is a critical point of f: the gradient ∇f vanishes at $\mathbf{a}$.*

Proof. If $\mathbf{a}$ is not a critical point there is a direction $\mathbf{v}$ in $\mathbb{R}^n$ in which the directional derivative of f is non-zero. Restricting f to the line $\ell = \mathbf{a} + t\mathbf{v}$ parametrized by a real variable t, you get a function $g(t) = f(\mathbf{a}+t\mathbf{v})$ in one variable t: the restriction of f to the line ℓ. The function $g(t)$ has a minimum or maximum at $t = 0$, since f does at $\mathbf{a}$. But then single-variable calculus tells us that g has a critical point (the derivative g' is 0) , which is a contradiction, since the derivative of g is the directional derivative of f in direction $\mathbf{v}$. □

13.1.2 Theorem. *Let $f(\mathbf{x})$ be a real-valued $\mathcal{C}^2$ function defined on an open set U in $\mathbb{R}^n$. Assume that $\mathbf{a} \in U$ is a local minimizer for f. Then the Hessian F of f at $\mathbf{a}$ is positive semidefinite.*

Proof. The proof is almost the same as that of the previous theorem: assume the Hessian at $\mathbf{a}$ is not positive semidefinite. Then by the Spectral Theorem f has a negative eigenvalue with an associated real eigenvector $\mathbf{e}$. Again restricting f to the line ℓ containing the vector $\mathbf{e}$, the restriction $g(t) = f(\mathbf{a} + t\mathbf{e})$ has derivative 0 and has negative second derivative, so $\mathbf{a}$ cannot be a minimizer for g by the well-known theorem from single-variable calculus: see Theorem 3.3.4. □

A similar theorem holds for maxima. Thus we have established easy *necessary* conditions for a local maximum or minimum. Next we turn to sufficient conditions, given by the next theorem, the main result of this section.

13.1.3 Theorem. *Let $f(\mathbf{x})$ be a real-valued $\mathcal{C}^2$ function defined on an open set U in $\mathbb{R}^n$. Assume that $\mathbf{a} \in U$ is a critical point of f and that the Hessian F of f is positive definite at $\mathbf{a}$. Then $f(\mathbf{x})$ has a strict local minimum at $\mathbf{a}$.*

Proof. This is a consequence of Taylor's theorem in several variables. Since $\mathbf{a}$ is a critical point, $\nabla f(\mathbf{a}) = 0$. Write $\mathbf{x} = \mathbf{a} + \mathbf{h}$ for an arbitrary point $\mathbf{x}$ near $\mathbf{a}$. Then

$$f(\mathbf{x}) = f(\mathbf{a}) + \frac{\mathbf{h}^T F \mathbf{h}}{2} + r_2(\mathbf{h}), \text{ where } \frac{r_2(\mathbf{h})}{\| \mathbf{h} \|^2} \to 0 \text{ as } \mathbf{h} \to \mathbf{0}.$$

The expression

$$\frac{\mathbf{h}^T F \mathbf{h}}{\| \mathbf{h} \|^2} \tag{13.1.4}$$

is the Rayleigh quotient associated to the quadratic form F, so its minimum value λ for $\mathbf{h} \neq \mathbf{0}$ is attained on the unit sphere U by Theorem 9.1.2. Since F is positive definite, this minimum value λ is strictly positive: by the Spectral Theorem it is the smallest eigenvalue. So Expression 13.1.4 is bounded below by λ for any $\mathbf{h} \neq \mathbf{0}$.

Thus for $\mathbf{h}$ sufficiently small this term dominates the remainder $r_2(\mathbf{h})$ which by Taylor's Theorem 12.4.8 goes to 0 as $\mathbf{h}$ goes to $\mathbf{0}$. In other words $(\mathbf{h}^T F\mathbf{h})/2 + r_2(\mathbf{h})$ remains positive. This means, precisely, that $f(\mathbf{x})$ has a strict local minimum at $\mathbf{a}$. □

Completely analogously, we have

13.1.5 Theorem. *Let $f(\mathbf{x})$ be a real-valued C^2 function defined on an open set U in $\mathbb{R}^n$. Assume that $\mathbf{a} \in U$ is a critical point of f and that the Hessian of f is negative-definite at* $\mathbf{a}$. *Then $f(\mathbf{x})$ has a strict local maximum at* $\mathbf{a}$.

13.2 Examples

13.2.1 Exercise. Let $f(x, y) = x^2 + y^2$. What is the signature of the Hessian F of f at the origin? What kind of form is the F? Conclusion?

Now do the same thing for $g(x, y) = x^4 + y^2$.

As we noted above, these theorems are multivariable analogs of the single-variable theorems stated in §3.3.

13.2.2 Example (The folium of Descartes)**.** In the plane let $f(x, y) = x^3 + y^3 - 3xy$. The curve C in the plane given by $f(x, y) = 0$, in other words, one of the level sets of $f(x, y)$ can be described by changing to polar coordinates $x = r\cos\theta$, $y = r\sin\theta$. Form the composite function $\varphi(r, \theta) = f(r\cos\theta, r\sin\theta)$. If you set it to 0, you get the solutions $r = 0$ and

$$r(\theta) = \frac{3\sin\theta\cos\theta}{\sin^3\theta + \cos^3\theta}.$$

For which values of θ is this not defined? What happens as θ approaches these values? The representation $r(\theta)$ is called a parametric representation of the curve C.

A different parametric equation for C is obtained by writing $y = tx$, which yields

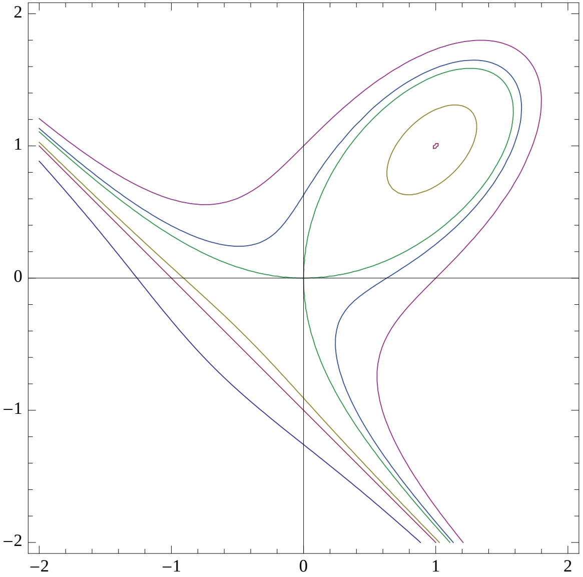
$$x = \frac{3t}{1+t^3}, \quad y = \frac{3t^2}{1+t^3}.$$

Thinking of t as time, this describes a trajectory on the level set. When t is nonnegative the trajectory describes a *folium*, meaning a leaf. See the graph below. What happens as $t \to \pm\infty$?

The partial derivatives of f are $3x^2 - 3y$ and $3y^2 - 3x$. Check that the only critical points are the origin and the point $(1, 1)$. The Hessian of f is

$$\begin{bmatrix} 6x & -3 \\ -3 & 6y \end{bmatrix}.$$

Thus at the point $(1, 1)$, the matrix is positive definite, guaranteeing a local minimum. At $(0, 0)$, the matrix is indefinite, with eigenvalues ± 3, so we have a saddle point. There is no global minimum or maximum: why?



Here is a graph showing the level curves of f, including C, the level curve at level 0. Make sure you can label each one of the level curves, including the leaf. Also see [28], IV.4.8, p. 324.

13.2.3 Example. Consider the positive definite matrix Q_n from Example 9.4.9 Suppose we want to solve the unconstrained optimization problem

$$\frac{1}{2}\mathbf{x}^T Q_n \mathbf{x} - \mathbf{b}^T \mathbf{x} + c = 0 \tag{13.2.4}$$

where $\mathbf{b}$ is a constant n-vector. To find the critical points, set the gradient to zero.

$$Q_n \mathbf{x} - \mathbf{b} = 0.$$

Since Q_n is invertible, multiplying by the inverse gives

$$\mathbf{x} = Q_n^{-1}\mathbf{b}.$$

So, using our explict formula (9.4.10) for Q_n^{-1}, we could solve this numerically for any choice of $\mathbf{b}$.

13.2.5 Exercise. Let $\mathbf{b} = (1, 0, \dots, 0)$ and work out the previous example numerically.

13.3 The Least Squares Method

Suppose you want to solve the system of linear equations

$$A\mathbf{x} = \mathbf{b}$$

where A is a $m \times n$ matrix with m larger - perhaps much larger - than n. So we have more equations than variables. The system is *overdetermined*. For simplicity we assume that A has maximal rank, which by hypothesis is n. Thus the columns $\mathbf{a}_j$ of A are linearly independent. and form a basis for the range $\mathcal{R}(A)$ of A, a subspace of dimension n of $\mathbb{R}^m$. If $\mathbf{b}$ is not in $\mathcal{R}(A)$, the system is *inconsistent*, which means it admits no solution. Instead we could ask for an approximate solution. For example, we look for the point $\mathbf{c}$ in $\mathcal{R}(A)$ that is closest to $\mathbf{b}$. We already solved this minimization problem in §7.5 using orthogonal projection and the Pythagorean Theorem. The $\mathbf{v}^i$ there are the $\mathbf{a}_j$ here. Here we want to give an explicit description of the solution in terms of A.

In our new notation, the problem is

13.3.1 Problem (Least Squares)**.** Assume $m > n$, Given a $m \times n$ matrix A of rank n, and a m-vector $\mathbf{b}$, the method of least squares finds the unique solution $\mathbf{x}$ of the minimization problem:

$$\text{Minimize the function } \|A\mathbf{x} - \mathbf{b}\|$$

Equivalently, we can minimize the square $f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|^2$ of this function. Writing it out, we see that

$$f(\mathbf{x}) = \langle A\mathbf{x} - \mathbf{b}, A\mathbf{x} - \mathbf{b}\rangle = \mathbf{x}^T A^T A\mathbf{x} - \mathbf{x}^T(2A^T\mathbf{b}) + \mathbf{b}^T\mathbf{b}. \qquad (13.3.2)$$

Thus $f(\mathbf{x})$ is a quadratic polynomial in $\mathbf{x}$.

This equation suggests we need to understand the $n \times n$ matrix $A^T A$.

13.3.3 Proposition. *Let A be a an $m \times n$ matrix, with $m \geq n$. If A has maximal rank n, then the $n \times n$ matrix $A^T A$ is positive definite.*

Proof. Because A has maximal rank, its nullspace is trivial, so the only n-vector $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{0}$ is the zero vector. So assume $\mathbf{x} \neq \mathbf{0}$. Then

$$\mathbf{x}^T (A^T A)\mathbf{x} = \|A\mathbf{x}\|^2 \geq 0$$

Now $\|A\mathbf{x}\|^2 = 0$ implies that $A\mathbf{x}$ is the zero vector, and this cannot be the case. Thus $\mathbf{x}^T (A^T A)\mathbf{x} > 0$ whenever $\mathbf{x} \neq \mathbf{0}$. Theorem 9.4.1 then says $A^T A$ is positive definite. □

Here is an example.

13.3.4 Exercise. Compute $A^T A$ for the rank 2 matrix

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \text{ to get } A^T A = \begin{bmatrix} 3 & 2 \\ 2 & 2 \end{bmatrix}.$$

Show this is positive definite. Let $\mathbf{b}$ be the vector $[0, 0, 1]$, Write the function $\|A\mathbf{x} - \mathbf{b}\|^2$ explicitly.

As we will see in §13.7, the level curves of f are ellipsoids.

Since $A^T A$ is positive definite, we find the unique minimum to Problem 13.3.1 by solving for its critical point: take its gradient using (13.3.2) and set it to zero:

$$2A^T A\mathbf{x} - 2A^T \mathbf{b} = \mathbf{0}.$$

Differentiating again, we see that the Hessian of f is the constant matrix $2A^T A$. As noted, this matrix is positive definite, and therefore invertible, so we can solve for $\mathbf{x}$ and conclude that our least squares problem has a unique minimizer at

$$\mathbf{x}^* = (A^T A)^{-1} A^T \mathbf{b}$$

Thus $A\mathbf{x}^*$ is the point in the range of A that minimizes the distance to $\mathbf{b}$. It can be written

$$A\mathbf{x}^* = A(A^T A)^{-1} A^T \mathbf{b}$$

The matrix $A(A^T A)^{-1} A^T$ is the key to understanding this situation. We now connect with Theorem 7.6.2 describing orthogonal projections.

13.3.5 Proposition. $A(A^T A)^{-1} A^T$ *is an orthogonal projection matrix.*

Proof. We use Theorem 7.6.2. First we show that $A(A^TA)^{-1}A^T$ is symmetric by computing its transpose:

$$\big(A(A^TA)^{-1}A^T\big)^T = A\big((A^TA)^{-1}\big)^TA^T = A\big((A^TA)^T\big)^{-1}A^T = A(A^TA)^{-1}A^T$$

Next we compute its square:

$$\big(A(A^TA)^{-1}A^T\big)^2 = A(A^TA)^{-1}A^TA(A^TA)^{-1}A^T = A(A^TA)^{-1}A^T$$

by cancelation of the matrix A^TA and its inverse in the middle of the expression, so we are done. □

Thus the transformation that takes an arbitrary $\mathbf{b} \in \mathbb{R}^m$ to

$$\mathbf{c} = (A^TA)^{-1}A^T\mathbf{b} \in \mathcal{R}(A)$$

is the orthogonal projection of $\mathbf{b}$ to $\mathcal{R}(A)$. As we proved in §7.5, $\mathbf{c}$ is the point at minimum distance on $\mathcal{R}(A)$.

This is a special case of the function D_S we studied in Exercise 11.1.5. When we study convex sets and convex functions, we will see that for any convex set C, the distance function D_C is a convex function: see Example 22.3.5. Now $\mathcal{R}(A)$ is a convex set, so that result applies. Indeed, linear functions are convex.

13.3.6 Exercise. Compute $A(A^TA)^{-1}A^T$ for the matrix A of 13.3.4 and verify that it is a projection matrix.

Do the same for matrix

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

Also confirm that A^TA is positive definite.

From yet another point of view, by Corollary 7.2.4 of the Four Subspaces theorem, we can write $\mathbf{b}$ uniquely as the orthogonal sum of an element $\mathbf{b}' \in \mathcal{R}(A)$ and an element $\mathbf{y}$ in the nullspace $\mathcal{N}(A^T)$:

$$\mathbf{b} = \mathbf{b}' + \mathbf{y}.$$

Now apply $A(A^TA)^{-1}A^T$ to this equation to get

$$A(A^TA)^{-1}A^T\mathbf{b} = A(A^TA)^{-1}A^T\mathbf{b}'.$$

Now $\mathbf{b}' = A\mathbf{c}$ for some n-vector $\mathbf{c}$. So

$$A(A^TA)^{-1}A^T\mathbf{b} = A(A^TA)^{-1}A^TA\mathbf{c} = A\mathbf{c}.$$

This is the desired expression of the projection of $\mathbf{b}$ as a linear combination of the columns of A.

13.4 The Cobb-Douglas Utility Function

Next we turn to an important example, which is actually a constrained optimization problem.

13.4.1 Example (Maximization of the Cobb-Douglas utility function)**.** This example is a 19th century favorite, found both in [58], §154, and [26], §137. In economics it is called the Cobb-Douglas function. Let the objective function be

$$f(\mathbf{x}) = x_0^{d_0} x_1^{d_1} x_2^{d_2} \dots x_n^{d_n}, \tag{13.4.2}$$

where the d_j are positive real numbers. We impose some constraints: all the x_j are non-negative, and they satisfy the linear condition

$$a = p_0 x_0 + p_1 x_1 + \dots + p_n x_n \tag{13.4.3}$$

for positive constants a and p_j. The goal is to find maximize f subject to these constraints.

In words, we write the positive number a as a weighted sum of $n+1$ non-negative numbers $x_0, x_1, \dots, x_n$, and we seek to maximize the product of the x_i raised to the d_i power, for fixed integers d_i.

Expressed this way. we have a *constrained* optimization problem, because the x_i are are non-negative and not independent variables: they satisfy (13.4.3). We can remove this constraint by solving for the variable x_0 in terms of the others. View this as a problem in n independent variables, substituting

$$x_0 = (a - p_1 x_1 - \dots - p_n x_n)/p_0$$

in 13.4.2. This is still a constrained problem, since we require $x_j \geq 0$, $1 \leq j \leq n$ and $a - p_1 x_1 - p_2 x_2 - \dots - p_n x_n \geq 0$: this last equation expresses the requirement that x_0 be non-negative. This is the feasible set: it is a n-simplex in $\mathbb{R}^n$. Its $n+1$ vertices are the origin and the n points with a/p_j in the j-th position, and 0 everywhere else. On the boundary of the simplex f takes the value 0, which is clearly the minimum value. So we know that the maximizer is an interior point of the simplex. We compute the partial derivatives of f considered as a function of x_1 through x_n alone, as explained above. Then we set the partials to 0. Because of the multiplicative nature of f, it is easier to write the "logarithmic derivative": in other words, instead of maximizing f, we maximize $\ln \mathbf{x}$. This makes sense since f takes on non-negative values in our feasible set, so $\ln$ is defined. The two problems are equivalent because $\ln$ is an increasing function. Now

$$\ln f = d_0 \ln x_0 + d_1 \ln x_1 + \dots d_n \ln x_n.$$

By the chain rule,

$$\frac{\partial \ln f}{\partial x_j} = \frac{d_j}{x_j} + \frac{d_0}{x_0}\frac{\partial x_0}{\partial x_j} = \frac{d_j}{x_j} - \frac{d_0}{x_0}\frac{p_j}{p_0}.$$

so at a maximizer $\mathbf{x}^*$, for $j \geq 1$ we have

$$\frac{d_j}{x_j} - \frac{d_0 p_j}{p_0 x_0} = 0.$$

Since we can let any one of the variables x_k play the role of x_0, all the partials vanish when

$$\frac{p_j x_j}{d_j} = \frac{p_k x_k}{d_k} \tag{13.4.4}$$

for all j and k between 0 and n. Set $D = \sum_{i=0}^{n} d_i$. Use (13.4.4) to eliminate all the x_j except for a fixed x_k from the constraint, getting

$$a = \frac{p_k x_k D}{d_k}, \text{ so } x_k = \frac{d_k a}{p_k D}.$$

Do this for each index k: we get a unique critical value $\mathbf{x}^*$ expressed in terms of the constants of the problem. Then

$$f(\mathbf{x}^*) = \big(\frac{a}{D}\big)^D \big(\frac{d_0}{p_0}\big)^{d_0} \dots \big(\frac{d_{n-1}}{p_{n-1}}\big)^{d_{n-1}} \big(\frac{d_n}{p_n}\big)^{d_n} > 0. \tag{13.4.5}$$

Since $\mathbf{x}^*$ is the unique interior point where all the partials vanish, and since f vanishes on the entire boundary of the region we are considering, this is both a local and the global maximum of the function. Thus we do not have to make the Hessian computation.

13.4.6 Exercise. Show that the Hessian F of f at $\mathbf{x}^*$ is negative definite.

Hint: It is easiest to compute F/f.

13.4.7 Example (Interpretation as a utility function)**.** In Example 13.4.1, think of the variables x_j as quantities of $n + 1$ commodities labeled by 0 to n, and p_j as the price of the j-th commodity. The quantity a represents the total income of a consumer, and f is the consumer's utility function. The consumer wants to maximize her utility subject to her income constraint. The computation above shows that there is a unique maximizer for the Cobb-Douglas utility function, and we have computed both the maximizer $\mathbf{x}^*$ and the maximum value $f(\mathbf{x}^*)$ of the utility function explicity. One often assumes that $D = 1$: for example in [56], p.19. $\mathbf{x}^*$ is called the vector of demands. If the consumer starts with quantities $\mathbf{x}^0$ of the commodities, then the difference vector $\mathbf{x}^* - \mathbf{x}^0$ is called the consumer's *excess demand*. The j-th coordinate of the excess demand is positive if the consumer wants to acquire more of the j-th commodity, and negative if the consumer wants to get rid of some of her holding of the j-th commodity.

13.5 The Ambiguous Case

Let f be a $\mathcal{C}^2$ function, and $\mathbf{p}$ an interior point of its domain where the gradient vanishes. By *ambiguous case*, we mean that the necessary condition for a minimum is satisfied at $\mathbf{p}$ (so the Hessian is positive semidefinite), but the sufficient condition is not satisfied (so it is not positive definite). Thus we cannot conclude that $\mathbf{p}$ is a minimizer. In the case of a function of a single variable, we proved an additional Theorem 3.3.6 that closes most of the distance between the two theorems. This is much harder even in the case of two variables, that we discuss now. The variables are called x and y.

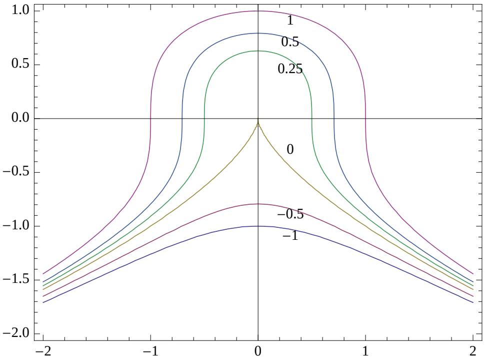
This issue created a long discussion in the 19-th century, because Lagrange thought that if one could prove that a function had a minimum at $\mathbf{x}^*$ when restricted to all lines passing through $\mathbf{x}^*$, it would have a minimum. This turned out to be false, as demonstrated by Peano in [26], Annotazioni N.133-136. Example 13.5.3 gives his key example. The issue of how to determine if there is an extremum when the Hessian is only semi-definite remains an active research area: see [3].

Let $z = f(x, y)$ be the graph of a $\mathcal{C}^2$ function of two variables in the neighborhood of the origin in $\mathbb{R}^2$. We assume that the function has a critical point at $\mathbf{0}$, so the partials vanish:

$$\frac{\partial f}{\partial x}(\mathbf{0}) = \frac{\partial f}{\partial y}(\mathbf{0}) = 0$$

By adjusting by a constant, we may assume that $f(\mathbf{0}) = 0$. We intersect the graph of the surface with its tangent plane $z = 0$ and analyze geometrically what can happen if the Hessian of f at the origin is positive semidefinite.

13.5.1 Example. $f(x, y) = x^2 + y^3$. The origin, a critical point, is neither a minimizer nor a maximizer. The intersection of the graph with the tangent plane $z = 0$ is a plane curve called a cuspidal cubic. We will use this curve later: see Definition 28.2.3 and Example 28.4.3. Here is a graph of the level curves near the origin.



13.5.2 Example. $f(x, y) = x^2 + y^4$ clearly has a minimum at the origin. The intersection of the graph with the tangent plane $z = 0$ is just one point.

13.5.1 Peano's Example

13.5.3 Example. This famous example is discussed in [26], Annotazioni N.133-136, [27], §57 and [30], §25. Let $f(x, y) = y^2 - 3x^2y + 2x^4$. Note that f factors as $(y - x^2)(y - 2x^2)$. The Hessian of f at the origin is positive semidefinite: indeed, it is the matrix

$$\begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \tag{13.5.4}$$

13.5.5 Lemma. *The origin is a local minimizer for f restricted to any line through the origin.*

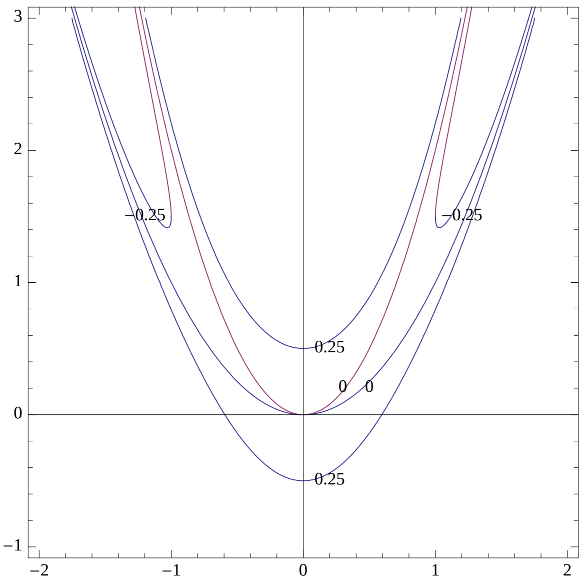
Proof. Indeed take the line $y = ax$, and use it to substitute y out of the equation of f. You get $a^2x^2 - 3ax^3 + 2x^4$. If $a \neq 0$, the restriction to the line is positive definite, since the quadratic term then has a positive coefficient. Therefore the origin is a strict local minimizer on that line. When $a = 0$, we are left with $2x^4$, so again the origin is a strict local minimizer. Finally we should check the line $x = 0$. The restriction of f to this line is y^2, so again the origin is a strict minimizer. □

This suggests that the origin is a local minimizer for f in a neighborhood of the origin. And yet

13.5.6 Lemma. *The origin is a maximizer for the restriction of f to the parabola $y = ax^2$, for any a between 1 and 2.*

Proof. Substituting out y, the restriction of f to the parabola is $x^4(a - 1)(a - 2)$. When $1 < a < 2$, the coefficient is negative, showing that there are points in the plane arbitrarily close to the origin where f takes on negative values, showing that the origin is not a minimizer. □

Here is a graph in the plane of the level curves near the origin: the level curves for level $-1/4$, 0 and $1/4$ are given and labeled.



Here is an analytic formula for a level curve with positive level ϵ^2:

$$(y - x^2)(y - 2x^2) - \epsilon^2 = 0$$

Solving as a quadratic polynomial in y we get

$$y = \frac{3x^2 \pm \sqrt{x^4 + 4\epsilon^2}}{2}$$

This defines two functions of x for all values of x. They intersect the y-axis at $\pm\epsilon$. On the graph, the level curves at level 0.25 are shown, so $\epsilon = 0.5$ so they intersect the y-axis at ± 0.5.

On the other hand if we consider a level set with negative level $-\epsilon^2$:

$$(y - x^2)(y - 2x^2) + \epsilon^2 = 0$$

Solving in the same way we get

$$y = \frac{3x^2 \pm \sqrt{x^4 - 4\epsilon^2}}{2}$$

which is only defined when $x^2 \geq 2\epsilon$. By symmetry, just consider the case $x \geq \sqrt{2\epsilon}$ and only look at the smaller of the two y values, which is

$$y = \frac{3x^2 - \sqrt{x^4 - 4\epsilon^2}}{2}$$

Thus when $x = \sqrt{2\epsilon}$, we get $y = 3\epsilon$, so in conclusion we get a point on the $-\epsilon^2$ level curve at $(\sqrt{2\epsilon}, 3\epsilon)$, which approaches the origin as $\epsilon \to 0$. This shows, in yet another way, that the origin is not a minimum. Notice that the point $(\sqrt{2\epsilon}, 3\epsilon)$ approaches the origin along the parabola $y = 3x^2/2$, one of the parabolas discussed earlier.

13.6 Gram Matrices

This section is connected to the *Gram-Schmidt* orthonormalization of a basis $\mathbf{a}_1$, $\mathbf{a}_2$, $\dots \mathbf{a}_n$ of $\mathbb{R}^n$. See [68], §4.4, or [60], §6.2. This material will be useful in understanding ellipsoids.

We do something slightly more general here, building on Proposition 9.3.2. Let R be a $n \times n$ matrix with columns $\mathbf{r}_1, \dots, \mathbf{r}_n$.

13.6.1 Definition. The *Gram matrix* of R, or of the n vectors $\mathbf{r}_i$, $1 \le i \le n$, is the symmetric matrix $A = R^T R$.

We do not assume that R is symmetric, so it is not the symmetric square root of A we studied in §9.3. Note that $a_{ij} = \langle \mathbf{r}_i, \mathbf{r}_j \rangle$, so we can take the square root of the diagonal elements:

$$\sigma_i = \sqrt{a_{ii}} = \|\mathbf{r}_i\|.$$

We also define

$$d_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\| = \sqrt{\sigma_i^2 + \sigma_j^2 - 2\langle \mathbf{r}_i, \mathbf{r}_j \rangle} = \sqrt{a_{ii} + a_{jj} - 2a_{ij}},$$

so that

$$a_{ij} = \frac{\sigma_i^2 + \sigma_j^2 - d_{ij}^2}{2}.$$

13.6.2 Definition. When $\mathbf{r}_i$ and $\mathbf{r}_j$ are non-zero, the *correlation coefficient* ρ_{ij} is

$$\rho_{ij} = \frac{\langle \mathbf{r}_i, \mathbf{r}_j \rangle}{\|\mathbf{r}_i\| \|\mathbf{r}_j\|}$$

If U is an orthogonal matrix, then UR has the same Gram matrix as R, as we saw in Proposition 9.3.3. An easy corollary is

13.6.3 Proposition. *A is the Gram matrix of a matrix R if and only if A is a (symmetric) positive semidefinite matrix.*

Proof. By Theorem 9.2.5, we can write the symmetric matrix A as $A = QDQ^T$, where D is diagonal and Q is orthogonal. So $R^TR = QDQ^T$. Multiply this equation on the left by Q^T and on the right by Q. Then

$$Q^TR^TRQ = D$$

using the defining property of an orthogonal matrix. But the left-hand side can be written $(RQ)^TRQ = D$, which shows that the diagonal entries of D are the dot products of the columns of RQ, so they are non-negative. □

13.6.4 Exercise. Find the symmetric square root of the 3×3 matrix in (8.5.11).

13.7 Ellipsoids and Level Sets

The simplest functions (other than linear functions) that we use as objective function is the quadratic

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^TA\mathbf{x} + \mathbf{b}^T\mathbf{x} \tag{13.7.1}$$

where A is an $n \times n$ symmetric matrix, and $\mathbf{b} \in \mathbb{R}^n$. Indeed this is what we do in §8.1. Because we are focusing on minimization, we generally assume that A is positive semidefinite, as per Proposition 8.1.9[1]. In this section we assume that A is positive definite and we examine the *level sets* $S_v = \{\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) = v\}$ of f.

The critical points of f are given by the solutions of the equation $A\mathbf{x} + \mathbf{b} = \mathbf{0}$. Since A is invertible, this has a unique solution $\mathbf{c} = -A^{-1}\mathbf{b}$. Because the Hessian of f is the positive definite A, Proposition 8.1.9 or its generalization Theorem 13.1.3 applies, so that f has a unique strict minimum at $\mathbf{c}$. Writing $\mathbf{b}$ in terms of $\mathbf{c}$, we get $\mathbf{b} = -A\mathbf{c}$. So we can write, completing the square:

$$f(\mathbf{x}) = \frac{1}{2}\big(\mathbf{x}^TA\mathbf{x} - 2\mathbf{c}^TA\mathbf{x}\big) = \frac{1}{2}\big((\mathbf{x}-\mathbf{c})^TA(\mathbf{x}-\mathbf{c}) - \mathbf{c}^TA\mathbf{c}\big) \tag{13.7.2}$$

We want to study the level sets S_v of f. Writing $\hat{v} = 2v + \mathbf{c}^TA\mathbf{c}$, this means looking at the solutions of

$$(\mathbf{x}-\mathbf{c})^TA(\mathbf{x}-\mathbf{c}) = \hat{v} \tag{13.7.3}$$

for fixed values of the constant $\hat{v}$.

Because A is positive definite, $(\mathbf{x}-\mathbf{c})^TA(\mathbf{x}-\mathbf{c})$ is non-negative, and it takes the value 0 only when $\mathbf{x} = \mathbf{c}$.

[1] We have already considered special cases of this: Examples 8.6.13 and 9.4.9.

13.7.4 Definition. If A is a positive definite $n \times n$ matrix, the set of $\mathbf{x} \in \mathbb{R}^n$ such that

$$(\mathbf{x} - \mathbf{c})^T A(\mathbf{x} - \mathbf{c}) = r^2 \tag{13.7.5}$$

is the *ellipsoid* $\mathcal{E}$ of center $\mathbf{c}$ based on A, with radius $r > 0$. We write it $\mathcal{E}(A, \mathbf{c}, r)$.

Thus the level sets of $f(\mathbf{x})$ are ellipsoids based on A with center $\mathbf{c}$ and varying radii r. In particular $\mathcal{E}(A, \mathbf{c}, r_1)$ and $\mathcal{E}(A, \mathbf{c}, r_2)$ do not meet if $r_1 \neq r_2$.

Warning. In the mathematics literature, the term ellipsoid denotes higher dimensional generalizations of the ellipse in the plane. See for example [63] p. 836 or [67], p. 335. This is the definition we use. In the recent optimization literature (see for instance [10],p.30), ellipsoid often refers to the region bounded by what we call an ellipsoid. Furthermore the matrix A is sometimes replaced by A^{-1}. Finally the radius r is often taken to be 1. Indeed, by rescaling the coordinate system by a factor of $1/r$ in each coordinate direction, it is clearly enough to consider the case $r = 1$.

13.7.6 Example. The standard way of writing the equation of an ellipsoid in $\mathbb{R}^3$ is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

We will see how to get this equation starting from a general positive definite matrix A.

Let R the symmetric square root of A as per Definition 9.3.1 .

13.7.7 Proposition. *The ellipsoid $\mathcal{E}(A, \mathbf{c}, r)$ can be written as the set of $\mathbf{x} \in \mathbb{R}^n$ of the form*

$$\mathbf{x} = \mathbf{c} + R^{-1}\mathbf{u} \quad \textit{for } \|\mathbf{u}\| = r. \tag{13.7.8}$$

Proof. To see the equivalence of these two representations, first change the coordinate system so the center of the ellipsoid is the origin. Then $\mathcal{E}(A, \mathbf{0}, r)$ is the set of $\mathbf{x}$ such that

$$\mathbf{x}^T A\mathbf{x} = \langle R\mathbf{x}, R\mathbf{x}\rangle = r^2 \tag{13.7.9}$$

So if we set $\mathbf{u} = R\mathbf{x}$, then $\|\mathbf{u}\| = r$, and $\mathbf{x} = R^{-1}\mathbf{u}$, so the two representations are equivalent. □

Since the set $\{\mathbf{u} \in \mathbb{R}^n \mid \|\mathbf{u}\| = r\}$ is just the sphere of radius r centered at the origin, we see that we get the ellipsoid $\mathcal{E}(A, \mathbf{c}, r)$ by distorting the sphere by the matrix R and then translating it by $\mathbf{c}$. To understand the distortion introduced by R, we need the eigenvectors and eigenvalues of A and R.

13.7.10 Definition. The directions of the eigenvectors of A are called the *principal axes* of the ellipsoid.

Recall from Definition 9.3.1 that the eigenvectors of A are the same as those of R. As usual we list the eigenvalues of A in increasing order: $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.

It is easiest to understand an ellipsoid in the basis of principal axes with as origin of the coordinate system the center of the ellipsoid. Remembering that the eigenvalues of R are $\sigma_i > 0$, with $\sigma_i^2 = \lambda_i$, we see that (13.7.5) can be written in the principal axes coordinate system as

$$\sum_{i=1}^{n} \sigma_i^2 x_i^2 = r^2 \tag{13.7.11}$$

Let $a_i = r/\sigma_i$. Then (13.7.11) becomes, after division by r^2:

$$\sum_{i=1}^{n} \frac{x_i^2}{a_i^2} = 1 \tag{13.7.12}$$

So we have generalized the equation of the ellipse and justified the name ellipsoid. The $\pm a_i$ are just the intercepts of the ellipse with the i-th coordinate axis, and we see how to write them in terms of the invariants of A and the level r of the level curve.

Then, going back to the original coordinate system and assuming that the σ_i are distinct, so strictly increasing, the two points on this ellipsoid that are closest to the center of the ellipsoid are $\mathbf{c} \pm \mathbf{e}_n/\sigma_n$, and two points that are furthest from the origin are $\mathbf{c} \pm \mathbf{e}_1/\sigma_1$. This follows immediately from Proposition 9.1.2.

13.7.13 Example. We start with the function

$$f() = \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x}$$

with

$$A = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}$$

as in Example 8.6.13 and $\mathbf{b} = (1, 2, 3)$.

We computed the eigenvalues and eigenvectors of A in Example 8.5.8. The eigenvectors are $(1/2, 1/2, 2)$ and the matrix Q of orthonormal eigenvectors is

$$Q = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{bmatrix}$$

Note that Q is not uniquely determined, because there is a repeated eigenvalue. Indeed, for the first two columns of Q you may pick any two mutually perpendicular vectors (x_1, x_2, x_3) of length 1 such that their coordinates add up to 1. You should check this by verifying that $Q^T AQ = D(1/2, 1/2, 2)$. In order to get the center $\mathbf{c}$, we need to compute the inverse of A. Since we have the eigenvalue and eigenvector matrices, we compute $A^{-1} = QD(2, 2, 1/2)Q^T$ and get

$$A^{-1} = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{3}{2} \end{bmatrix}$$

so

$$c = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{3}{2} \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 3 \end{bmatrix}$$

Here is the Mathematica code to graph a level surface, with the center of the ellipse at the origin, in the original coordinates:

```
ContourPlot3D[
{{x, y, z}.{{1, 1/2, 1/2}, {1/2, 1, 1/2}, {1/2, 1/2, 1}}.{x, y, z}
== 2},
{x, -2, 2}, {y, -2, 2}, {z, -2, 2}]
```

and here it is now in the principal axes coordinates. Notice that the vertical axis corresponds to the eigenvalue 2, while the other two axes correspond to eigenvalues 1/2.

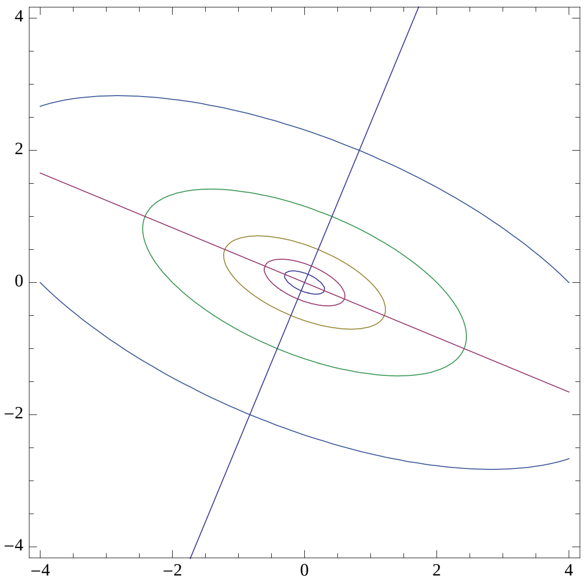
```
ContourPlot3D[
{{x, y, z}.{{1/2, 0, 0}, {0, 1/2, 0}, {0, 0, 2}}.{x, y, z}
== 2},
{x, -2, 2}, {y, -2, 2}, {z, -2, 2}]
```

13.7.14 Example. Now let's take the positive definite matrix

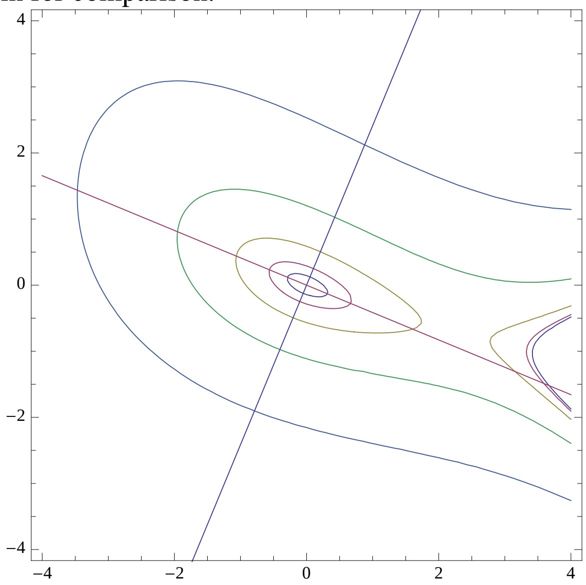
$$A = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}$$

The eigenvalues are $\lambda_1 = 2 - \sqrt{2}$ and $\lambda_2 = 2 + \sqrt{2}$, and the corresponding eigenvectors (not normalized to length 1) are $\mathbf{v}_1 = (-1 - \sqrt{2}, 1)$ and $\mathbf{v}_2 = (-1 + \sqrt{2}, 1)$. Thus the principal axes are $(-1-\sqrt{2})x+y = 0$ and $(-1+\sqrt{2})x+y = 0$.

Let us graph the level curves at levels $1/16$, $1/4$, 1, 4 and 16 to the function $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$, and the principal axes. We get:



Now let us perturb f near the origin by adding terms $-x^3/5 - y^3/5$. By Theorem 13.1.3 we know that there will still be a minimum at the origin, and that the level curves near the origin will not change much. Here is what we get with the same level curves, and with the principal axes of the previous quadratic shown again for comparison.



Notice the saddle point somewhere near the point $(2.5, 1)$.

Part V

Analysis

Lecture 14

Open and Closed Sets in Real Vector Spaces

This lecture starts our study of real analysis, which for us will mean the study of continuous or differentiable functions on the real vector space $\mathbb{R}^n$. In the multivariable calculus course you have taken, you have studied the case of the plane $\mathbb{R}^2$ and space $\mathbb{R}^3$, but perhaps not the general case that we will study here.

This chapter has two independent goals. The first is to establish, without proof, the completeness of the real numbers. We use this all the time: in fact we already needed it in the proof of the Mean Value Theorem 3.2.1. It is important that you understand it. The second goal is to define open and closed subsets of $\mathbb{R}^n$. This is something you know well already in the case of the real line (open and closed intervals), but becomes more interesting as the dimension gets larger.

This is the first step in the study of the *topology* of $\mathbb{R}^n$, namely the structure of its open and closed sets. In a later lecture, Theorem 16.1.4 relates open sets to continuous functions, which explains their importance to us and to mathematics more generally.

14.1 Ordered Fields

In this section we review the properties that the real numbers $\mathbb{R}$ share with the rational numbers $\mathbb{Q}$: each one forms is an *ordered field*.

To say that F is a *field* is just to formalize the law of arithmetic: F has two operations, addition, noted $+$, and multiplication, noted $\cdot$, satisfying the following properties:

1. Associativity: $(a+b)+c = a+(b+c)$ and $(a \cdot b) \cdot c = a \cdot (b \cdot c)$;

2. Commutativity: $a + b = b + a$ and $a \cdot b = b \cdot a$.

3. Existence of a neutral element 0 for addition: $a + 0 = a$, and a neutral element $1 \neq 0$ for multiplication: $a \cdot 1 = a$.

4. Existence of inverses: for all $a \in F$ there is an element b such that $a+b = 0$. b is then written $-a$. For all $a \neq 0$ in F, there is a c such that $a \cdot c = 1$. c is written $1/a$ or a^{-1}.

5. Multiplication distributes over addition: $a \cdot (b + c) = a \cdot b + a \cdot c$.

All this you know well. All the properties of arithmetic can be deduced from these rules.

14.1.1 Definition. To say that the field F is *ordered* means that any two elements a and b can be compared: $a > b$, such that

1. If $a > b$, then $a + c > b + c$, for all $c \in F$.

2. If $a > 0$ and $b > 0$ then $a \cdot b > 0$.

It is perhaps more surprising that all the familiar rules for manipulating inequalities come from these two plus the laws of arithmetic. If you are working out examples, you will need to use the fact that $a - b$ means $a + (-b)$, and that a/b means $a \cdot (1/b)$.

14.1.2 Example. We prove, using only the rules above:

1. $-a = (-1) \cdot a$.

   By definition $1 + (-1) = 0$. Multiply by a and use distributivity of multiplication over addition: $a + (-1) \cdot a = 0$. Add $-a$ to get the result.

2. $a > 0 \Leftrightarrow -a < 0$.

   By the first inequality property, we may add $-a$ to both sides of $a > 0$ and keep the inequality, so $a + (-a) > -a$. By definition of $-a$ we get $0 > -a$, which is what we want.

3. $a > b \Leftrightarrow -a < -b$.

   First add $-a$ to $a > b$: we get $0 > b - a$ by the first inequality property, and using $a - a = 0$. Now add $-b$ to this, getting $-b > -a$, as required. This time you use $b - b = 0$.

4. $a > b$ and $c > d \Rightarrow a + c > b + d$.

   Use the first inequality property twice.

5. $1 > 0$.

   Proof by contradiction. Since $0 \neq 1$, we would have $0 > 1$. We show this leads to a contradiction. By a previous result, this would imply $-1 > 0$. Take a positive a: $a > 0$. By the second inequality result, letting $b = -1$, and a previous item, $-a > 0$ is positive. But then $a + (-a) > 0$, but this is absurd, since the left-hand side is 0.

14.1.3 Exercise. Prove the following implications from the rules above:

1. $a > b$ and $c < 0 \Rightarrow ac < bc$;

2. $a > 0 \Leftrightarrow 1/a > 0$;

3. $a > b > 0 \Leftrightarrow 0 < 1/a < 1/b$.

4. $a \neq 0 \Rightarrow a \cdot a > 0$.

5. $0 < 1$.

To conclude, note that $\geq$ is an example of a binary relation (on $\mathbb{Q}$ or $\mathbb{R}$), discussed in §2.2. It is

- reflexive: $a \geq a$;

- antisymmetric: $a \geq b$ and $b \geq a$ implies $a = b$;

- complete: for any two elements a and b, we have either $a \geq b$ or $b \geq a$;

- transitive: $a \geq b$ and $b \geq c$ implies $a \geq c$.

14.2 The Completeness of the Real Numbers

Next we turn to an important property of real numbers $\mathbb{R}$: the *completeness* of $\mathbb{R}$ that $\mathbb{Q}$ does not share. Formalizing this concept serves as a preliminary to our study of $\mathbb{R}^n$.

14.2.1 Definition. Let S be a set of real numbers.

1. A real number u is called an *upper bound* for S if for all $s \in S$, $s \leq u$. If S admits an upper bound, then it is said to be *bounded above*.

2. A real number l is called a *lower bound* for S if for all $s \in S$, $l \leq s$. If S admits a lower bound, then it is said to be *bounded below*.

3. If S admits both an upper and lower bound, we say simply that it is *bounded*.

14.2.2 Example. The set $\mathbb{R}$ is neither bounded above nor bounded below. The set of positive reals is bounded below, while the set of negative reals is bounded above. The interval $[0, 1]$ is bounded both above and below—or just bounded.

We generalize our notion of boundedness to $\mathbb{R}^n$ in §15.1.

14.2.3 Definition. Again let S be a set of real numbers.

1. Assume S is bounded above. The real number u_0 is called the *least upper bound* or *lub* for S, if u_0 is an upper bound for S and $u_0 \leq u$ for any upper bound u of S.

2. Assume S is bounded below. The real number l_0 is called the *greatest lower bound* or *glb* for S, if l_0 is an lower bound for S and $l \leq l_0$ for any lower bound l of S.

14.2.4 Remark. It is convenient to extend this definition to unbounded sets S: we let the *supremum* or *sup* of S be ∞ if S is not bounded above, and the $lub(S)$ otherwise. Similarly, we let the *infimum* or *inf* of S be $-\infty$ if S is not bounded below, and the $glb(S)$ otherwise.

The needed property of $\mathbb{R}$, the *completeness* of $\mathbb{R}$, is the assertion that the following statement holds.

14.2.5 Theorem. *Every set $S \subset \mathbb{R}$ that is bounded above has a least upper bound, and every set $S \subset \mathbb{R}$ that is bounded below has a greatest lower bound.*

A proof of this difficult result can be found in [55], ch. 1 or [70], §2.3. For a historical account, see [28] §III.1.

Here are two corollaries that show its power.

14.2.6 Theorem (The Archimedean Property)**.** *If x and y are in $\mathbb{R}$, with $x > 0$, then there is a positive integer n such that $nx > y$.*

Equivalently, if $x > 0$ in $\mathbb{R}$, then there is a positive integer n such that $x > 1/n$.

Proof. The result is obvious if $y \leq 0$, so we assume $y > 0$. Proof by contradiction. Given a positive x, let S be the set of nx, for all positive integers n. If the result is false, then for all positive integers n, $nx \leq y$, so that y is an upper bound for S. But then by Theorem 14.2.5, S has a least upper bound that we call M. Since $x > 0$ by hypothesis, $M - x < M$, and so $M - x$ is not an upper bound, meaning that there is an integer m such that $M - x < mx$. Thus $M < (m + 1)x \in S$, which is a contradiction since M is an upper bound and $(m + 1)x$ is in S.

To get the equivalent version just let y be 1. To go back from the equivalent version, replace x by x/y. □

14.2.7 Exercise. Let (a, b) be a open interval in $\mathbb{R}$, with $b-a > 1$. Write a careful proof that (a, b) contains an integer.

Hint: Let m be the largest integer less than or equal to a, and show that $m + 1$ is in (a, b).

14.2.8 Theorem (Density of the Rationals)**.** *If x and y are in $\mathbb{R}$, $x < y$, then there is a rational number between them: $\exists p \in \mathbb{Q}$ such that $x < p < y$.*

Proof. Let $z = y - x$. By the archimedean property we can find an integer n so that $nz > 1$. So there is an integer m strictly between nx and ny:

$$nx < m < ny.$$

Divide this equation by n to see that the rational number $p = m/n$ meets the requirement. □

This theorem says that between two real numbers, no matter how close together, there is a rational number. We say that the rational numbers are *dense*.

14.2.9 Exercise. With the notation of Theorem 14.2.8, show that there are an infinite number of rationals in the open interval (x, y).

14.3 Cells and Balls

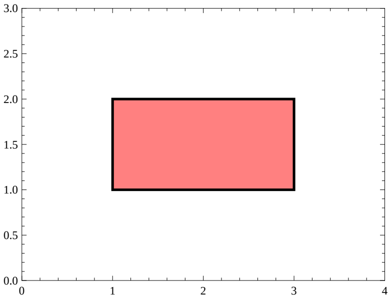
Now we move from $\mathbb{R}$ to $\mathbb{R}^n$.

In preparation for defining open and closed sets, we define two important collections of subsets of $\mathbb{R}^n$: cells and balls.

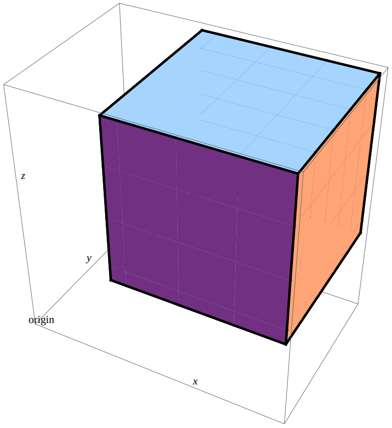
14.3.1 Definition (Cells)**.**

The generalization of an interval in $\mathbb{R}$ to a subset of $\mathbb{R}^n$ is called an n-cell. n-cells are defined by n pairs of real numbers $(\alpha_1, \beta_1), \dots (\alpha_n, \beta_n)$ with $\alpha_i \leq \beta_i$ for all i, $1 \leq i \leq n$. The n-cell delimited by (α_i, β_i), $1 \leq i \leq n$, is the collection of points $\mathbf{x} \in \mathbb{R}^n$ such that $\alpha_i \leq x_i \leq \beta_i$ for all i.

When $n = 1$ (a 1-cell), we have a closed interval $[\alpha_1, \beta_1]$. When $n = 2$ (a 2-cell), we have a closed rectangle, with sides parallel to the coordinate axes. When $n = 3$ (a 3-cell), we have a closed rectangular box, with sides parallel to the coordinate axes.



This figure depicts a 3-cell (a rectangular prism) with (α, β) pairs $(0.5, 1.5)$, $(0, 1)$, $(0.5, 1.5)$:



14.3.2 Definition (Balls)**.** For each point $\mathbf{p}$ in $\mathbb{R}^n$, called the *center*, and each strictly positive real number r, called the *radius*, we define the *open ball* $N_r(\mathbf{p})$ as the set of points $\mathbf{y}$ at distance less than r from $\mathbf{p}$: $N_r(\mathbf{p}) = \{\mathbf{y}|d(\mathbf{p}, \mathbf{y}) < r\}$.

The *closed ball* $\overline{N}_r(\mathbf{p})$ is $\overline{N}_r(\mathbf{p}) = \{\mathbf{y}|d(\mathbf{p}, \mathbf{y}) \leq r\}$. Thus it is the union of the open ball $N_r(\mathbf{p})$ as well as the $n-1$-sphere of radius r centered at $\mathbf{p}$, defined next.

14.3.3 Definition (Spheres)**.** The $(n-1)$-*sphere* in $\mathbb{R}^n$ centered at $\mathbf{p}$ and of radius r is the set of points $\mathbf{y}$ at distance exactly r from $\mathbf{p}$:

$$S_r(\mathbf{p}) = \{y \in \mathbb{R}^n \mid d(\mathbf{p}, \mathbf{y}) = r\}.$$

14.3.4 Example. When $n = 1$, a ball is just an interval of length $2r$ centered at p, with 'closed' and 'open' meaning the same thing for balls and intervals.

A 2-ball is a disk of radius r centered at $\mathbf{p}$, with the closed disk including the circle of radius r. In $\mathbb{R}^3$, a ball is just an ordinary ball of radius r centered at $\mathbf{p}$, with the closed ball including the ordinary sphere of radius r centered at $\mathbf{p}$.

In higher dimensions, we continue to use the words ball and sphere.

Our definition of balls serves as the basis for our definitions of neighborhoods and open sets.

14.4 Open and Closed Sets

14.4.1 Definition.

1. The *r-neighborhood*, $r > 0$, of a point $\mathbf{p}$ is the open ball of radius r centered at $\mathbf{p}$: formally, $N_r(\mathbf{p})$.
2. A set S in $\mathbb{R}^n$ is *open* if for every $\mathbf{x} \in S$ there is a radius $r > 0$ with $N_r(\mathbf{x}) \subset S$. That is, for any point in an open set, there exists an open ball with positive radius that is wholly contained within the set. The empty set, denoted $\emptyset$, is open by convention.

Note that we can write the definition that S is open in $\mathbb{R}^n$, only using the distance function $d(\mathbf{x}, \mathbf{y})$ on $\mathbb{R}^n$ discussed in §5.4:

S is open in $\mathbb{R}^n$ if for every $\mathbf{x} \in S$ there is a real $r > 0$ such that every $\mathbf{y} \in \mathbb{R}^n$ with $d(\mathbf{x}, \mathbf{y}) < r$ is in S.

Indeed, if $d(\mathbf{x}, \mathbf{y})$ is the Euclidean distance function, this is the same definition.

14.4.2 Exercise. Prove that any open ball $N_r(\mathbf{p})$ is an open set.

Hint: Prove that any point $\mathbf{x}$ in $N_r(\mathbf{p})$ has a neighborhood contained in the ball. Draw the picture in $\mathbb{R}^2$ to see what you need to do. You will need the triangle inequality (Definition 5.4.1).

14.4.3 Definition. A point $\mathbf{p}$ is a *limit point* of a set S if for every $\epsilon > 0$, the neighborhood $N_\epsilon(\mathbf{p})$ of $\mathbf{p}$ contains a point of S other than $\mathbf{p}$. Note that $\mathbf{p}$ can be a limit point of S without being contained in S, as the following examples show.

14.4.4 Example. Limit points of open balls.

1. The points 1 and -1 are limit points of the interval $(-1, 1)$, as are all the points in $(-1, 1)$. The set of limit points of $(-1, 1)$ is the closed interval $[-1, 1]$.
2. In $\mathbb{R}^n$, the limit points of $N_r(\mathbf{p})$ are $\overline{N}_r(\mathbf{p})$.

14.4.5 Exercise. Take the set of points in $\mathbb{R}$ of the form $\frac{1}{n}$ for all positive integers n. What are the limit points of this set?

14.4.6 Exercise. Prove that if $\mathbf{p}$ is a limit point of S, every neighborhood of $\mathbf{p}$ contains an infinite number of points in S.

Hint: Prove this by contradiction.

14.4.7 Definition. A set S is *closed* if it contains all its limit points. The empty set is closed by convention.

14.4.8 Exercise. Prove that $\overline{N}_r(\mathbf{p})$ is a closed set.

We will use the following definitions in later lectures:

14.4.9 Definition. A point $\mathbf{p}$ is an *interior point* of a set S if there is a neighborhood $N_r(\mathbf{p})$ of $\mathbf{p}$ contained in S. The set of all interior points of S is called the *interior* of S, written $intS$.

Note that this implies that all the points in an open set S are interior points of S.

14.4.10 Definition. The closure $\bar{S}$ of S is the union of S and all its limit points. It is a closed set.

14.4.11 Definition. A point $\mathbf{p}$ is a *boundary point* of S if $\mathbf{p} \in \bar{S}$ and $\mathbf{p}$ is not in the interior of S. The set of all boundary points of S is called the *boundary*.

Note that a boundary point $\mathbf{p}$ can belong to S or not, depending on S. A closed set contains all its boundary points. An open set has no boundary points. The boundary of the closed n-ball $\overline{N}_r(\mathbf{p})$ is the $(n-1)$-sphere $S_r(\mathbf{p})$.

14.4.12 Definition. The *complement* of a set S in $\mathbb{R}^n$, denoted S^c, is the set of points $\mathbf{x}$ in $\mathbb{R}^n$ that are not in S.

Note that the complement of a complement is the original set: $(S^c)^c = S$.

14.4.13 Theorem. *The complement of an open set is closed, and the complement of a closed set is open.*

Proof. We begin by proving that if a set $S \subset \mathbb{R}^n$ is open, its complement S^c is closed. We pick an arbitrary point $\mathbf{x}$ outside of S^c, which by definition is in S. Definition 14.4.7 provides that a closed set contains all its limit points, so we need to prove that $\mathbf{x}$ is not a limit point of S^c. Since $\mathbf{x}$ is in S and S is an open set, Definition 14.4.1 tells us that there exists a radius $r > 0$ such that $N_r(\mathbf{x})$ is contained in S. Accordingly, no point in $N_r(\mathbf{x})$ is in the complement S^c. Because $\mathbf{x}$ is a positive distance r away from S^c, $\mathbf{x}$ is not a limit point of S^c. So S^c is closed.

We now assume that S is closed, proving that its complement S^c is open. Pick any point $\mathbf{x}$ in the complement S^c, meaning any point that is not in S. S is closed, so by Definition 14.4.7, $\mathbf{x}$ cannot be a limit point of S. We know, then, that there exists a neighborhood of $\mathbf{x}$ in S^c. By Definition 14.4.1, S^c is open. □

14.4.14 Exercise. Show that a non-empty set S in $\mathbb{R}^n$, apart from $\mathbb{R}^n$ itself, cannot be both open and closed.

Hint: Prove this by contradiction, so assume S is both open and closed. Since S is non-empty, it contains a point $\mathbf{s}$. Since its complement S^c is non-empty, it contains a point $\mathbf{p}$. Consider the line segment L joining $\mathbf{s}$ and $\mathbf{p}$. For $\mathbf{x} \in L$, consider the distance function $d(\mathbf{x}, \mathbf{s})$, and let $m = \operatorname{glb}_{\mathbf{x} \in S^c} d(\mathbf{x}, s)$ giving the greatest lower bound of the distance of a point $\mathbf{x} \in L \cap S^c$ to the point $\mathbf{s} \in S$. Let $\mathbf{q}$ be the point realizing this distance. Discuss what happens if $\mathbf{q} \in S$, or not.

14.4.15 Theorem.

1. *Any union of open sets is open, and a finite intersection of open sets is open.*
2. *Any intersection of closed sets is closed, and a finite union of closed sets is closed.*

Note the difference between taking a finite union, meaning a union over a finite number of sets, and "any union," meaning a union over a possibly infinite number of sets. If we write the union as $\cup_{i \in I} U_i$, where the U_i are the open sets, that the set I indexing the opens is called the *indexing set*. A typical and very large index set might be all the points in the set, for example the unit disk.

Proof. We start with an arbitrary union $\cup_{i \in I} U_i$ of open sets. Pick a point $\mathbf{p}$ in the union: we must show it is an interior point of the union. Now $\mathbf{p}$ belongs to one of the open sets in the union, say U_0. So by definition a neighborhood N of $\mathbf{p}$ is in U_0, and therefore in the union. We are done.

Now we show that a finite intersection $\cap_{i=1}^{n} U_i$ of open sets is open. Pick a point $\mathbf{p}$ in the intersection: so $\mathbf{p}$ is in each U_i. Since U_i is open, there is a open ball $N_{r_i}(\mathbf{p})$ of radius $r_i > 0$ around $\mathbf{p}$ in U_i. Let r be the smallest of the r_i. Then the ball $N_r(\mathbf{p})$ is in all the U_i, and therefore in the intersection. Note that this argument would break down if we had an infinite number of sets. Then we would have to take for r the infimum of the r_i: even though all the r_i are positive, their infimum might be 0.

We could prove the results for closed sets in the same way. Or we could get them from Theorem 14.4.13, using DeMorgan's Law (see §2.1.4): for an arbitrary collection I of sets C_i, $i \in I$, we have

$$(\bigcup_{i \in I} C_i)^c = \bigcap_{i \in I} C_i^c \tag{14.4.16}$$

□

14.4.17 Exercise. Prove the equality in (14.4.16) and fill in the details in the second part of the proof.

14.4.18 Exercise.

1. For any positive integer m, let G_m be the open set $G_m = N_{\frac{1}{m}}(\mathbf{0})$ in $\mathbb{R}^n$. Thus, the indexing set is infinite. Determine both the union U and the intersection I of all the G_m. What is U? What is I? Check that Theorem 14.4.15 is satisfied.

2. For any positive integer m, let F_m be the closed set $F_m = \overline{N}_{\frac{1}{m}}(\mathbf{0})$ in $\mathbb{R}^n$. Determine both the union $\overline{U}$ and the intersection $\overline{I}$ of all the F_m. What is $\overline{U}$? What is $\overline{I}$? Check that Theorem 14.4.15 is satisfied.

Hint: Try this in $\mathbb{R}$ first.

14.5 The Nested Interval Theorem

For use in the next lecture (see the proof of Theorem 15.1.6), we prove a result about intervals in $\mathbb{R}$ that depends on the existence of least upper bounds and greatest lower bounds discussed in §14.2. We also use the definition of infinite sequence, discussed in the next lecture.

14.5.1 Theorem. *Let I_k, $1 \leq k < \infty$. be a nested sequence of closed intervals in $\mathbb{R}$, meaning that $I_{k+1} \subset I_k$ for all $k \geq 1$. The intersection of the I_k is non-empty. Furthermore, if the lengths ℓ_k of the intervals I_k converge to 0 as k approaches infinity, the intersection is a single point.*

Proof. Let $I_k = [a_k, b_k]$ such that $a_k \leq b_k$. The hypothesis that the I_k are nested means that

$$a_1 \leq a_2 \leq \cdots \leq a_k \leq a_{k+1} \leq \cdots \leq b_{k+1} \leq b_k \leq \cdots \leq b_2 \leq b_1$$

This expression demonstrates that the collection of a_k is bounded above (by the b_k) and therefore has a least upper bound a. By the definition of the least upper bound, $a_k \leq a$ for all k; the inequalities above, meanwhile, confirm that $a \leq b_k$ for all k. a therefore belongs to the interval I_k for all k. By the same reasoning, the collection of b_k is bounded below (by the a_k) and therefore has a greatest lower bound b. As before, $a_k \leq b \leq b_k$ for all k, and we may conclude that b belongs to the interval I_k for all k.

With the a and b terms defined, we can flesh out the train of inequalities:

$$a_1 \leq a_2 \leq \dots a_k \leq a_{k+1} \leq \cdots \leq a \leq b \leq \cdots \leq b_{k+1} \leq b_k \cdots \leq b_2 \leq b_1$$

This expression shows that the interval $[a, b]$ belongs to all the I_k, proving that the intersection of all I_k is non-empty.

The length ℓ_k of I_k is $b_k - a_k$. Note that $\{\ell_k\}$ is a decreasing sequence, so it converges by Theorem 10.2.4.

In the second part of the theorem,we assume the sequence $\{\ell_k\}$ converges to zero. Thus the a_k and b_k converge to the same value c. Since a and b are between a_k and b_k, they must be equal to c, which is the unique point of intersection. □

Now we generalize our theorem to $\mathbb{R}^n$. We only deal with the special case we need later.

14.5.2 Definition. A *hypercube* in $\mathbb{R}^n$ is an n-cell delimited by $[\alpha_i, \beta_i]$, $1 \leq i \leq n$, where $\beta_i - \alpha_i$ are all equal to the same value ℓ called the length of the side of the hypercube.

The hypercube is the n-dimensional generalization of the line segment in $\mathbb{R}$, the square in $\mathbb{R}^2$ and the cube in $\mathbb{R}^3$.

14.5.3 Exercise. Show that the maximum distance between two points in the hypercube C in $\mathbb{R}^n$ of side length ℓ is $\ell\sqrt{n}$.

14.5.4 Theorem. *Let C_k, $0 \leq k < \infty$ be a nested sequence of closed hypercubes in $\mathbb{R}^n$, meaning that $C_{k+1} \subset C_k$ for all $k \geq 0$. The intersection of the C_k is non-empty. Furthermore, if the lengths ℓ_k of the side of C_k converge to 0 as k approaches infinity, the intersection is a single point* $\mathbf{c}$.

Proof. By hypothesis $C_k = \{\mathbf{x} \in \mathbb{R}^n | \alpha_{k,i} \leq x_i \leq \beta_{k,i}\}$, and the nesting means that $\alpha_{k,i} \leq \alpha_{k+1,i}$ and $\beta_{k+1,i} \leq \beta_{k,i}$ for all k. Thus in each coordinate direction we have a nested sequence of intervals, so by Theorem 14.5.1 in the limit we get a non-empty interval, so we have a non-empty hypercube. If the lengths ℓ_k converge to 0, in the limit we have a hypercube with side length 0, and that is the desired point $\mathbf{c}$. □

14.5.5 Exercise. As with Theorem 14.5.1, consider intervals $I_k = [a_k, b_k]$, $k \in \mathbb{N}$. Start with a segment of length 1, $a_1 = -1/2$ and $b_1 = 1/2$, so $I_1 = [-1/2, 1/2]$. Define the other intervals recursively by the following probabilistic rule:

1. With probability p, $a_{k+1} = a_k$ and $b_{k+1} = \frac{a_k+b_k}{2}$.
2. With probability $1-p$, $a_{k+1} = \frac{a_k+b_k}{2}$ and $b_{k+1} = b_k$.

Note that I_k is a set of nested intervals, regardless of p. Show that the length of the intervals goes to 0 as k goes to ∞. Thus in the limit there is only one point q nested in all the intervals

Now some extra credit for those who know some probability theory. Assume that $p = \frac{1}{2}$. Describe the probabilities of various outcomes of where q ends up. You should first try doing this in the case you only perform n nestings, so that you are left with an interval I_n of length 2^{-n}.

14.6 Relatively Open Sets

We start out with a simple example showing the potential pitfalls of the definition of open set. Let S be an open set in $\mathbb{R}^2$. View $\mathbb{R}^2$ as a subspace of $\mathbb{R}^3$: for example, make $\mathbb{R}^2$ the xy coordinate plane. Is S open in $\mathbb{R}^3$? No. In fact its interior is empty. To clarify the situation, we define relatively open sets, using the distance function $d(\mathbf{x}, \mathbf{y})$ from §5.4.

14.6.1 Definition. S is *open relative to* Y if for all $\mathbf{x} \in S$ there is a real $r > 0$ such that every $\mathbf{y} \in Y$ with $d(\mathbf{x}, \mathbf{y}) < r$ is in S.

14.6.2 Definition. Assume S is an arbitrary set in Y. We can define the *relative interior* of S in Y: it is the set of all $\mathbf{x} \in S$ such that there is a real $r > 0$ such that every $\mathbf{y} \in Y$ with $d(\mathbf{x}, \mathbf{y}) < r$ is in S. We write the relative interior as $\operatorname{relint}_Y S$.

14.6.3 Example. Let S be the closed unit disk in a plane Y. If we put Y in $\mathbb{R}^3$, and view S as a set there, its interior is empty. On the other hand, as a set in Y, the interior of S is the open disk U. So U is the relative interior of S in $\mathbb{R}^3$.

The next theorem tells us exactly how to determine what sets are relatively open.

14.6.4 Theorem. *A set S in $\mathbb{R}^n$ that is contained in a subset Y of $\mathbb{R}^n$ is open relative to Y if and only if S is the intersection with Y of an open set U in $\mathbb{R}^n$.*

Proof. First we assume that S is open relative to Y. We must construct an open set U in $\mathbb{R}^n$ such that $U \cap Y = S$. By Definition 14.6.1, for each $p \in Y$ there is a $r_p > 0$ such that the set $S_p = \{\mathbf{y} \in Y \mid d(\mathbf{p}, \mathbf{y}) < r_p\}$ is in S. Then take for U the union of all the r_p neighborhoods of p in $\mathbb{R}^n$. As we saw in Theorem 14.4.15, this union is an open set: by construction its intersection with Y is S, which is what we need.

In the other direction, assume that $S = Y \cap U$ for some open set $U \subset \mathbb{R}^n$. Now pick any point $\mathbf{p} \in S$. Then $\mathbf{p}$ is in U. Since U is open, there is a $r > 0$ such that $U_p = \{\mathbf{x} \in \mathbb{R}^n \mid d(\mathbf{p}, \mathbf{x}) < r\}$ is contained in U. Since $U \cap Y = S$, $U_p \cap Y \subset S$, which says that S is open relative to Y. □

Here is an application of linear algebra. As usual $S \subset Y \subset \mathbb{R}^n$. For convenience we assume that the origin of $\mathbb{R}^n$ is in S. Let Y be the linear subspace of smallest dimension in $\mathbb{R}^n$ containing S.[1] Let m be the dimension of Y, so $Y = \mathbb{R}^m$ for some m, $m \leq n$. Then if $m < n$, the interior of S in $\mathbb{R}^n$ is empty, but it need not be empty in Y. This is what happens in Example 14.6.3.

[1] This is a concept we will study later on: the affine hull of S: see Definition 18.2.4 and Corollary 18.4.5 where the relative interior is used.

Lecture 15

Compact Sets

In this lecture we study compact sets, which, in finite dimensional vector spaces, are simply the sets that are both closed and bounded. Compact sets form a powerful generalization of finite sets, and many of the desirable properties of finite sets remain true for compact sets. Compact sets are an essential ingredient for optimization theory, so this chapter (often viewed by students as the most difficult in this course) needs to be studied very carefully.

The lecture starts with the key section of the chapter: the definition and and central properties of compactness in a finite dimensional vector space

Then, after a review of infinite sequences in Chapter 10.1 for those who need it, we introduce Cauchy sequences, a useful tool for discussing convergence of sequences, thanks to Theorem 15.2.2. We use standard results about the convergence of sequences proved §10.2. to connect issues of convergence of sequences to the metric space structure of $\mathbb{R}^n$ studied in §5.1. The second main result of the chapter is the the notion of sequential compactness. Theorem 15.4.2 shows that it is equivalent to compactness. We will use this result in the next lecture.

The lecture concludes with an optional section 15.5 showing how Cauchy sequence allows the construction of the real numbers starting from the rational numbers. This gives some additional insight on the completeness of $\mathbb{R}$, studied in §14.2

15.1 Compact Sets

Compact sets play a key role in optimization. The definition given here for compact sets is not the most general one (see Rudin [55], Definition 2.32 or Strichartz [70], §9.2.4), but it works in $\mathbb{R}^n$, where it is called the Heine-Borel Theorem. In the Exercises 15.1.10 and 15.1.11 the equivalence of the definition given here and the more general one is established.

15.1.1 Definition. A set S in $\mathbb{R}^n$ is *compact* if it is closed and bounded.

We just learned about closed sets. We now define boundedness.

15.1.2 Definition. A set S in $\mathbb{R}^n$ is *bounded* if it is contained in a ball $N_r(\mathbf{0})$ of some finite radius r.

The key examples of compact sets are the ones given in Definitions 14.3.1 and 14.3.2: Cells are compact, and closed balls are compact. Open balls are *not* compact because they are not closed.

The next theorem will be proved using a variant of the *pigeon hole principle*:

15.1.3 Proposition. *If an infinite number of objects are fit into a finite number of slots, then at least one of the slots will contain an infinite subset of the objects.*

Proof. Proof by contradiction: assume that the objects have been put into a finite number n of slots, and assume that in each slot there is only a finite number of elements. Let m be the maximum number of elements in a given slot (note that since there are only a finite number of slots, it makes sense to talk about the maximum). Then clearly the total number of elements is bounded by nm, a finite number. So we have a contradiction. □

15.1.4 Exercise. Let the infinite set be the integers. Devise a method to put them in m slots, for any positive integer m, such that each slot contains an infinite number of integers. Now do it so that only one slot contains an infinite number of integers.

15.1.5 Theorem. *Every infinite subset of a compact set S in $\mathbb{R}^n$ has a limit point in S.*

We will also need a slight variant of this theorem.

15.1.6 Theorem. *Every infinite subset of a bounded set S in $\mathbb{R}^n$ has a limit point in $\mathbb{R}^n$.*

Proof. We prove both theorems simultaneously. Let I denote an infinite subset of a bounded set S. Since S is bounded, it is contained in an n-cell C_0, which we might as well assume is a hypercube with sides of length ℓ. We divide C_0 into smaller cells by slicing it in half in each dimension. So now we have 2^n smaller n-cells, each a hypercube (see Theorem 14.5.4) with side length $\ell/2$. The union of these smaller cells is C_0. There is overlap on the boundary of each one of the closed hypercubes, but that doesn't matter: we only care that the union of the smaller cells is equal to C_0.

Since $I \subset S$ has an infinite number of elements, the pigeon hole principle (Proposition 15.1.3) tells us that there must be an infinite number of elements of I

in at least one of the smaller n-cells. Pick one such smaller cell—call it C_1—and repeat the same subdivision process on it. Of the $(2^n)^n$ resulting n-cells, select one of them—$C_2 \subset C_1$—that also contains an infinite number of elements of I. The hypercube C_2 has sides of length $\ell/2^2$. Continuing in this way, for any positive integer k we can find a hypercube C_k of side length $\ell/2^k$ containing an infinite number of elements of I. Note that the C_k form an infinite sequence of nested hypercubes with side length converging to 0, so by Theorem 14.5.4 the intersection of the C_k is a single point $\mathbf{p}$. We now show that given any open neighborhood of $\mathbf{p}$, no matter how small, we can find a k such that C_k is contained in the neighborhood. More formally

15.1.7 Proposition. *Given a neighborhood $N_\epsilon(\mathbf{p})$, there is an integer k such that $C_k \subset N_\epsilon(\mathbf{p})$.*

Proof. Since the hypercubes C_i are nested, we see that $\mathbf{p}$ is contained in any hypercube of the sequence. The length of the side of C_i is $\ell/2^i$, so by Exercise 14.5.3 the maximum distance between any two points in C_i is $\sqrt{n}\ell/2^i$. So to find a C_i contained in a ϵ neighborhood of $\mathbf{p}$, we simply need to solve for i in

$$\sqrt{n}\frac{\ell}{2^i} < \epsilon \quad \text{or} \quad \sqrt{n}\frac{\ell}{\epsilon} < 2^i.$$

This can be done by taking i large enough, since the left-hand side of the second equality is fixed. Note that we are using the archimedean property. □

By Definition 14.4.3, $\mathbf{p}$ is a limit point of the set $I \subset S$. This proves Theorem 15.1.6. For Theorem 15.1.5, we just add that S is closed. This means that the limit point $\mathbf{p}$ is in S (Definition 14.4.7), which completes the proof. □

15.1.8 Exercise. Draw a figure for the theorem in $\mathbb{R}^2$.

15.1.9 Definition. An *open cover* of a set S in $\mathbb{R}^n$ is a collection of open sets U_i, $i \in I$, where i is an arbitrary index set such that $S \subset \cup_{i \in I} U_i$.

By Theorem 14.4.15, an arbitrary union $\cup_{i \in I} U_i$ of opens is open. The standard definition of compactness says that a set S is compact if every open cover of S admits a finite subcover, meaning that one can find a finite number of U_i that continue to cover S. To avoid confusion with the definition of compactness used in these notes, we will refer to a set with this property as a set with the *finite-open-cover* property. This definition is the right definition of compactness in infinite dimensional spaces, and is equivalent to Definition 15.1.1 in $\mathbb{R}^n$, as the next four exercises establish.

15.1.10 Exercise. In this exercise, you are asked to show that the finite open cover definition of compactness implies Definition 15.1.1. Thus you must show that any set S in $\mathbb{R}^n$ that satisfies the finite-open-cover property is closed and bounded. To show that S is bounded, cover $\mathbb{R}^n$ by $U_i = N_i(\mathbf{0})$, open ball centered at the origin, for all positive integers i. Convince yourself that any point in $\mathbb{R}^n$ is in a $N_i(\mathbf{0})$, for large enough i. If a finite cover for S can be extracted from this cover, there there is a ball with largest radius containing S, which is therefore bounded. To show that S is closed, assume by contradiction that $\mathbf{p}$ is a limit point of S that is not in S. For any positive integer i, let U_i be the complement of the closed ball $\overline{N}_{1/i}(\mathbf{p})$. By Theorem 14.4.13 each U_i is open, and their union covers $\mathbb{R}^n \smallsetminus \mathbf{p}$ so it covers S. Show that no finite subcover of the $\{U_i\}$ covers S.

15.1.11 Exercise. Now assume that S is a closed hypercube. Let U_i, $i \in I$, be an arbitrary open cover of S. Assume that it does not admit a finite subcover, and derive a contradiction as follows. Divide S into smaller and smaller n-cells as in the proof of Theorem 14.5.4. At each step there must be at least one of the smaller hypercubes that does not admit a finite cover by the U_i, $i \in I$. This constructs a nested sequence of hypercubes C_k where the side length goes to 0, which all fail to have a finite subcover. Now there in a unique point **c** in the intersection of all these hypercubes by Theorem 14.5.4. It lies in one of the open sets of the cover, call it U_0. Therefore a small enough ball centered at **c** is in U_0, which means that for a large enough K, the entire hypercube C_K is in U_0. This contradicts the construction of the C_k as hypercubes that do not admit a finite subcover, so we are done.

15.1.12 Exercise. Next show that if C is a closed subset of S, where S has the finite-open-cover property, then C also has the property. Hint: if U_i, $i \in I$, is an open cover of C, then add to this open cover the complement C^c of C, which is open by Theorem 14.4.13. In this way you get an open cover for S, which has the finite-open-cover property.

15.1.13 Exercise. Use the last two exercises to show that any compact set C in $\mathbb{R}^n$ has the finite-open-subcover property, by putting C in a closed hypercube.

Before reading further, you may want to read Chapter 10 to review sequences.

15.2 Cauchy Sequences

As the main theorem of this section shows, Cauchy sequences[1] provide a way of discussing convergent sequences without mentioning the value the sequence converges to. This section is not essential for what follows.

[1] See [12] p.123 for historical details.

15.2.1 Definition. A sequence $\{\mathbf{a}_i\}$ in $\mathbb{R}^n$ is a *Cauchy sequence* if for every $\epsilon > 0$ there is an integer N such that $d(\mathbf{a}_i, \mathbf{a}_j) < \epsilon$ for all $i \geq N$ and $j \geq N$.

15.2.2 Theorem. *Any Cauchy sequence in $\mathbb{R}^n$ converges, and every convergent sequence is Cauchy.*

Proof. First we prove that any convergent sequence $\{\mathbf{a}_i\}$, is Cauchy. Let $\mathbf{a}$ be the limit of the sequence. Thus for any $\epsilon > 0$ there is an integer N so that $i \geq N$ implies $d(\mathbf{a}_i, \mathbf{a}) < \epsilon/2$ for $i \geq N$. But then by the triangle inequality (5.4.14)

$$d(\mathbf{a}_i, \mathbf{a}_j) \leq d(\mathbf{a}_i, \mathbf{a}) + d(\mathbf{a}_j, \mathbf{a}) \leq \epsilon$$

Now we go in the other direction: we assume the sequence is Cauchy. For each positive integer N let C_N be the smallest hypercube containing all the $\mathbf{a}_i$ for $i \geq N$. Because $\{\mathbf{a}_i\}$ is a Cauchy sequence, it is bounded. Thus a C_N exists for each N: be sure you see that, it is the key point of the proof. By elementary logic, since C_{N+1} does not need to contain the point $\mathbf{a}_N$, but otherwise contains the same points in the sequence as C_N, $C_{N+1} \subset C_N$ for all N. So the C_N form a nested sequence of hypercubes. Finally we show that the length of the side of the hypercubes converges to 0 - this is the second use of the Cauchy criterion. Indeed, for every $\epsilon > 0$ there is a N such that $d(\mathbf{a}_i, \mathbf{a}_j) < \epsilon$ for all $i \geq N$ and $j \geq N$. By Exercise 14.5.3 this means that the side of any hypercube C_i, for $i \geq N$ is less than or equal to $\epsilon/\sqrt{n}$. So the limit of the sequence of hypercubes is a single point $\mathbf{a}$, and the Cauchy sequence $\{\mathbf{a}_i\}$ converges to $\mathbf{a}$. □

Note that this last step requires the completeness of $\mathbb{R}$: otherwise there might be nothing to converge to.

15.2.3 Exercise. Write a careful proof to show that Cauchy sequences are bounded.

See §15.5 for an extended example of Cauchy sequences.

15.3 Subsequences of sequences

Next, we need the notion of a subsequence of a sequence. While the formal definition is a little forbidding, the idea is simple: out of the terms of a sequence, just pick out any infinite subset of terms. First an example.

15.3.1 Example. Start with the sequence $\{1/i\}$. We could pick out

- the even numbered terms $1/2, 1/4, 1/6, 1/8, \dots, 1/(2i)$
- or the powers of two: $1/2, 1/4, 1/8, 1/16, \dots, 1/(2^i)$.

- or the terms where i is a prime number: $1/2, 1/3, 1/5, 1/7, 1/11, \dots$ Note that in this case there is no formula for the terms.

There are many other possibilities, of course. An infinite number, in fact.

Here is how one defines a subsequence formally.

15.3.2 Definition. Take any infinite sequence $\{i_j\}$ of increasing positive integers, indexed by j:

$$i_1 < i_2 < i_3 < \cdots < i_j < \cdots .$$

Then $\{x_{i_j}\}$ is a *subsequence* indexed by j of the sequence $\{x_i\}$, indexed by i.

So in the first sequence in Example 15.3.1, $i_1 = 2$, $i_2 = 4$, and $i_j = 2j$, so the j-th term of the subsequence is $1/(2j)$. In the second example, $i_j = 2^j$.

Here is the main way we use compactness in these lectures.

15.3.3 Theorem.

1. *Let $\{\mathbf{a}_i\}$ be an arbitrary sequence in a compact set S in $\mathbb{R}^n$. Then it is possible to find a convergent subsequence $\{\mathbf{a}_{i_j}\}$ of $\{\mathbf{a}_i\}$, converging to a point in S.*
2. *Let $\{\mathbf{a}_i\}$ be a bounded sequence in $\mathbb{R}^n$. Then it is possible to find a convergent subsequence $\{\mathbf{a}_{i_j}\}$ of $\{\mathbf{a}_i\}$.*

Proof. Part (2) of the theorem follows easily from part (1), so let's prove it first, assuming part(1): Since the sequence is bounded, by definition all its value lie in a some closed ball B. B is compact, so the result follows from part (1).

Now we prove part (1). If we denote the collection of points in the sequence by S, then S could be a finite or an infinite set.

If S is a finite set, then, for an infinite number of indices in the sequence, the sequence takes on the same value a. So just take the subsequence corresponding to these indices: we get the constant sequence which obviously converges to its unique value, and we are done.

If S is infinite, we use Theorem 15.1.5, which tells us that the infinite set S has a limit point $\mathbf{p}$.

Here is how we pick the subsequence required by the theorem, which we call $\{\mathbf{b}_j\}$. We first pick an element $\mathbf{b}_1$ from the full sequence $\{\mathbf{a}_i\}$. We can assume that $\mathbf{b}_1 \neq \mathbf{p}$, since the sequence has an infinite number of values. Let $r_1 = d(\mathbf{b}_1, \mathbf{p})$. Then r_1, the distance between $\mathbf{b}_1$ and $\mathbf{p}$, is positive.

We build the sequence recursively as follows, assuming that r_j and $\mathbf{b}_j$, where $\mathbf{b}_j = \mathbf{a}_{i_j}$, for some index i_j, have been defined. Let $r_{j+1} = \frac{r_j}{2}$. By Exercise 14.4.6, we know that the neighborhood $N_{r_{j+1}}(\mathbf{p})$ contains an infinite number of elements of the sequence $\{\mathbf{a}_i\}$. Thus we can find one with index i greater than any given index, in particular greater than i_j. Call that index i_{j+1}, and let $\mathbf{b}_{j+1} = \mathbf{a}_{i_{j+1}}$. This gives a subsequence $\mathbf{a}_{i_j}$, which converges to $\mathbf{p}$ since the distance to $\mathbf{p}$ halves at each step. □

15.4 Sequential Compactness

One of the key uses we have for sequences is the following definition.

15.4.1 Definition. A subset S of $\mathbb{R}^n$ is *sequentially compact* if every sequence in S has a subsequence that converges to a point in S.

15.4.2 Theorem. *A subset S of $\mathbb{R}^n$ is sequentially compact if and only if it is compact.*

Proof. Theorem 15.3.3 shows that compactness implies sequential compactness. To prove the other implication assume that S is sequentially compact. First we show S is closed: to do this we need only show that it contains all its limit points. A limit point of S gives rise to an infinite sequence in S. One can then construct a subsequence converging to the limit point, which by the definition of sequential compactness is in S, so S is closed. Finally we need to show that S is bounded. Assume it is not. Then one can construct a sequence $\{\mathbf{x}_n\}$ with $\|\mathbf{x}_n\| \geq n$. Sequential compactness would imply that one can find a convergent subsequence, but this is impossible since the lengths go to infinity. □

15.5 An Equivalence Relation on Cauchy Sequences

We conclude with an example of Cauchy sequences of rational numbers. This gives an interesting example of an equivalence relation that also illustrates some proof techniques, and sheds some additional light on the completeness of $\mathbb{R}$: §14.2

Let $\{a_i\}$ be an infinite sequence where all the a_i are rational numbers. We start with the index set of all positive integers i , and to each index i we associate a rational number a_i.

We restate the definition of Cauchy sequence using only rational numbers. The only change is the use of $1/M$ where M is a natural number, in place of ϵ.

15.5.1 Definition. We say the sequence $\{a_i\}$ of rational numbers is a *Cauchy sequence* if

$$\forall M \in \mathbb{N}, \exists N \in \mathbb{N} \text{ such that } \forall n \geq N \text{ and } m \geq N \text{ then } |a_n - a_m| < \frac{1}{M}$$

In words: for any rational number of the form $1/M$ there is an positive integer N such that for all integers m and n, both at least N, a_m and a_n are less than $1/M$ apart.

15.5.2 Definition. We say that two Cauchy sequences of rationals, $\{a_i\}$ and $\{b_i\}$, are *equivalent* if

$$\forall M \in \mathbb{N}, \exists N \in \mathbb{N} \text{ such that } \forall n \geq N \text{ then } |a_n - b_n| < \frac{1}{M} \tag{15.5.3}$$

In words they are equivalent if the n-th term of both sequences are arbitrarily close if n is large enough.

15.5.4 Theorem. *Definition 15.5.2 gives an equivalence relation on Cauchy sequences.*

Proof. We need to prove three things. The first two (the fact that it is reflexive and symmetric) are simple and left to the reader. The third one (transitivity) is harder, and gives a good example of a proof involving sequences.

Suppose that we have three Cauchy sequences of rationals $\{a_i\}$, $\{b_i\}$ and $\{c_i\}$. We assume that $\{a_i\}$ and $\{b_i\}$ are equivalent and that $\{b_i\}$ and $\{c_i\}$ are equivalent. We must show that this implies that $\{a_i\}$ and $\{c_i\}$ are equivalent.

What do we know? We fix a positive integer M. Since $\{a_i\}$ and $\{b_i\}$ are equivalent

$$\exists N_1 \in \mathbb{N} \text{ such that } \forall n \geq N_1 \text{ then } |a_n - b_n| < \frac{1}{2M}$$

Since $\{b_i\}$ and $\{c_i\}$ are equivalent

$$\exists N_2 \in \mathbb{N} \text{ such that } \forall n \geq N_2 \text{ then } |b_n - c_n| < \frac{1}{2M}$$

So let N be the larger of N_1 and N_2. Then for $n \geq N$

$$\begin{aligned} |a_n - c_n| &\leq |a_n - b_n| + |b_n - c_n| && \text{by the triangle inequality in } \mathbb{R} \\ &\leq \frac{1}{2M} + \frac{1}{2M} = \frac{1}{M} && \text{by hypothesis} \end{aligned}$$

so we are done. □

Thus equivalence classes of Cauchy sequences of rationals are well defined. So we may ask: what are the equivalence classes? The beautiful answer is that each equivalence class corresponds to a distinct real number, and all real numbers are accounted for in this way.

Indeed, the method can be used to construct the real numbers. This is done in [70], §2.3.1. Note that this approach makes the density of the rationals in the reals (Theorem 14.2.8) almost obvious.

15.5.5 Example. Let x be a real number. Let a_i be the rational number which is the decimal expansion of x truncated after i terms. Then $\{a_i\}$ is a Cauchy sequence representing the real number x. Now modify $\{a_i\}$ by replacing the first N terms, for any fixed integer N, by any numbers whatsoever. Call this sequence $\{b_i\}$. Show that $\{b_i\}$ is equivalent to $\{a_i\}$. If x has a terminating decimal expansion, then the sequence $\{a_i\}$ is constant after a finite number of terms, so the modification proposed does nothing.

Lecture 16

The Maximum Theorem

The main theorem of this chapter is the Weierstrass Maximum Theorem 16.2.2, certainly the most-used theorem of this course. It guarantees that a continuous function on a compact set in $\mathbb{R}^n$ has a global maximum and a global minimum. The theorem builds on the results on compact sets in Lecture 15, and then on some results on continuity that we establish early in the lecture. The reader may want to review more elementary properties of continuity in several variables in Chapter 11, if only to establish notation and to see the statements of the key theorems.

One intermediate result, Theorem 16.2.1, is worth mentioning here: it says that the image of a compact set under a continuous function is compact. The major steps in the proof of the Weierstrass Theorem are conceptually simple and worth understanding. Be sure to remember the key examples: 16.2.3, 16.2.4 and 16.2.5, which show why all the hypotheses of the Weierstrass Theorem are needed.

The lecture concludes with two optional sections. The first, §16.3, gives two generalizations of the notion of continuity, called upper and lower semicontinuity. They come up now because we use them to formulate a generalization of the Weierstrass Theorem that will prove useful later: see §21.5.

16.1 Continuous Functions

The notion of continuity of a real-valued function at a point, using the language of distance functions.[1] is reviewed in Chapter 11. Here we generalize to the case of a function from $\mathbb{R}^n$ to $\mathbb{R}^m$.

From now on, $f(\mathbf{x})$ is a function from a subset $D \subset \mathbb{R}^n$ to $\mathbb{R}^m$. Thus we have two different Euclidean spaces with different Euclidean distance functions.

[1] Stewart [63] gives a good discussion of both the one-variable case (§2.5) and the multi-variable case (§14.2).

To keep them distinct, we call the distance in the domain d_n and the distance in the range d_m.

16.1.1 Definition. A $\mathbb{R}^m$-valued function f defined on a set $D \subset \mathbb{R}^n$ is *continuous* at a point $\mathbf{a} \in D$ if for every $\epsilon > 0$ there exists a $\delta > 0$ such that the set $M = \{\mathbf{x} \in D \mid d_n(\mathbf{x}, \mathbf{a}) < \delta\}$ is mapped by f to the set $M = \{\mathbf{y} \in \mathbb{R}^m \mid d_m(\mathbf{y}, f(\mathbf{a})) < \epsilon\}$. Thus $f(N) \subset M$.

When $\mathbf{a}$ is a point in the interior of D, we can take for N a neighborhood of $\mathbf{a}$ in $\mathbb{R}^n$. However, if $\mathbf{a}$ is on the boundary of D, then no ball around $\mathbf{a}$, no matter how small the radius, lies completely inside D. So we are forced to take instead the set N of the definition, which is the intersection of D with the δ-neighborhood of $\mathbf{a}$ in $\mathbb{R}^n$.

16.1.2 Remark. We could repeat Theorem 11.2.1 in this context. Instead just note that if f is not continuous at $\mathbf{p}$, then there is a sequence of points $\{\mathbf{p}_i\}$ converging to $\mathbf{p}$, but such that the sequence $\{f(\mathbf{p}_i)\}$ does not converge to $f(\mathbf{p})$.

Finally here is a theorem connecting continuity to the metric structure of $\mathbb{R}^n$. This gives us another definition of continuity: a function is continuous if and only if the inverse image under f of every open set is open. First a definition.

16.1.3 Definition. Let f be a function defined on $D \subset \mathbb{R}^n$ and mapping into $\mathbb{R}^m$. Let U be a subset of $\mathbb{R}^m$. Then the *inverse image* $f^{-1}(U)$ under f of a set $U \subset \mathbb{R}^m$ is the set of $\mathbf{x} \in \mathbb{R}^n$ such that $f(\mathbf{x}) \in U$.

Note that we *do not* assume that U is contained in the range of f.

16.1.4 Theorem. *Let $f\colon D \to \mathbb{R}^m$ be a continuous function from a set $D \subset \mathbb{R}^n$ into $\mathbb{R}^m$. Then for any open set U in $\mathbb{R}^m$, $f^{-1}(U)$ is open relative to D.*

Conversely, if the inverse image under f of every open set is open, then f is continuous.

According to Definition 14.6.1 this means that for any $\mathbf{x} \in f^{-1}(U)$, then every $\mathbf{x}' \in D$ sufficiently close to $\mathbf{x}$ is in $f^{-1}(U)$. Because we only use metric properties of $\mathbb{R}^m$, and because any subset of a metric space is a metric space, as we noticed in Exercise 5.4.16, we could avoid the use of relative openness by restricting from $\mathbb{R}^n$ to D.

Proof. To show that $f^{-1}(U)$ is open in D, it is enough to consider a $\mathbf{p} \in D$, with its image $\mathbf{q} = f(\mathbf{p})$ in $\mathbb{R}^m$. Since U is open, for sufficiently small $\epsilon > 0$, the neighborhood $N_\epsilon(\mathbf{q})$ is contained in U. Since f is continuous at $\mathbf{p}$, there is a $\delta > 0$ such that the points $\mathbf{x} \in D \cap N_\delta(\mathbf{p})$ get mapped into $N_\epsilon(\mathbf{q})$. Since $\mathbf{p}$ is an arbitrary point of $f^{-1}(U)$, this says that $f^{-1}(V)$ is open relative to D.

For the converse, pick a point $\mathbf{p} \in D$, and its image $\mathbf{q} = f(\mathbf{p})$ in $\mathbb{R}^m$. Let $U = N_\epsilon(\mathbf{q})$, for an arbitrary $\epsilon > 0$. Consider its inverse image $f^{-1}(U)$. By hypothesis it is open, so it contains a δ-neighborhood of $\mathbf{p}$ in D. By construction this δ-neighborhood of $\mathbf{p}$ is mapped under f to U, which is the statement that f is continuous at $\mathbf{p}$. □

16.1.5 Exercise. Prove that f is continuous if the inverse image under f of any closed set is closed.

Hint: Use the fact that the complement of a closed set is open.

16.1.6 Exercise. Assume that the inverse image under f of any compact set is closed. Show that this does not imply that f is continuous by producing a counterexample. There are counterexamples even when $n = m = 1$. Show that there still are counterexamples when the domain of f is compact. When the range of f is compact, f is continuous by Exercise 16.1.5. Explain.

16.2 The Weierstrass Theorem

The main step in the proof of the Weierstrass theorem is interesting in its own right.

16.2.1 Theorem. *If S is a compact set in $\mathbb{R}^n$, and f a continuous function from S to $\mathbb{R}^m$, then $f(S)$ is compact.*

Proof. Take an infinite sequence $\{\mathbf{y}_n\}$ in $f(S)$. Then we can find a sequence $\{\mathbf{x}_n\}$ in S, such that $f(\mathbf{x}_n) = \mathbf{y}_n$ for each n. Because S is compact and therefore sequentially compact, we can find a subsequence of $\{\mathbf{x}_n\}$ that converges to a $\mathbf{x} \in S$. Continuity of f says that the image of this subsequence under f converges to $f(\mathbf{x})$, which shows that $f(S)$ is sequentially compact and therefore compact. □

Next we turn to the the Weierstrass theorem[2], also called the maximum theorem.

16.2.2 Theorem (Weierstrass Theorem)**.** *If S is a compact set in $\mathbb{R}^n$, and f a continuous function from S to $\mathbb{R}$, then f has a global maximum and a global minimum on S*

This means that there is a point $\mathbf{x}_m \in S$ such that for all $\mathbf{x} \in S$, $f(\mathbf{x}_m) \leq f(\mathbf{x})$, and a point $\mathbf{x}_M \in S$ such that for all $\mathbf{x} \in S$, $f(\mathbf{x}_M) \geq f(\mathbf{x})$. This is sometimes described by the following inaccurate language: "f attains its maximum

[2] References: In Rudin [55] the theorem appears as Theorem 4.16. In Stewart's Calculus book [63] the two variable case is discussed in §14.7, p.959.

and minimum values on S". One should really say that f is bounded on S and that it attains its least upper bound and greatest lower bound on S.

The function may take on its maximum or minimum value at several points, indeed, an infinite number of points. Consider, for example, the functions

- $f(x) = x^2$ on the interval $[-1, 1]$ or
- $g(x) = -x^2 - y^2$ on the unit disk centered at **0** in $\mathbb{R}^2$.

We first produce "counterexamples" to Theorem 16.2.2 when we drop any one of the hypotheses.

16.2.3 Example. The theorem can fail if f is not continuous. Let $f(x)$ be the function on the interval $[-1, 1]$ defined in (11.2.5). Does f has a minimum on the interval? If it has does, the minimum has to be 0. But the function does not take on the value 0, as you can easily see, there is no minimum and the theorem fails.

16.2.4 Example. The theorem can fail if the set S is not closed. Consider the open interval $S = (0, 1)$ and the function $f(x) = x$. As $x \to 0$ from the right, $f(x)$ decreases to 0, but it never takes on the value 0 on S, even though 0 is the glb of the values of f on the interval. As $x \to 1$ from the left, $f(x)$ increases to 1, but it never takes on the value 1 on S, even though 1 is the lub of the values of f on the interval. So this function has neither a minimum nor a maximum on S, and the theorem fails.

16.2.5 Example. The theorem can fail if S is not bounded. Take for S the unbounded interval $[0, \infty)$ of non-negative numbers, and let $f(x) = x$. f gets arbitrarily large as $x \to \infty$, so again the theorem fails.

Next an example in two variables.

16.2.6 Exercise. Consider the function $f(x, y) = x^2 + 2y^2$.

1. Find the minimum value of this function on the closed disk $\overline{N}_1(\mathbf{0})$. Where is it attained?
2. Restrict the function $f(x, y)$ to the boundary of $\overline{N}_1(\mathbf{0})$, which is the unit circle. Using a trigonometric substitution, study the behavior of the function f there: in particular find its minima and its maxima.
3. Finally find the maxima on $\overline{N}_1(\mathbf{0})$. Justify your answer.

We now prove the Weierstrass Theorem 16.2.2. As noted in the first lecture, Remark 1.1.6, it is enough to show the function f has a minimum, since to get the maximum we simply take $-f$.

Theorem 16.2.1 tells us that $f(S)$ is compact in $\mathbb{R}$.

Thus, by Theorem 14.2.5, the set of all values of $f(\mathbf{x})$ on S has a greatest lower bound m, which by compactness of $f(S)$ is in $f(S)$, since it is by definition a limit point of a sequence in $f(S)$. Then there is a $\mathbf{x} \in S$ with $f(\mathbf{x}) = m$, and the theorem is proved.

16.3 Lower Semicontinuous Functions

In Lecture 21, in order to understand the continuity of convex functions of the boundary of their domain, after establishing their continuity on the interior, we generalize the notion of continuity of a function to that of lower and upper semicontinuity. We mainly consider lower semicontinuity, because, as we will see in §16.4 lower semicontinuous functions achieve their minimum on compact sets, so they satisfy a generalized Weierstrass Theorem 16.4.1.

We first define the $\liminf$ of a function at a point. Recall the notation $\overline{N}_\epsilon(\mathbf{x}_0)$ for the closed ball of radius ϵ around the point $\mathbf{x}_0$ from Definition 14.4.1. We write $\overline{\mathbb{R}}$ for the real numbers extended by ∞ and $-\infty$.

16.3.1 Definition. The *limit inferior* of a sequence $\{x_n\}$ in $\mathbb{R}$ is its smallest limit point (see Definition 14.4.3) in $\overline{\mathbb{R}}$. It is written $\liminf\{x_n\}$. Let f be a function from an open set $S \subset \mathbb{R}^n$ to $\mathbb{R}$, and let $\mathbf{x}_0$ be a point in the closure $\overline{S}$ of S. Then let

$$\liminf_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) = \lim_{\epsilon\searrow 0} \left(\inf\{f(\mathbf{x}) \mid \mathbf{x} \in \overline{N}_\epsilon(\mathbf{x}_0) \cap S\} \right)$$

Notice that the right-hand side defines an increasing sequence (since we are taking the $\inf$ over smaller and smaller sets), so the limit exists (if the sequence is unbounded, we say its limit is ∞).

16.3.2 Definition. We say that f, a function from a set $S \subset \mathbb{R}^n$ to $\mathbb{R}$, is *lower semicontinuous* if for all $\mathbf{x}_0 \in S$, $f(\mathbf{x}_0) = \liminf_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x})$.

16.3.3 Exercise. Check that an alternate way of defining lower semicontinuity is:

f is lower semicontinuous at $\mathbf{x}_0 \in S$, if for all $\epsilon > 0$ there is a $\delta > 0$ such that for all $\mathbf{x} \in N_\delta(\mathbf{x}_0) \cap S$, $f(\mathbf{x}) > f(\mathbf{x}_0) - \epsilon$.

This exercise makes it clear that continuous functions are lower semicontinuous: for continuous functions, the conclusion is that $f(\mathbf{x})$ is in the open interval $(f(\mathbf{x}_0) - \epsilon, f(\mathbf{x}_0) + \epsilon)$, while lower semicontinuity only requires the that $f(\mathbf{x})$ be in the unbounded interval $(f(\mathbf{x}_0) - \epsilon, \infty)$.

16.3.4 Theorem. *The function f is lower semicontinuous at $\mathbf{x}_0 \in S$ if and only if the following two conditions are satisfied. Let $c_0 = f(\mathbf{x}_0)$.*

- *For any sequence $\{\mathbf{x}_n\}$ in S converging to $\mathbf{x}_0$ such that the sequence $\{f(\mathbf{x}_n)\}$ has a limit c, then $c_0 \le c$.*
- *There is a sequence $\{\mathbf{x}_n\}$ in S converging to $\mathbf{x}_0$ such that* $\lim_n f(\mathbf{x}_n) = c_0$.

This follows immediately from the definition.

16.3.5 Example. Let $\mathcal{I}$ be the closed interval $[-1/2, 1/2]$ in $\mathbb{R}$. Let $f(x)$ be the function on $\mathcal{I}$ given by

$$f(x) = \begin{cases} x+1, & \text{if } -1/2 \le x < 0; \\ x, & \text{if } 0 \le x \le 1/2. \end{cases}$$

Then $f(x)$ is lower semicontinuous, but not continuous at 0.

Similarly, if $g(x)$ is the function defined on all of $\mathbb{R}$, given by

$$g(x) = \begin{cases} \frac{1}{x^2}, & \text{if } -1/2 \le x < 0; \\ x, & \text{if } 0 \le x \le 1/2. \end{cases}$$

Then $g(x)$ is lower semicontinuous but not continuous at 0.

By a simple change in the definition of f:

$$h(x) = \begin{cases} x+1, & \text{if } -1/2 \le x \le 0; \\ x, & \text{if } 0 < x \le 1/2. \end{cases}$$

we get a function that is not lower semicontinuous at 0. Note that both f and g attain their minimum value 0 on $\mathcal{I}$, while h does not. As we see in the next section, lower semicontinuous function attain their minimum on compact sets. On the other hand h is upper semicontinuous and attains its maximum.

A more interesting example is given below in Example 16.3.8.

In Definition 21.1.14 we define the sublevel sets S_c of any real-valued function $f(\mathbf{x})$ defined on a set S. S_c is just the set where f takes values at most c. They are connected to lower semicontinuity by the following theorem.

16.3.6 Theorem. *The function f is lower semicontinuous on S if and only if $\forall c \in \mathbb{R}$, the sublevel set S_c is closed relative to S.*

Proof. First assume f is lower semicontinuous. Take a sequence of points $\{\mathbf{x}_n\}$ in S_c converging to a point $\mathbf{x}_0 \in S$. We need to show that $\mathbf{x}_0$ is in S_c. Since $f(\mathbf{x}_n) \leq c$, then $\liminf_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) \leq c$. Then lower semicontinuity of f says that $f(\mathbf{x}_0) \leq c$, so $\mathbf{x}_0 \in S_c$ as required.

Next assume that f is not lower semicontinuous at $\mathbf{x}_0$. Then there is a sequence $\{\mathbf{x}_n\}$ in S converging to $\mathbf{x}_0$ with $\liminf_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) = c < f(\mathbf{x}_0)$. Let $\epsilon = (f(\mathbf{x}_0) - c)/2$. Then, for large enough n, $\mathbf{x}_n$ is in $S_{c+\epsilon}$, while $\mathbf{x}_0$ is not. This contradicts the hypothesis that $S_{c+\epsilon}$ is closed.

Remark: we also have to consider the case where c is $-\infty$. The proof above goes through by taking for $(c+\epsilon)$ any number less than $f(\mathbf{x}_0)$. □

16.3.7 Example. The sublevel set S_c of the function $f(x)$ from Example 16.3.5 is the closed interval $0 \leq x \leq c$, when $c < 1/2$, the union of the point $x = -1/2$ and the interval $0 \leq x \leq 1/2$, when $c = 1/2$, and the union of the two intervals $-1/2 \leq x \leq c - 1$ and $0 \leq x \leq c$ when $c \leq 1/2$, confirming that f is lower semicontinuous. On the other hand, the sublevel set S_c of the function $h(x)$, for $0 < c < 1/2$, is the interval $0 < x \leq c$, which is not closed, confirming that h is not lower semicontinuous.

16.3.8 Example. Here is a multivariable example of a lower semicontinuous function that is not continuous. We will study this example in detail in §9.1: it is the Rayleigh quotient of a quadratic form. Start with the function $f(x,y)$ on $\mathbb{R}^2 \setminus \{\mathbf{0}\}$ given by

$$f(x,y) = \frac{3x^2 - 2xy + 3y^2}{x^2 + y^2} \tag{16.3.9}$$

For any positive real number λ, $f(\lambda x, \lambda y) = f(x,y)$, so f is homogenous of degree 0: see §12.3. Since f is clearly continuous on the unit circle

$$U = \{x^2 + y^2 = 1\} \subset \mathbb{R}^2,$$

by the Weierstrass Theorem 16.2.1 it attains its minimum m and its maximum M on U. Parametrizing U by $x = \cos t$ and $y = \sin t$, $0 \leq t \leq 2\pi$, we get

$$g(t) = f(\cos t, \sin t) = 3 - 2\cos t \sin t = 3 - \sin 2t,$$

using the usual trigonometric identities. Then $m = 2$ and $M = 4$. Then extend $f(x,y)$ to all of $\mathbb{R}^2$ by setting $f(\mathbf{0}) = 2$. This extended function is a lower semicontinuous function on $\mathbb{R}^2$ but is not continuous at the origin. Indeed, if you take a sequence of points (x_i, y_i) approaching the origin along suitable rays you can get any limit between 2 and 4. If instead we extend f by setting $f(\mathbf{0}) = 4$, the extended function is upper semicontinuous.

For more complicated Rayleigh quotients the determination of m and M is an exercise in linear algebra: the computation of the eigenvalues and eigenvectors of a symmetric matrix: See Theorem 9.1.2.

16.3.10 Exercise. Be sure that you can graph $z = f(x, y)$ in $\mathbb{R}^3$.

It is left to you to define the $\limsup$, the notion of upper semicontinuous function, and the analog of Theorem 16.3.6 for the upper level set $S^c = \{\mathbf{x} \in S \mid f(\mathbf{x}) \geq c\}$.

16.4 A Generalization of the Weierstrass Theorem

In these lectures we will only apply this generalization when we know that f is convex. Lemma 21.3.8 then implies that f takes values in $\mathbb{R} \cup \infty$, which is what we assume in the next theorem.

16.4.1 Theorem (Generalized Weierstrass Theorem)**.** *If S is a compact set in $\mathbb{R}^n$, and f a lower semicontinuous function from S to $\mathbb{R} \cup \infty$, then f has a global minimum on S.*

Proof. Consider the set C of $c \in \mathbb{R}$ such that the sublevel set S_c is non empty. We first show that C is bounded below. Suppose not: then there is a sequence $\{c_n\}$ of values of f on S getting arbitrarily negative. Thus we get a sequence of points $\{\mathbf{x}_n\}$ in S, with $f(\mathbf{x}_n) = c_n$. Since S is compact, we can extract a converging subsequence that we still call $\{\mathbf{x}_n\}$. Lower semicontinuity then says that the value of f at the limit $\mathbf{x}$ of the sequence is $-\infty$. This contradicts our hypothesis that f takes values in $\mathbb{R} \cup \infty$.

Thus C is bounded below. So Theorem 14.2.5 says that C has a greatest lower bound, that we call c_0 Let $\{c_n\}$ be a decreasing sequence of real numbers converging to c_0. Then each S_{c_n} is non-empty, so there is a $\mathbf{x}_n \in S_{c_n}$. Because S is compact we can extract a subsequence from the sequence $\{x_n\}$ that converges to some $\mathbf{x}_0$. The fact that f is lower semicontinuous implies that $f(\mathbf{x}_0) \leq c_0$, so we must have $f(\mathbf{x}_0) = c_0$, showing that the minimum value is actually attained, and we are done. □

We will use this result in the proof of Theorem 29.7.2.

There is of course a second theorem obtained by letting f be an upper semicontinuous function from a compact set S to $\mathbb{R} \cup -\infty$. The conclusion is that f has a global maximum on S. The proof is left to you: just replace f by $-f$.

Lecture 17

Function Graphs and the Implicit Function Theorem

With this lecture we begin our study of nonlinear optimization by examining the geometry of the feasible set. We have already looked at convex feasible sets in Lecture 22, and on polyhedral feasible sets in Lectures 19 and 25. Here we look at a set $F \subset \mathbb{R}^n$ defined by the vanishing of m functions $h_i(x_1, \dots, x_n)$, $m < n$. When the equations $h_i = 0$ are of degree 1, this is what we studied in Lectures 19 and 25. When the functions h_i are more complicated: e.g. polynomials of degree > 1, it can be difficult even to find a single point $\mathbf{x}^*$ in F. Once we have found such a point $\mathbf{x}^*$, we need to determine what F looks like in a neighborhood of $\mathbf{x}^*$.

This turns out to be a difficult question, except when the point $\mathbf{x}^*$ is *regular*: see Definition 17.1.4. When $\mathbf{x}^*$ is regular, and the functions h_i are $\mathcal{C}^1$, we can write a neighborhood of $\mathbf{x}^*$ in F as the graph of a function, using the Implicit Function Theorem 17.6.6. An important mathematical object emerges from this discussion: the tangent space to F at a regular point. We study it in §17.2. We also discuss vector fields.

The main result of this lecture is the Implicit Function Theorem, stated in §17.6. It is worth first studying the special case given in §17.5, and even more importantly the linear case stated as Theorem 17.6.3. They both help understand the full theorem. Work through the examples in §17.8 to understand the statement of what is probably the most important theorem in multivariable calculus. Then in §17.9 there are corollaries to be used in later lectures on nonlinear optimization, starting with §28.6 and then §31.2.

17.1 Subsets of a Real Vector Space

We are given m functions h_i, $1 \leq i \leq m$, in n variables x_j, $1 \leq j \leq n$, where $m < n$.

17.1.1 Definition. Let U be the intersection of the domains of all the h_i. We assume that U is not empty. Write $\mathbf{h}(\mathbf{x})$ for the vector function from $U \subset \mathbb{R}^n \to \mathbb{R}^m$ whose coordinate functions are the h_i. The function $\mathbf{h}(\mathbf{x})$ defines a map from $U \subset \mathbb{R}^n$ to $\mathbb{R}^m$.

Let

$$F = \{\mathbf{x} \in U \subset \mathbb{R}^n \mid \mathbf{h}(\mathbf{x}) = \mathbf{0}\} \tag{17.1.2}$$

We call this subset F because we are thinking of it as the feasible set for an optimization problem.

Simply deciding whether the set F is empty or not is already difficult. Consider, for example, the locus given by the vanishing of the single equation

$$h(x_1, \dots, x_n) = \sum_{j=1}^{n} x_j^2 - t = 0$$

where t is a real number. When t is greater than 0, we get the sphere of radius $\sqrt{t}$ in $\mathbb{R}^n$. However, if $t < 0$ the locus is empty.

Assume F is non-empty, indeed, that we have found a point $\mathbf{x}^*$ in it. The next question is: what does F look like in a neighborhood U of $\mathbf{x}^*$? Assume the functions h_i are $\mathcal{C}^1$. Write

$$\nabla\mathbf{h} = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \dots & \frac{\partial h_1}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial h_m}{\partial x_1} & \dots & \frac{\partial h_m}{\partial x_n} \end{bmatrix}. \tag{17.1.3}$$

The i-th row of the $m \times n$ matrix $\nabla\mathbf{h}$ is the gradient ∇h_i of the scalar function h_i. The ij-th entry of the matrix is $\frac{\partial h_i}{\partial x_j}$. When we evaluate the matrix $\nabla\mathbf{h}$ at $\mathbf{x}^*$, we write $\nabla\mathbf{h}(\mathbf{x}^*)$.

17.1.4 Definition. A point $\mathbf{x}^*$ satisfying $\mathbf{h}(\mathbf{x}^*) = \mathbf{0}$ is *regular* if the $m \times n$ matrix $\nabla\mathbf{h}(\mathbf{x}^*)$ has rank m.

Here is an equivalent way of stating regularity, that readers who have studied differential geometry may have seen.

17.1.5 Definition. A *submersion* is a $\mathcal{C}^1$-map $\mathbf{h}$ from an open set U in $\mathbb{R}^n$ to $\mathbb{R}^m$ such that the $m \times n$ matrix $\nabla\mathbf{h}$ has (maximal) rank m at all $\mathbf{x} \in U$.

If $\mathbf{x}^*$ is regular, a $m \times m$ minor of the matrix $\nabla \mathbf{h}$ does not vanish at $\mathbf{x}^*$. Since the entries of this matrix are continuous (by the assumption that $\mathbf{h}$ is $\mathcal{C}^1$), this minor does not vanish in a small enough neighborhood of $\mathbf{x}^*$. So regularity is an *open condition* on the zero locus of a $\mathcal{C}^1$ map.

Thus

17.1.6 Theorem. *Regularity at a point $\mathbf{x}^*$ in the set $F = \{\mathbf{x} | \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$ implies regularity for all $\mathbf{x}$ in a small enough neighborhood of $\mathbf{x}^*$ in F.*

17.2 Tangent Vectors and Vector Fields

To each point $\mathbf{p} \in \mathbb{R}^n$, we attach a new vector space $T_{\mathbf{p}}$ of dimension n, which we call the *tangent space* to $\mathbb{R}^n$ at $\mathbf{p}$.

Thus we get an n-dimensional collection of n-dimensional vector spaces. To specify an element of $\bigcup_{\mathbf{p} \in \mathbb{R}^n} T_{\mathbf{p}}$ requires two n-tuples of real numbers: first, the point $\mathbf{p}$ and second an n-tuple $(v_1, v_2, \dots, v_n)$ determining a vector $\mathbf{v} \in T_{\mathbf{p}}$, which we called a tangent vector to $\mathbb{R}^n$ at $\mathbf{p}$. We write it $\mathbf{v}_{\mathbf{p}}$, and view it as an arrow starting at $\mathbf{p}$ with tip at $\mathbf{p} + \mathbf{v}$.

Equivalently, we can consider an ordered pair of points $\mathbf{x}_0$ and $\mathbf{x}_1$ in $\mathbb{R}^n$ as the element $\mathbf{v} = \mathbf{x}_1 - \mathbf{x}_0$ in $T_{\mathbf{x}_0}$.

17.2.1 Example. Here are some examples of tangent vectors at different points in $\mathbb{R}^2$. First a tangent vector at $\mathbf{p} = (1, 0)$. Note that addition is performed in $T_{\mathbf{p}}$ as if the origin of the vector space had been transferred to $\mathbf{p}$. If $\mathbf{v}_{\mathbf{p}} = (1, 1)_{\mathbf{p}}$ and $\mathbf{w}_{\mathbf{p}} = (1, -1)_{\mathbf{p}}$ then $\mathbf{v}_{\mathbf{p}} + \mathbf{w}_{\mathbf{p}} = (2, 0)_{\mathbf{p}}$. and then a tangent vector at $\mathbf{q} = (-1, 1)$: $\mathbf{v}_{\mathbf{q}} = (-1, 0)_{\mathbf{q}}$.

There is how this concept will come up in this course.

17.2.2 Definition. Let f_1, f_2, $\dots f_n$ be n functions of the n variables x_1, $\dots$, x_n with common domain U. Then, $\mathbf{e}_i$ being the standard coordinate vectors, write

$$f_1\mathbf{e}_1 + f_2\mathbf{e}_2 + \cdots + f_n\mathbf{e}_n \tag{17.2.3}$$

defines a *vector field* on U, namely an assignment for every $\mathbf{x}^* \in U$ of the tangent vector $(f_1(\mathbf{x}^*), \dots, f_n(\mathbf{x}^*))_{\mathbf{x}^*}$.

The most important vector field for us is the gradient ∇f of a differentiable function $f(\mathbf{x})$, where

$$f_i = \frac{\partial f}{\partial x_i}$$

It is a vector field where the tangent vector assigned to $T_{\mathbf{p}}$ is $\nabla f(\mathbf{p})$.

When is a vector field (17.2.3) the gradient of a function f? When it is, it is called a *conservative* vector field, and f is called the *potential function* of the vector field. A necessary condition when the f_i are differentiable is given by Clairault's Theorem 12.1.26.

$$\frac{\partial f_i}{\partial x_j} = \frac{\partial f_j}{\partial x_i} \quad \text{for all } 1 \leq i, j \leq n$$

Further analysis of this issue is given, for example, in [63], §16.3.

17.2.4 Example. For the function $f(x, y) = x^2 + 2x - y^3$ draw the gradient field of $f(x, y)$ at a few points. Now consider the curve in the plane given by the equation $f(x, y) = 0$. Add it to the gradient field graph. What can you say?

17.3 The Tangent Space of a Subspace

We define the tangent space of a subspace of $\mathbb{R}^n$ at a point. Recall from §17.2 the definition of the n-dimensional tangent space $T_{\mathbf{p}}$ of $\mathbb{R}^n$ at $\mathbf{p}$. It allows us to define the tangent space of a subspace as follows:

17.3.1 Definition. If M is a subspace of $\mathbf{x} \in \mathbb{R}^n$ given by $h_i(\mathbf{x}) = 0$, $1 \leq i \leq m$, and $\mathbf{x}^* \in M$, then the *tangent space* $T_{M,\mathbf{x}^*}$ of M at $\mathbf{x}^*$ is the subspace of $T_{\mathbf{x}^*}$ orthogonal to the m vectors $\nabla h_i(\mathbf{x}^*)$.

This definition is mainly useful when x^* is regular. Then, since the matrix $\nabla \mathbf{h}(\mathbf{x}^*)$ has rank m, the nullspace, which is the tangent space, has dimension $n - m$. This describes $T_{M,\mathbf{x}^*}$ as a linear subspace of $T_{\mathbf{x}^*}$. We are more interested in seeing what it looks like inside of $\mathbb{R}^n$ itself. For that, we consider the tangent space from the point of view of the Taylor polynomial of the functions h_i.

First the case where there is just one constraint $h(\mathbf{x}) = \mathbf{0}$. Let $P_1(\mathbf{x}^*, \mathbf{x})$ be the Taylor polynomial of degree 1 of $h(x_1, \dots, x_n)$ centered at a point $\mathbf{x}^*$ where $h(\mathbf{x}^*) = 0$. Then

$$P_1(\mathbf{x}^*, \mathbf{x}) = h(\mathbf{x}^*) + \nabla h(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*) = \nabla h(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*).$$

Assume that $\nabla h(\mathbf{x}^*)$, the gradient of h evaluated at $\mathbf{x}^*$, is a non-zero vector. Then we get a non-trivial affine equation

$$\nabla h(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*) = 0. \tag{17.3.2}$$

This equation is the best linear approximation of $h(\mathbf{x}) = 0$ at the point $\mathbf{x}^*$, and it is the equation of the tangent space of $h = 0$ at $\mathbf{x}^*$ in $\mathbb{R}^n$.

Now make the same construction for the m- vector $\mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), \dots, h_m(\mathbf{x}))$ of $\mathcal{C}^1$ functions. Assuming none of gradients are the zero vector, we get a collection of m affine equations

$$\nabla \mathbf{h}(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*) = \mathbf{0}.$$

These equations given the best linear approximation to $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ near $\mathbf{x}^*$, and generalize the equation found in Definition 17.3.1.

We can rewrite these equations as

$$\nabla \mathbf{h}(\mathbf{x}^*) \cdot \mathbf{x} = \mathbf{b}, \text{where } \mathbf{b} \text{ is the constant } \nabla \mathbf{h}(\mathbf{x}^*) \cdot \mathbf{x}^* \tag{17.3.3}$$

The coefficients $\nabla \mathbf{h}(\mathbf{x}^*)$ on the left-hand side form a $m \times n$ matrix of rank m if and only if the point $\mathbf{x}^*$ is regular for the equations $\mathbf{h}(\mathbf{x}) = \mathbf{0}$.

17.3.4 Theorem. *$T_{M,\mathbf{x}^*}$ is an affine space of dimension $\geq n - m$. It has dimension $n - m$ if and only if $\mathbf{x}^*$ is regular for the constraints $\mathbf{h}$.*

Proof. This is just a way of restating the rank-nullity theorem of linear algebra, quoted in §7.2, giving the dimension of the nullspace of a linear map in terms of its rank. Since m is the maximum possible rank, it corresponds to the nullspace of smallest possible dimension, as the theorem states. □

For more details on tangent spaces, see [47].

17.3.5 Example. Assume we are in $\mathbb{R}^2$ with coordinates x_1 and x_2. Let $h(x_1, x_2)$ be the function $x_1^2 + x_2^2 - 1$, and let M be the subset of $\mathbb{R}^2$ given by $h(x_1, x_2) = 0$, so that M is the unit circle centered at the origin. The matrix of partials of h is just the gradient matrix $[2x_1, 2x_2]$. If we evaluate at any point on the unit circle, we do not get the zero matrix, which just says that the matrix has rank 1, the maximal rank possible. We write the Taylor polynomial of degree 1 of h at (x_1^*, x_2^*), and set it to zero to give us the linear equation of the tangent line: $2x_1^*(x_1 - x_1^*) + 2x_2^*(x_2 - x_2^*) = 0$ or

$$x_1^* x_1 + x_2^* x_2 = 1.$$

So if (x_1^*, x_2^*) is the point $(1, 0)$, we get as equation for the tangent line $x_1 = 1$, while if it is $(-1, 0)$, we get $x_1 = -1$. If the point $\mathbf{x}^*$ is $(\cos\theta, \sin\theta)$, then the equation of the tangent line to the circle at $\mathbf{x}^*$ is

$$\cos(\theta) x_1 + \sin(\theta) x_2 = 1.$$

The Implicit Function Theorem will allow us to transform a feasible set F into the graph of a function in a neighborhood of a regular point $\mathbf{x}^*$, so the next thing we do is study such graphs in §17.4.

17.4 Graphs of functions

Assume we have a function $\mathbf{g}\colon \mathbb{R}^{n-m} \to \mathbb{R}^m$. We denote its m coordinate functions by g_i, $1 \le i \le m$. You may wonder why we have changed our indexing convention: why does $\mathbf{g}$ take its values in $\mathbb{R}^{n-m}$ instead of $\mathbb{R}^n$? This is because $\mathbf{g}$ will serve later at the implicit function of the Implicit Function Theorem, and the indexing above is the indexing we will need. It is also because we want the graph of $\mathbf{g}$ to be a subset of $\mathbb{R}^n$.

You of course know what the graph of a real-valued function $f(\mathbf{x})$ is. In the current context we would write it at the subset of $\mathbb{R}^{n+1}$ given as $(f(\mathbf{x}), \mathbf{x})$. We now generalize this to a function mapping to $\mathbb{R}^m$.

17.4.1 Definition. The graph of a function $\mathbf{g}\colon \mathbb{R}^{n-m} \to \mathbb{R}^m$ is the subset Γ of $\mathbb{R}^n$ given by

$$\Gamma = (g_1(\mathbf{x}), \dots, g_m(\mathbf{x}), x_1, \dots, x_{n-m})$$

for all $\mathbf{x} = (x_1, \dots, x_{n-m})$ in the domain of $\mathbf{g}$.

It is a matter of convenience to list the values of $\mathbf{g}$ before the variables. In this representation the last $n-m$ variables can vary freely and are therefore called the free variables. while the first m variables are uniquely determined by the free variables, and are therefore called the bound variables.

17.4.2 Example. Let $\mathbf{g}(x_1, x_2)$ be the function from $\mathbb{R}^2$ to $\mathbb{R}^2$ with coordinate functions $g_1(x_1, x_2) = x_1^2 + x_2^2$, and $g_2(x_1, x_2) = x_1^3 - x_2^3 - 1$. The graph of $\mathbf{g}(x_1, x_2)$ is the collection of points in $\mathbb{R}^4$ $(x_1^2 + x_2^2, x_1^3 - x_2^3 - 1, x_1, x_2)$, for all x_1 and x_2 in $\mathbb{R}^2$. So $n = 4$ and $m = 2$.

We get a $m \times (n-m)$ matrix $\nabla \mathbf{g}$ of partial derivatives of the m real-valued functions g_i with respect to the $n-m$ free variables.

Expressing a part of a subset of $\mathbb{R}^n$ in this way is known as *parametrizing* it. We will only be considering the *implicit function parametrizations* derived from the Implicit Function Theorem. We want to see what the chain rule tells us when we are dealing with the graph of a function.

17.4.3 Definition. In Theorem 22.1.2, we used the well-known expression for the tangent hyperplane to the graph of a real-valued function at a point $(x_1^*, \dots, x_{n-m}^*)$. It generalizes readily to the expression for the tangent space to the graph of a function $\mathbf{g}$ to $\mathbb{R}^m$. It is the intersection of the tangent hyperplanes to each one of the g_i, so we write

$$y_i = g_i(\mathbf{x}^*) + \nabla g_i|_{\mathbf{x}^*}(\mathbf{x} - \mathbf{x}^*).$$

In this notation the variables on $\mathbb{R}^n$ are $y_1, \dots, y_m, x_1, \dots, x_{n-m}$. Move the terms with the variables $\mathbf{x}$ $\mathbf{y}$ to the left hand side, and the constants to the right hand side.

In block matrix notation, this gives the system of m linear equations in n variables $\mathbf{y}$ and $\mathbf{x}$

$$\begin{bmatrix} -I_m & \nabla \mathbf{g}|_{\mathbf{x}^*} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} -I_m & \nabla \mathbf{g}|_{\mathbf{x}^*} \end{bmatrix} \begin{bmatrix} -\mathbf{g}(\mathbf{x}^*) \\ \mathbf{x}^* \end{bmatrix}$$

This matrix of coefficients of this system contains the identity matrix, so it has maximal rank m. So the set of solutions is an affine space of dimension $n - m$, and this is the tangent space to the graph.

17.4.4 Example. Continuing with Example 17.4.2, when $x_1^* = 1$ and $x_2^* = 0$, we get the point $(1, 0, 1, 0)$ on the graph. $\nabla g_1|_{(1,0)} = (2, 0)$ and $\nabla g_2|_{(1,0)} = (3, 0)$. The two tangent equations are

$$\begin{aligned} y_1 &= 1 + 2(x_1 - 1) \\ y_2 &= 3(x_1 - 1), \end{aligned}$$

so the variable x_2 does not appear.

17.5 The Implicit Function Theorem: the Simplest Case

Let $h(x, y)$ be a $\mathcal{C}^1$ function of two variables. Let C be the set of solutions in $\mathbb{R}^2$ of the equation $h(x, y) = 0$, and let (x_0, y_0) be a point on C. Since $h(x, y)$ is $\mathcal{C}^1$, we can compute its partials, and evaluate then at (x_0, y_0). Set

$$\frac{\partial h}{\partial x}(x_0, y_0) = a \text{ and } \frac{\partial h}{\partial y}(x_0, y_0) = b.$$

In calculus[1] you learned how to compute the derivative of $h(x_0, y_0)$ implicitly in a neighborhood of the point (x_0, y_0), if certain conditions are satisfied.

Namely, if $a \neq 0$, then in a neighborhood of (x_0, y_0) the curve C can be written as a graph $(g(y), y)$ for a $\mathcal{C}^1$ function $g(y)$ such that $g(y_0) = x_0$ and $g'(y_0) = -b/a$.

17.5.1 Example. The simplest case is the affine function $h(x, y) = ax + by - c$, for constants a, b and c. Note that a and b are the appropriate partials of h as per the notation above. In this case we can solve explicitly for x as a function of y: $x = -\frac{b}{a}y$. So this is $g(y)$, and the derivative is $-\frac{b}{a}$ as predicted by (17.5.8).

17.5.2 Example. The next example is $h(x, y) = x^2 + y^2 - 1$. The vanishing locus of h is the unit circle centered at the origin in the plane. We assume that this

[1] See [63], §3.6 and also p. 936.

relation makes x an implicit function g of y in a small neighborhood of a point (x_0, y_0) such that h is satisfied, meaning $h(x_0, y_0) = x_0^2 + y_0^2 - 1 = 0$. In fact in this case, we can write g explicitly. If x is non-negative, $g(y) = \sqrt{1-y^2}$ and if x is negative, $g(y) = -\sqrt{1-y^2}$. When $y = \pm 1$, so $x = 0$ there is a problem: indeed

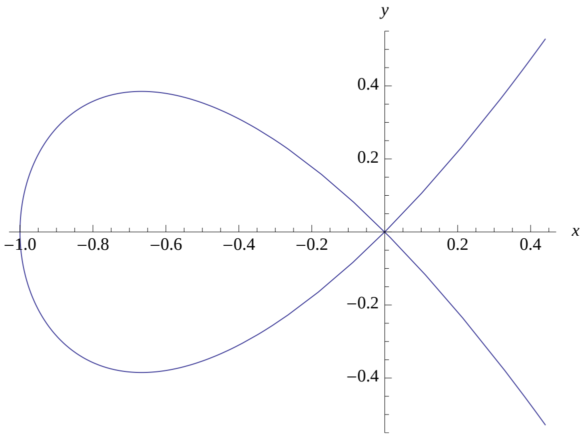
$$\frac{\partial h}{\partial x} = 2x$$

so (17.5.7) is not satisfied at $(0, \pm 1)$. At all other points on the circle, it is satisfied, the points of the circle are regular for the constraint of being on the circle. By reversing the roles of x and y, you can see that the remaining two points $(0, \pm 1)$ are also regular, but you need to write y as a function of x.

17.5.3 Example (The Nodal Cubic)**.** Consider the equation

$$h(x, y) = y^2 - x^2(x + 1) = 0.$$

Here is a graph of the curve, called the *nodal cubic*, because of the singular point at the origin, which is called a node. By definition a point is singular if it is not regular.

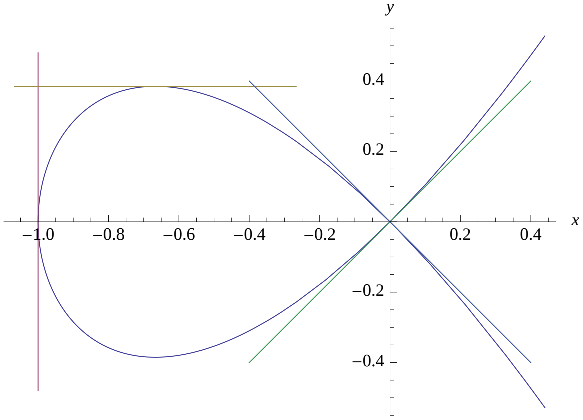


The gradient of h is $\nabla h = (-2x - 3x^2, 2y)$. Consider the following cases:

1. First let us locate all the points of C where the gradient is the zero vector. When x and y are both 0, the gradient is 0, and the point is on the curve. There is also the solution $x = -2/3$, $y = 0$, but since $h(-2/3, 0) \neq 0$, this is not a point on the curve C.

2. Now consider the points where only the first coordinate of the gradient vanishes, so the gradient is vertical. There is just one solution in x, giving $-2/3$, and that determines two values of y, $y = \pm\sqrt{1/3}2/3$. The gradient is perpendicular to the tangent to the curve, drawn on the next graph for the point with positive y coordinate.

3. Finally consider the points where only the second coordinate of the gradient vanishes: $y = 0$, so $x^2(x+1) = 0$. We have already considered the point $(0,0)$, so we are left with $(-1,0)$ where we have also drawn the tangent line.



We ask: in the neighborhood of which points on the locus $y^2 - x^2(x+1) = 0$ does the graph make x a function of y? at which points does is make y a function of x? It is clear that no matter how small a neighborhood you take of the origin, the locus is not the graph of a function.

Here is a different way of describing C: we can parametrize it in terms of the parameter t,

$$x(t) = t^2 - 1 \quad \text{and } y(t) = t(t^2 - 1)$$

Indeed, plug this values into h and see what happens when you expand $h(t^2 - 1, t(t^2-1))$. We think of t as time, and the curve C is then the trajectory of a particle, starting at the bottom right, going through the origin a first time at $t = -1$, going around the loop, going through the origin a second time at $t = 1$ and exiting the picture at the upper right. The velocity vector $v(t) = (x'(t), y'(t)) = (2t, 3t^2 - 1)$, so we can draw the tangent line at any point of the curve.

Back to the general case. Assume the relation given by $h(x,y) = 0$ makes x an implicit function of y in a small neighborhood. Writing this unknown function $x = g(y)$, we have

$$h(g(y), y) = 0 \tag{17.5.4}$$

so take the derivative with respect to y using the chain rule: you get

$$\frac{\partial h}{\partial x}(g(y), y)\frac{dg}{dy} + \frac{\partial h}{\partial y}(g(y), y) = 0 \tag{17.5.5}$$

which can be evaluated at (x_0, y_0), where $x_0 = g(y_0)$:

$$\frac{\partial h}{\partial x}(x_0, y_0)\frac{dg}{dy}(y_0) + \frac{\partial h}{\partial y}(x_0, y_0) = 0 \tag{17.5.6}$$

If

$$\frac{\partial h}{\partial x}(x_0, y_0) \neq 0 \tag{17.5.7}$$

we can solve for

$$\frac{dg}{dy}(y_0) = -\frac{\frac{\partial h}{\partial y}(x_0, y_0)}{\frac{\partial h}{\partial x}(x_0, y_0)} \tag{17.5.8}$$

and this is what is known as the implicit differentiation of h at (x_0, y_0). As we saw in §17.3, the equation of the tangent line to $h(x, y) = 0$ at (x_0, y_0) is:

$$\frac{\partial h}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial h}{\partial y}(x_0, y_0)(y - y_0) = 0.$$

The generalization of this result to several variables is called the *implicit function theorem*: see Rudin [55], p.223. We state it in Theorem 17.6.6 after setting up our notation in Remark 17.6.1.

Now we differentiate (17.5.5) with respect to y using the chain rule again, in order to get a formula for the second derivatives. We assume that both h and g are $\mathcal{C}^2$. This computation is a special case of that of Theorem 12.2.9, so we redo it from scratch. Write h_x and h_y for the first partials of h, and h_{xx}, h_{xy}, h_{yy} for the second partials. Also write g' and g'' for the derivatives of $g(y)$. Then the derivative of (17.5.5) is:

$$\begin{aligned}\Big(h_{xx}(g(y), y)g'(y) + h_{xy}(g(y), y)\Big)g'(y) + h_x(g(y), y)g''(y) \\ + h_{xy}(g(y), y)g'(y) + h_{yy}(g(y), y) = 0.\end{aligned}$$

This can be written in block matrix form (suppressing the points where the functions are evaluated) as:

$$\begin{bmatrix} g' & 1 \end{bmatrix} \begin{bmatrix} h_{xx} & h_{xy} \\ h_{xy} & h_{yy} \end{bmatrix} \begin{bmatrix} g' \\ 1 \end{bmatrix} + h_x g'' = 0. \tag{17.5.9}$$

Since $g' = -h_y/h_x$, we get (since $h_x \neq 0$), solving for the second derivative of g:

$$\begin{aligned} g'' = \frac{-1}{h_x} \begin{bmatrix} -\frac{h_y}{h_x} & 1 \end{bmatrix} \begin{bmatrix} h_{xx} & h_{xy} \\ h_{xy} & h_{yy} \end{bmatrix} \begin{bmatrix} -\frac{h_y}{h_x} \\ 1 \end{bmatrix} \\ = -\frac{1}{h_x^3}\big(h_x^2 h_{yy} - 2h_x h_y h_{xy} + h_y^2 h_{xx}\big). \end{aligned} \tag{17.5.10}$$

17.6 The Implicit Function Theorem

17.6.1 Remark. What follows is a generalization of the computation leading to (17.5.8): h becomes a function $\mathbf{h}$ from $\mathbb{R}^n$ to $\mathbb{R}^m$, $n > m.$ so it has m coordinates h_i which are functions of n variables $(x_1, x_2, \dots, x_n)$. The case examined above is $n = 2$, $m = 1$. We mimic the split (x, y) that we had above by writing the variables in two groups: the first group of m variables replacing x we write $\mathbf{x}_b$ (b for basic or bound) and the second group of $(n - m)$ variables replacing y, we write $\mathbf{x}_f$ (f for free).

Let us see what we need to generalize the one variable computation.

- For (17.5.4) we need a function $\mathbf{g}$ of $(n - m)$ variables $\mathbf{x}_f$. $\mathbf{g}(\mathbf{x}_f)$ has m coordinates g_i, $1 \leq i \leq m$ that we put into the first m slots of $\mathbf{h}$.
- For (17.5.5) we need to perform a chain rule computation.
- For (17.5.6) we evaluate at a point $\mathbf{x}^* = (\mathbf{x}_b^*, \mathbf{x}_f^*)$ at which $\mathbf{h}$ vanishes. The point $\mathbf{x}^*$ is the point of interest for the entire computation and has been given in advance.
- For (17.5.7) and (17.5.8), since we are now dealing with vector functions, we see that $\frac{\partial h}{\partial x}(x_0, y_0)$ needs to be replaced by the $m \times m$ matrix of partials we will call $\nabla_b \mathbf{h}$, defined by

$$\nabla_b \mathbf{h} = \left[\frac{\partial h_i}{\partial x_j}(\mathbf{x}^*) \right]$$

  where i, $1 \leq i \leq m$ is the row index and j is the column index, so the i-th row of $\nabla_b \mathbf{h}$ is the gradient of h_i with respect to $\mathbf{x}_b$. Throughout the whole computation, gradients will therefore be row vectors, leading to some peculiar-looking matrix multiplications.
- We let $\nabla_b \mathbf{h}^*$ denote the value of the gradient $\nabla_b \mathbf{h}$ at $\mathbf{x}^*$. In order to generalize the key (17.5.7), the matrix $\nabla_b \mathbf{h}^*$ must be invertible: division by the non-zero derivative is replaced by multiplication by the inverse matrix $(\nabla_b \mathbf{h}^*)^{-1}$.

Before dealing with the general case of the implicit function theorem, let's state and prove the easy linear case. Indeed, we already used this case in the study of the asymmetric form of linear optimization, when we looked for a basic submatrix of a $m \times n$ matrix of rank m.

First our notation. Let H be a $m \times n$ matrix of rank m, and $\mathbf{x}$ a n-vector of variables. We study the linear function $H\mathbf{x}$. After changing the order of the columns of H, if necessary, we may write $H = \begin{bmatrix} H_b & H_f \end{bmatrix}$ and $\mathbf{x} = (\mathbf{x}_b, \mathbf{x}_f)$, where H_b is a $m \times m$ matrix of rank m, H_f a $m \times (n-m)$ matrix, $\mathbf{x}_b$ a m vector, and $\mathbf{x}_f$ a $n-m$ vector. By block multiplication (see (**??**)),

$$H\mathbf{x} = H_b\mathbf{x}_b + H_f\mathbf{x}_f. \tag{17.6.2}$$

17.6.3 Theorem (The Linear Case of the IFT)**.** *Consider the collection of m equations $H\mathbf{x} = \mathbf{0}$. Since the $m \times m$ matrix H_b is invertible, then for every $\mathbf{x}_f \in \mathbb{R}^{n-m}$ there is a unique $\mathbf{x}_b \in \mathbb{R}^m$ such that*

$$\mathbf{x}_b = -H_b^{-1}H_f\mathbf{x}_f. \tag{17.6.4}$$

Proof. Indeed, since

$$H \begin{bmatrix} \mathbf{x}_b \\ \mathbf{x}_f \end{bmatrix} = H_b\mathbf{x}_b + H_f\mathbf{x}_f = \mathbf{0},$$

and H_b is invertible, we can solve for $\mathbf{x}_b$ by left multiplication by the inverse of H_b. □

Our goal is to generalize this to the nonlinear case. We cannot hope to do this over the entire vector space: we only hope for an open neighborhood of the point of interest $\mathbf{x}^*$. Furthermore (17.6.4) expresses the $\mathbf{x}_b$ variables as a function (which we call $\mathbf{g}$ in the general case) of the $\mathbf{x}_f$ variables. What will generalize is the computation of the gradient of $\mathbf{g}$ with respect to its variables $\mathbf{x}_f$: in the linear case, the gradient is clearly $-H_b^{-1}H_f$; in the general case it is given by (17.6.9).

We start by setting up our notation.

17.6.5 Notation. Let $\mathbf{h}(\mathbf{x}_b, \mathbf{x}_f)$ be a continuously differentiable ($\mathcal{C}^1$) function defined on an open set U in $\mathbb{R}^n$ and mapping to $\mathbb{R}^m$, $m < n$, where the variables are written as compound vectors $\mathbf{x}_b = (x_1, \dots, x_m)$ and $\mathbf{x}_f = (x_{m+1}, \dots, x_n)$. Write $\nabla_b\mathbf{h}$ for the $m \times m$ matrix of partials of $\mathbf{h}$ with respect to the first m variables $\mathbf{x}_b$. Thus the (i,j)-th entry of $\nabla_b\mathbf{h}$, $1 \le i, j \le m$, is

$$\frac{\partial h_i}{\partial x_j}(\mathbf{x}).$$

In the same way we define $\nabla_f\mathbf{h}$ as the $m \times (n-m)$ matrix of partials with respect to the remaining $n-m$ variables.

17.6.6 Theorem (The Implicit Function Theorem)**.** *Let $\mathbf{h}(\mathbf{x}_b, \mathbf{x}_f)$ be a $\mathcal{C}^1$ function from an open set U in $\mathbb{R}^n$ to $\mathbb{R}^m$, $m < n$. Suppose there is a point $(\mathbf{x}_b^*, \mathbf{x}_f^*) \in U$ satisfying $\mathbf{h}(\mathbf{x}_b^*, \mathbf{x}_f^*) = 0$ such that the $m \times m$ matrix*

$$\nabla_b\mathbf{h}(\mathbf{x}_b^*, \mathbf{x}_f^*) \textit{ is invertible.}$$

Then there exists a $\mathcal{C}^1$ map $\mathbf{g}(\mathbf{x}_f)$ from $\mathbb{R}^{n-m}$ to $\mathbb{R}^m$ such that

$$\mathbf{h}(\mathbf{g}(\mathbf{x}_f), \mathbf{x}_f) = \mathbf{0} \quad \textit{and} \quad \mathbf{g}(\mathbf{x}_f^*) = \mathbf{x}_b^*, \tag{17.6.7}$$

defined on a sufficiently small open set M in $\mathbb{R}^{n-m}$ containing $\mathbf{x}_f^*$. Furthermore if $\mathbf{h}(\mathbf{x})$ is $\mathcal{C}^k$ (and not just $\mathcal{C}^1$ as above) then $\mathbf{g}(\mathbf{x}_f)$ is $\mathcal{C}^k$.

We will not prove this important result, but we derive the following two corollaries.

17.6.8 Corollary. *Since the $m \times m$ matrix $\nabla_b \mathbf{h}(\mathbf{g}(\mathbf{x}_f^*), \mathbf{x}_f^*)$ is invertible, for $\mathbf{x}_f$ close enough to $\mathbf{x}_f^*$, the $m \times m$ matrix $\nabla_b \mathbf{h}(\mathbf{g}(\mathbf{x}_f), \mathbf{x}_f)$ remains invertible. Then the $m \times (n-m)$ matrix*

$$\nabla \mathbf{g}(\mathbf{x}_f) = \left[\frac{\partial g_i}{\partial x_{m+j}}(\mathbf{x}_f)\right]$$

of partials of the implicit function $\mathbf{g}(x_{m+1}, \dots, x_n)$ can be written as

$$\nabla \mathbf{g}(\mathbf{x}_f) = -\Big(\nabla_b \mathbf{h}\big(\mathbf{g}(\mathbf{x}_f), \mathbf{x}_f\big)\Big)^{-1} \nabla_f \mathbf{h}\big(\mathbf{g}(\mathbf{x}_f), \mathbf{x}_f\big) \tag{17.6.9}$$

In particular, evaluating at $\mathbf{x}_b^* = g(\mathbf{x}_f^*)$, we get

$$\nabla \mathbf{g}(\mathbf{x}_f^*) = -\Big(\nabla_b \mathbf{h}\big(\mathbf{g}(\mathbf{x}_f^*), \mathbf{x}_f^*\big)\Big)^{-1} \nabla_f \mathbf{h}\big(\mathbf{g}(\mathbf{x}_f^*), \mathbf{x}_f^*\big) \tag{17.6.10}$$

Proof. This is a chain rule computation on (17.6.7), similar to the one done in Theorem 12.2.3. Take the partial $\partial/\partial x_j$ of (17.6.7), where x_j is a free variable. We get

$$\sum_{k \in B} \frac{\partial \mathbf{h}}{\partial x_k}\big(\mathbf{g}(\mathbf{x}_f), \mathbf{x}_f)\big)\frac{\partial \mathbf{g}}{\partial x_j} + \frac{\partial \mathbf{h}}{\partial x_j}(\mathbf{g}(\mathbf{x}_f), \mathbf{x}_f) = 0. \tag{17.6.11}$$

So summing over the free variables we get

$$\nabla_b \mathbf{h}\big(\mathbf{g}(\mathbf{x}_f), \mathbf{x}_f\big)\nabla \mathbf{g}(\mathbf{x}_f) + \nabla_f \mathbf{h}\big(\mathbf{g}(\mathbf{x}_f), \mathbf{x}_f\big) = \mathbf{0}. \tag{17.6.12}$$

Conclude by inverting the matrix $\nabla_b \mathbf{h}(\mathbf{g}(\mathbf{x}_f), \mathbf{x}_f)$. □

17.6.13 Remark. We now have several ways of describing the tangent space of $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ at the regular point $\mathbf{x}^*$. The first uses Definition 17.3.1: it is the set of vectors perpendicular to the m vectors $\nabla h_i(\mathbf{x}^*)$, in other words the vectors perpendicular to the rows of the $m \times n$ matrix $\nabla \mathbf{h}(\mathbf{x}^*)$. By regularity we know that this matrix has rank m, so by reordering the columns, we may assume that it can be written in blocks as $\begin{bmatrix}A_b & A_f\end{bmatrix}$ where A_b is an invertible $m \times m$ matrix and A_f is a $m \times (n-m)$ matrix. If we write the n vector $\mathbf{x} = (\mathbf{x}_b, \mathbf{x}_f)$ as a m vector $\mathbf{x}_b$ followed by an $(n-m)$ vector $\mathbf{x}_f$, we get $A_b\mathbf{x}_b + A_f\mathbf{x}_f = \mathbf{0}$, so, inverting A_b, we get $\mathbf{x}_b = -A_b^{-1}A_f\mathbf{x}_f$, so that the tangent vectors can be written

$$\begin{bmatrix}-A_b^{-1}A_f \\ I_{n-m}\end{bmatrix}\mathbf{x}_f \tag{17.6.14}$$

for any choice of $(n-m)$ vector $\mathbf{x}_f$.

The second uses the definition of the tangent space of the graph of the implicit function $\mathbf{g}$ given in Definition 17.4.3: it is the set of vectors perpendicular to the rows of the $m \times n$ matrix

$$\begin{bmatrix}-I_m & \nabla\mathbf{g}|_{\mathbf{x}^*}\end{bmatrix} \tag{17.6.15}$$

The chain rule computation of (17.6.12) shows that the column vectors of the $n \times (n-m)$ matrix

$$\begin{bmatrix}\nabla\mathbf{g}|_{\mathbf{x}^*} \\ I_{n-m}\end{bmatrix} \tag{17.6.16}$$

are in the tangent space according to the first definition. As these columns are linearly independent, they form a basis for the tangent space - assuming, as always, that $\mathbf{x}^*$ is regular. By Corollary 17.6.8, this is the same matrix as the one in (17.6.14).

To close the loop, notice that the matrices (17.6.15) and (17.6.16) can be multiplied in blocks, giving

$$\begin{bmatrix}-I_m & \nabla\mathbf{g}|_{\mathbf{x}^*}\end{bmatrix}\begin{bmatrix}\nabla\mathbf{g}|_{\mathbf{x}^*} \\ I_{n-m}\end{bmatrix} = \nabla\mathbf{g}|_{\mathbf{x}^*} - \nabla\mathbf{g}|_{\mathbf{x}^*} = \mathbf{0}_{m,n-m},$$

the expected result according to the second definition.

We also get a formula for the Hessian of $\mathbf{g}$, in exactly the way we obtained (17.5.9) by applying Theorem 12.2.9. We treat a special case in §17.7, and handle the general case at the beginning of Lecture 29.

A good reference for the implicit function theorem and its history is Krantz-Parks [35]. The proof of the Implicit Function Theorem given in Rudin [55], chapter 9 is recommended.

17.7 Application to Level Sets

The Implicit Function Theorem is useful in understanding level sets of functions.

We start with a real-valued $\mathcal{C}^2$ function $f(\mathbf{x})$ of n variables, defined on an open neighborhood U of a point $\mathbf{x}^*$. Let $c = f(\mathbf{x}^*)$, and let L_c be the c-level set of f, namely

$$L_c = \{\mathbf{x} \in U \mid f(\mathbf{x}) = c\}, \tag{17.7.1}$$

so we have just one equation. We use the implicit function theorem to understand the structure of L_c near $\mathbf{x}^*$, when the gradient $\nabla f(x^*)$ is not the zero vector, so that $\mathbf{x}^*$ is not a critical point for f. Since there is only one equation, that is all we need to be able to apply the IFT at $\mathbf{x}^*$.

Without loss of generality assume $\partial f / \partial x_1(\mathbf{x}^*) \neq 0$, so x_1 can be used as the bound variable. We write $\mathbf{x}^* = (x_1^*, \mathbf{x}_f^*)$, where $\mathbf{x}_f^*$ denotes the free variables x_2, $\dots$, x_n.

We can write x_1 as an implicit function $g(\mathbf{x}_f)$ in a neighborhood of $\mathbf{x}_f^*$. By Corollary 17.6.8, setting $\partial f / \partial x_1(\mathbf{x}^*) = a \neq 0$, we have

$$\nabla g(\mathbf{x}_f^*) = -\frac{\nabla_f f(\mathbf{x}^*)}{\partial f / \partial x_1(\mathbf{x}^*)}.$$

Next we compute the Hessian G of g at $\mathbf{x}^*$. From Theorem 12.2.9, we get

$$\begin{bmatrix} \nabla g^T & I_{n-1} \end{bmatrix} \begin{bmatrix} F_{11} & F_{12} \\ F_{12}^T & F_{22} \end{bmatrix} \begin{bmatrix} \nabla g \\ I_{n-1} \end{bmatrix} + \frac{\partial f}{\partial x_1}(\mathbf{x}^*) G = 0 \tag{17.7.2}$$

where we have suppressed the point where the gradients and the Hessians are evaluated. Here ∇g is a row vector with $n-1$ entries. F, the $n \times n$ Hessian of f, is broken into square diagonal blocks F_{11} of size 1 (corresponding to the bound variable), and F_{22} of size $n-1$ (corresponding to the free variables), and F_{12} is a row vector with $n-1$ entries. You should convince yourself that the sizes are appropriate for block multiplication.

The $(n-1) \times (n-1)$ symmetric matrix

$$F_{\perp}(\mathbf{x}) = \begin{bmatrix} \nabla g(\mathbf{x}_f)^T & I_{n-1} \end{bmatrix} \begin{bmatrix} F_{11}(\mathbf{x}) & F_{12}(\mathbf{x}) \\ F_{12}(\mathbf{x})^T & F_{22}(\mathbf{x}) \end{bmatrix} \begin{bmatrix} \nabla g(\mathbf{x}_f) \\ I_{n-1} \end{bmatrix} \tag{17.7.3}$$

is interesting. In Definition 17.4.3 we saw that the $n-1$ columns of the $n \times (n-1)$ matrix

$$\begin{bmatrix} \nabla g(\mathbf{x}) \\ I_{n-1} \end{bmatrix}$$

are the generators of the tangent space of the graph of g (or of f: it is the same) at the point $\mathbf{x}_f$. This means that $F_{\perp}(\mathbf{x})$ is the restriction of the Hessian of f at $\mathbf{x}^*$ to the tangent space of the level set at $\mathbf{x}$, or, which is the same, to the orthogonal complement of $\nabla f(\mathbf{x})$, for any $\mathbf{x}$ on the level set close enough to $\mathbf{x}^*$.

17.8 Examples of the IFT

In order to understand the Implicit Function Theorem, it is important to see how it extends the linear case discussed in Theorem 17.6.3. It is also important to work out some examples where the implicit function $\mathbf{g}$ cannot be found explicitly.

17.8.1 Example. We take $n = 3$ and $m = 1$. We consider

$$h(x_1, x_2, x_3) = x_1^2 + x_1x_2 + x_1x_3 + x_2^2 + x_2x_3 + x_3^2 - 25 \tag{17.8.2}$$

near the point $\mathbf{x}^* = (1, 2, 3)$. Note that $h(1, 2, 3) = 0$ as required.

The gradient of h is

$$\nabla h = (2x_1 + x_2 + x_3, x_1 + 2x_2 + x_3, x_1 + x_2 + 2x_3)$$

so evaluated at $\mathbf{x}^*$ we get

$$\nabla h^* = (7, 8, 9).$$

Thus the equation of the tangent plane at $\mathbf{x}^*$ is

$$7(x_1 - 1) + 8(x_2 - 2) + 9(x_3 - 3) = 0 \tag{17.8.3}$$

Here $\mathbf{x}_b = x_1$, $\mathbf{x}_f = (x_2, x_3)$, so $\nabla_b h^* = 7 \neq 0$ so that we can apply the IFT. Furthermore $\nabla_f h^* = (8, 9)$. So (17.6.10) gives:

$$\nabla \mathbf{g}^* = -\frac{1}{7}(8, 9) \tag{17.8.4}$$

The linear approximation of the implicit function g is given by the tangent plane 17.8.3. If you solve for x_1 in the tangent plane, and take the gradient with respect to x_2 and x_3 you get (17.8.4).

17.8.5 Example. We take $n = 3$ and $m = 2$. We add to (17.8.2), which we now call h_1, a second

$$h_2(x_1, x_2, x_3) = 36x_1^2 + 9x_2^2 + 4x_3^2 - 108$$

So

$$\nabla \mathbf{h}^* = \begin{bmatrix} 7 & 8 & 9 \\ 72 & 36 & 24 \end{bmatrix} \quad \text{and } \nabla_b \mathbf{h}^* = \begin{bmatrix} 7 & 8 \\ 72 & 36 \end{bmatrix}$$

and the determinant of $\nabla_b \mathbf{h}^* = -180 \neq 0$ so we can apply the IFT. The inverse of $\nabla_b \mathbf{h}^*$ is

$$(\nabla_b \mathbf{h}^*)^{-1} = -\frac{1}{180} \begin{bmatrix} 36 & -8 \\ -72 & 7 \end{bmatrix}$$

Note that the linear approximation of $\mathbf{h}$ at $(1,2,3)$ is given by

$$\begin{bmatrix} 7 & 8 & 9 \\ 72 & 36 & 24 \end{bmatrix} \begin{bmatrix} x_1 - 1 \\ x_2 - 2 \\ x_3 - 3 \end{bmatrix} = 0 \tag{17.8.6}$$

Now (17.6.10) gives:

$$\nabla \mathbf{g}^* = -\frac{1}{180} \begin{bmatrix} 36 & -8 \\ -72 & 7 \end{bmatrix} \begin{bmatrix} 9 \\ 24 \end{bmatrix} \tag{17.8.7}$$

which does result from (17.8.6)

17.9 Extensions of the IFT

Here is a typical extension of the IFT, used in the proof of the Lagrange Multiplier Theorem in §28.6.

In $\mathbb{R}^n$, suppose given a collection of $m < n$ function $\mathbf{h}(\mathbf{x})$, and a point $\mathbf{x}^*$ in the set $F = \{\mathbf{x} \mid \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$. Consider the $m \times n$ matrix A of partials of $\mathbf{h}(\mathbf{x})$ evaluated at $\mathbf{x}^*$. Assume it has maximal rank m. As always, after reordering the columns of A, which simply amounts to changing the labels of the variables, we may assume that the square submatrix formed by the first m columns of A is invertible. Now apply the IFT as stated. The tangent space $T_{F,\mathbf{x}^*}$ to F at $\mathbf{x}^*$, defined in 17.3.1 is the linear space of $\mathbf{v}$ satisfiying

$$A\mathbf{v} = \mathbf{0}. \tag{17.9.1}$$

17.9.2 Corollary. *Assume the IFT applies to the set F at $\mathbf{x}^*$, where $\mathbf{h}$ is $\mathcal{C}^1$. For any nonzero vector $\mathbf{v}$ in the tangent space $T_{F,\mathbf{x}^*}$, there is a parametrized curve $\mathbf{x}(t)$ defined on a small enough interval I given by $-\epsilon \le t < \epsilon$ for some $\epsilon > 0$, satisfying*

1. $\mathbf{x}(0) = \mathbf{x}^*$.
2. $\dot{\mathbf{x}}(0) = \mathbf{v}$, *where $\dot{x}$ denotes the derivative with respect to t.*
3. $\mathbf{x}(t)$ *is $\mathcal{C}^1$.*
4. $\mathbf{x}(t) \in F$, *for all $t \in I$. In other words,* $\mathbf{h}(\mathbf{x}(t)) = \mathbf{0}$.

Proof. Take the line $\mathbf{x}_f(t)$ in $\mathbb{R}^{n-m}$ given by $\mathbf{x}_f(t) = \mathbf{x}_f^* + \mathbf{v}t$. It satisfies

$$\mathbf{x}_f(0) = \mathbf{x}_f^* \text{ and } \dot{\mathbf{x}}_f(0) = \mathbf{v}. \tag{17.9.3}$$

Then, using the implicit function $\mathbf{g}$ associated to $\mathbf{h}$ at $\mathbf{x}^*$, we get a curve in $\mathbb{R}^n$ by taking for the first m coordinates $\mathbf{g}(\mathbf{x}_f(t))$ and for the last $n-m$ coordinates $\mathbf{x}_f(t)$. In other words we are taking the graph of $\mathbf{g}$ restricted to the curve $\mathbf{x}_f(t)$. By the IFT we have:

$$\mathbf{h}\Big(\mathbf{g}\big(\mathbf{x}_f(t)\big), \mathbf{x}_f(t)\big)\Big) = \mathbf{0},$$

so that the parametrized curve

$$\mathbf{x}(t) = \big(\mathbf{g}(\mathbf{x}_f(t)), \mathbf{x}_f(t)\big)$$

lies on F as required. The remaining conditions are immediate from (17.9.3) and the chain rule. □

The following extension will be used in §31.2.

17.9.4 Corollary. *In addition to the hypotheses of Corollary 17.9.2, assume for convenience that coordinates have been chosen so* $\mathbf{x}_b^* = \mathbf{0}$*. Also assume that the implicit function* $\mathbf{g}(\mathbf{x})$ *satisfies* [2]

$$\nabla\mathbf{g}(x^*)\mathbf{v} \prec \mathbf{0}.$$

Then the curve $\mathbf{x}(t)$ *can be chosen so that, on a (perhaps smaller) interval* I*, it satisfies not only the conclusions of Corollary 17.9.2 but also*

$$\mathbf{g}(\mathbf{x}(t)) \prec \mathbf{0}, \textit{for all } t \in I.$$

Proof. We compute the derivative of $\mathbf{g}(\mathbf{x}(t))$ at $t=0$ by the chain rule, and get $\nabla\mathbf{g}(x^*)\mathbf{v} \prec \mathbf{0}$ by hypothesis. Thus each one of the coordinate functions of $\mathbf{g}(\mathbf{x}(t))$ is strictly decreasing at $t=0$, so for t positive and small enough, since $\mathbf{g}(\mathbf{x}(0)) = \mathbf{0}$, we have $\mathbf{g}(\mathbf{x}(t)) \prec \mathbf{0}$ as required. □

17.10 The Contraction Principle

17.10.1 Definition. Let X be a finite dimensional vector space over $\mathbb{R}$, together with its distance function d (see Definition 5.4.1). Let φ be a mapping from X to itself. If there exists a real number $c<1$ such that

$$d(\varphi(\mathbf{x}), \varphi(\mathbf{y})) \leq cd(\mathbf{x}, \mathbf{y}) \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbf{X}, \tag{17.10.2}$$

then φ is a *contraction* of $\mathbf{X}$.

[2] Recall that the $\prec$ sign in the equation means that for every index k, $1 \leq k \leq m$, $\nabla g_k(x^*)\mathbf{v} < 0$.

17.10.3 Proposition. *Contractions are continuous.*

Proof. According to Definition 16.1.1 we have to show that for every ϵ-neighborhood 14.4.1 $N_\epsilon(\varphi(\mathbf{a}))$ of $\varphi(\mathbf{a}) \in X$ there is a δ-neighborhood $N_\delta(\mathbf{a})$ of $\mathbf{a}$ such that

$$\varphi(N_\delta(\mathbf{a})) \subset N_\epsilon(\varphi(\mathbf{a}))$$

It is clear that for a contraction $\delta = \epsilon$ will do the trick. □

17.10.4 Definition. Let φ be a mapping from X to itself, and $\mathbf{x}$ a point of X. Then $\mathbf{x}$ is a *fixed point* of φ if $\varphi(\mathbf{x}) = \mathbf{x}$.

17.10.5 Theorem. *If X is a finite dimensional vector space over $\mathbb{R}$ and φ a contraction of X, then φ has a unique fixed point $\mathbf{x}$.*

Proof. First we prove uniqueness by contradiction. If $\mathbf{x}$ and $\mathbf{y}$ are distinct fixed points and φ a contraction, then by (17.10.2)

$$d(\mathbf{x},\mathbf{y}) = d(\varphi(\mathbf{x}),\varphi(\mathbf{y})) \le cd(\mathbf{x},\mathbf{y}) < d(\mathbf{x},\mathbf{y})$$

an impossibility unless $d(\mathbf{x},\mathbf{y}) = 0$ by the definition of a distance function (see 5.4.1).

Next we prove existence by constructing a sequence $\{\mathbf{x}_n\}$ converging to the fixed point. We start with an arbitrary point $\mathbf{x}_0 \in X$, and define $\mathbf{x}_n = \varphi(\mathbf{x}_{n-1})$ for any n. So we have a sequence. We show it is a Cauchy sequence (see Definition 15.2.1).

Indeed, by construction

$$d(\mathbf{x}_n,\mathbf{x}_{n-1}) = d(\varphi(\mathbf{x}_{n-1}),\varphi(\mathbf{x}_{n-2})) \le cd(\mathbf{x}_{n-1},\mathbf{x}_{n-2}) \le \cdots \le c^n d(\mathbf{x}_1,\mathbf{x}_0)$$

so for $n < m$

$$\begin{aligned} d(\mathbf{x}_n,\mathbf{x}_m) &\le \sum_{i=n+1}^{m} d(\mathbf{x}_i,\mathbf{x}_{i-1}) \\ &\le \Big(\sum_{i=n+1}^{m} c^{i-1}\Big) d(\mathbf{x}_1,\mathbf{x}_0 \\ &\le \frac{c^n}{1-c} d(\mathbf{x}_1,\mathbf{x}_0). \end{aligned}$$

Thus it converges by Theorem 15.2.2 and we are done. □

Part VI

Convexity and Optimization

Lecture 18

Convex Sets

Why do we study convexity in optimization? Calculus helps us find local extrema, but we are really interested in finding global extrema, a harder problem. When the objective function is convex or concave, finding global extrema is an easier task. In economics and applied science, meanwhile, it is often reasonable to assume that the objective function is convex (or concave, depending on the situation). The domain of definition of a convex or concave function is a convex set, so we start by studying convex sets.

The core of the lecture is the three named theorems on convex sets :

- Carathéodory's Theorem 18.5.1,
- The Separating Hyperplane Theorem for disjoint convex sets 18.6.8,
- Minkowski's Theorem 18.7.1.

The basic definitions and the key theorems for convex sets are given in §18.1. The main thrust is to show that the convex hull of a set is the same as the set of convex combinations of points in the set: this is the content of Theorem 18.1.28. Other than the many examples, all the intermediate results in this section build up to this theorem.

In §18.2 we study affine geometry, a slight generalization of linear geometry that is useful in understanding convex geometry. In particular, it allows us to define the dimension of a convex set. Then in §18.3 we study the two key examples we will need for optimization theory: polytopes and polyhedra. Next in §18.4 we study the topology of convex sets. We are especially interested in knowing when a convex set is closed, and when it is compact. Theorem 18.4.9 is important.

The lecture concludes (§18.8) with an application to permutation matrices and doubly stochastic matrices. This section will not be used later in the course.

The standard (encyclopedic) reference for convexity is Rockafellar's treatise [53]. I quite like [10]. Other sources for convexity applied to optimization are [5], [7], and [22]. An excellent reference for convex sets is Barvinok [4]. A more elementary approach, with applications to optimization, is given in Lay [40].

18.1 The Convex Hull and Convex Combinations

18.1.1 The Convex Hull

Take two distinct points $\mathbf{p}$ and $\mathbf{q}$ in $\mathbb{R}^n$. There is a unique straight line L passing through both of them. By extension of the notation in $\mathbb{R}$ we denote $[\mathbf{p}, \mathbf{q}]$ and $(\mathbf{p}, \mathbf{q})$ the closed and open segments of points on L bounded by $\mathbf{p}$ and $\mathbf{q}$. The points $\mathbf{r}$ of $(\mathbf{p}, \mathbf{q})$ can be parametrized by

$$\mathbf{r} = \lambda \mathbf{p} + (1-\lambda)\mathbf{q}, \text{ for } \lambda \in \mathbb{R}, 0 < \lambda < 1. \tag{18.1.1}$$

18.1.2 Definition. A point $\mathbf{r}$ is *between* $\mathbf{p}$ and $\mathbf{q}$ if it satisfies (18.1.1), so that it is in the open segment $(\mathbf{p}, \mathbf{q})$.

18.1.3 Definition. A set S in $\mathbb{R}^n$ is *convex* if for every pair of points $\mathbf{p}$ and $\mathbf{q}$ in S, every point between $\mathbf{p}$ and $\mathbf{q}$ is in S.

In other words, if two points are in S, then any point of the open segment joining the two points is in S.

Before giving some examples of convex sets, we make some definitions and set up some notation that will be used throughout this course.

18.1.4 Definition. An *affine hyperplane* in $\mathbb{R}^n$ is the set of points $\mathbf{x} = (x_1, \dots, x_n)$ satisfying an equation $\sum_{i=1}^n a_i x_i = c$. We can rewrite this by thinking of the coefficients $(a_1, \dots, a_n)$ as the coordinates of a vector $\mathbf{a}$, so the equation becomes, using the inner product:

$$\langle \mathbf{a}, \mathbf{x} \rangle = c. \tag{18.1.5}$$

We denote this hyperplane by $H_{\mathbf{a},c}$. If $\mathbf{a} = \mathbf{0}$ this equation is uninteresting, indeed contradictory if $c \neq 0$, so we assume $\mathbf{a}$ is non-zero. The vector $\mathbf{a}$ is called the *normal* to the hyperplane. By dividing the equation by the length of $\mathbf{a}$, we may, if we wish, assume that $\mathbf{a}$ has length 1. This still does not prescribe $\mathbf{a}$ uniquely: you can still multiply the equation by -1.

18.1.6 Definition. Start with the hyperplane (18.1.5). Then the two closed half-spaces associated to this hyperplane are:

$$H^+_{\mathbf{a},c} = \{\mathbf{x} \mid \langle \mathbf{a}, \mathbf{x} \rangle \geq c\}, \text{ and } H^-_{\mathbf{a},c} = \{\mathbf{x} \mid \langle \mathbf{a}, \mathbf{x} \rangle \leq c\}.$$

Note that $H^+_{\mathbf{a},c}$ is the half-space the normal vector $\mathbf{a}$ points into. The open half-spaces: $\mathring{H}^+_{\mathbf{a},c}$ and $\mathring{H}^-_{\mathbf{a},c}$ are obtained by replacing $\geq$ (resp.$\leq$) by $>$ (resp. $<$) in the definitions of the closed half-spaces. The hyperplane $H_{\mathbf{a},c}$ is called the *face* of all these half-spaces.

18.1.7 Example. The following sets are convex:
The empty set;[1]
A point;
A line or a segment on a line;
Any affine hyperplane $H_{\mathbf{a},c}$;
More generally, any linear space;
A closed or open half-space.

18.1.8 Example. We show that every closed ball is convex. Change coordinates so that its center is at the origin. Then the closed ball $\overline{N}_r(\mathbf{0})$ is just the set of points $\mathbf{x} \in \mathbb{R}^n$ such that $\|\mathbf{x}\| \leq r$. Given two points $\mathbf{p}$ and $\mathbf{q}$ such that $\|\mathbf{p}\| \leq r$ and $\|\mathbf{q}\| \leq r$, we must show that $\|\lambda\mathbf{p} + (1-\lambda)\mathbf{q}\| \leq r$ for all λ, $0 < \lambda < 1$. By the triangle inequality we have

$$\|\lambda\mathbf{p} + (1-\lambda)\mathbf{q}\| \leq \lambda\|\mathbf{p}\| + (1-\lambda)\|\mathbf{q}\| \leq \lambda r + (1-\lambda)r = r,$$

so we are done.

18.1.9 Exercise. Prove that closed half-spaces are closed sets and open half-spaces are open sets.

18.1.10 Definition. A point $\mathbf{r}$ of a convex set S is an *extreme point* of S if it is not between two points of S.

In other words, one cannot find distinct points $\mathbf{p}$ and $\mathbf{q}$ in S so that (18.1.1) is satisfied. Extreme points are very useful in solving optimization problems: see for example Theorem 22.4.10.

18.1.11 Example. Let T be the closed region in $\mathbb{R}^2$ bounded by a triangle. Convince yourself T is convex. The extreme points of T are the vertices of the triangle.

18.1.12 Remark. An extreme point of C must be a boundary point of C, but a boundary point need not be an extreme point. Indeed if you take T as above, only the vertices are extreme points, but the edges of the triangle are boundary points.

[1] In a few texts, the empty set is not taken to be convex: see for example [7], p. 36. The majority of references say that the empty set is convex: [4], [10], [22], [33], [40], [52], [53], [66]. This is simply a matter of convention.

18.1.13 Example. The extreme points of the closed ball $\overline{N}_r(\mathbf{p})$ in $\mathbb{R}^n$ are all the points of the boundary, namely the $(n-1)$-sphere $S_r(\mathbf{p})$.

18.1.14 Exercise. If you remove an extreme point from a convex set, what remains is convex. Conversely, if you remove a point from a convex set, and the remainder is convex, the removed point was an extreme point. Combining Example 18.1.13 and this exercise, we see that open balls are convex.

18.1.15 Theorem. *The intersection of any collection of convex sets in $\mathbb{R}^n$ is convex.*

Proof. Let C_α, $\alpha \in I$, be such a collection, where the index set I may be infinite. If the intersection is empty, we are done; if there is just one point in the intersection, likewise. So take any two points $\mathbf{p}$ and $\mathbf{q}$ in the intersection. For every $\alpha \in I$, the segment $[p, q]$ is in C_α, so it is in the intersection, which is therefore convex. □

18.1.16 Exercise. Show that to prove that a point $\mathbf{p}$ is an extreme point of the convex set S, it is enough to show that it is an extreme point of the intersection of S with a ball $N_r(\mathbf{p})$ of arbitrarily small radius $r > 0$.

18.1.17 Definition. The *convex hull* of a set $S \in \mathbb{R}^n$ is the intersection of all convex sets containing S. It is denoted $\operatorname{Co} S$.

18.1.18 Corollary. *The convex hull of any set S is convex.*

18.1.19 Example. The convex hull of a regular polygon in $\mathbb{R}^2$ is convex; so are n-cells. See Definition 14.3.1.

18.1.20 Exercise. Find the convex hull of the set of points (x, y) in $\mathbb{R}^2$ satisfying $x^2 + y^2 = 1$ and $x \leq 0$. Draw the picture.

18.1.2 Convex Combinations

18.1.21 Definition. Let $\mathbf{x}^1, \dots, \mathbf{x}^r$, be a collection of r points in $\mathbb{R}^n$, where r is any positive integer. Then $\mathbf{x}$ is a *convex combination* of the points $\mathbf{x}^i$ if there exist non-negative real numbers λ_i, $\sum_{i=1}^r \lambda_i = 1$ such that

$$\mathbf{x} = \sum_{i=1}^{r} \lambda_i \mathbf{x}^i \tag{18.1.22}$$

18.1.23 Exercise.

- if $\mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^2$ are three distinct, non-aligned points in the plane, then the set of convex combinations of $\mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^2$ is the triangle and the inside of the triangle formed by the three points.

- if $\mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3$ are four distinct points in $\mathbb{R}^3$, such that any three span a plane, and the four points do not lie in a plane, then the set of convex combinations of $\mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3$ is a tetrahedron[2] and its interior.

18.1.24 Theorem (The Convex Combinations Theorem). *A set S is convex if and only if all finite convex combinations of points of S are in S.*

Proof. By definition, S is convex if convex combinations of two points of S are in S. So half of the theorem is clear, and we only need to show that a convex combination of r points of a convex set S is in S, for any $r \geq 2$. We do this by induction on r. We start the induction at $r = 2$: this is the definition of convexity, so there is nothing to do.

Next we assume that the result is known for $r \geq 2$, namely that any convex combination of r points is in S, and we prove it for $r + 1$. Let $\mathbf{x}^1, \dots, \mathbf{x}^{r+1}$ be $r + 1$ arbitrary points of S, and let

$$\mathbf{x} = \sum_{i=1}^{r+1} \lambda_i \mathbf{x}^i \text{ , where all } \lambda_i \geq 0 \text{ and } \sum_{i=1}^{r+1} \lambda_i = 1.$$

We need to show $\mathbf{x} \in S$. We may assume that $\lambda_i > 0$ for all i, since otherwise there is nothing to prove since there are only r terms. Let $\gamma = \sum_{i=1}^{r} \lambda_i$, so by the last remark $0 < \gamma < 1$. Then let

$$\gamma_i = \lambda_i / \gamma \text{ , } 1 \leq i \leq r,$$

so that the point $\mathbf{y} = \sum_{i=1}^{r} \gamma_i \mathbf{x}^i$ is a convex combination of r points of S, and is therefore in S by induction. Then $\mathbf{x} = \gamma \mathbf{y} + \lambda_{r+1} \mathbf{x}^{r+1}$, and $\gamma + \lambda_{r+1} = 1$, so $\mathbf{x}$ is a convex combination of two points of S and is therefore in S, since S is convex. □

18.1.25 Definition. For any set S, let $K(S)$ be the set of all finite convex combinations of points of S.

By taking just one point in the convex combination, so $r = 1$ and $\lambda_1 = 1$, we see that $S \subset K(S)$. When S is empty, $K(S)$ is empty.

18.1.26 Theorem. *For any set S, $K(S)$ is a convex set.*

Proof. To show that $K(S)$ is convex, we need to show that if $\mathbf{k}_1$ and $\mathbf{k}_2$ are points of $K(S)$ then for any λ, $0 \leq \lambda \leq 1$, $\lambda \mathbf{k}_1 + (1 - \lambda)\mathbf{k}_2$ is in $K(S)$. Since $\mathbf{k}_1$ is a

[2] If you do not remember what a tetrahedron is, you can use this as a definition.

convex combination of points of S, we have

$$\mathbf{k}_1 = \sum_{i=1}^{n} \mu_i \mathbf{x}_i, \text{ for } \mu_i \geq 0\ ,\ \sum_{i=1}^{n} \mu_i = 1,$$

and similarly for $\mathbf{k}_2$:

$$\mathbf{k}_2 = \sum_{j=1}^{m} \nu_j \mathbf{y}_j, \text{ for } \nu_j \geq 0\ ,\ \sum_{j=1}^{m} \nu_j = 1,$$

where the $\mathbf{x}_i$ and $\mathbf{y}_j$ are all in S. Then

$$\lambda \mathbf{k}_1 + (1-\lambda)\mathbf{k}_2 = \sum_{i=1}^{n} \lambda \mu_i \mathbf{x}_i + \sum_{j=1}^{m} (1-\lambda)\nu_j \mathbf{y}_j. \tag{18.1.27}$$

To show that the right-hand side is a convex combination of the $n+m$ points $\{x_i\}$ and $\{y_j\}$ we need to show that all the coefficients in (18.1.27) are non-negative, which is easy, and that they sum to 1, which we check:

$$\sum_{i=1}^{n} \lambda \mu_i + \sum_{j=1}^{m} (1-\lambda)\nu_j = \lambda \sum_{i=1}^{n} \mu_i + (1-\lambda) \sum_{j=1}^{m} \nu_j = \lambda + 1 - \lambda = 1,$$

so this is in $K(S)$. □

18.1.28 Theorem. *For any set S, the convex hull is equal to the set of convex combinations:* $\operatorname{Co} S = K(S)$.

Proof. By Theorem 18.1.26 $K(S)$ is convex, and it contains S. Since CoS is the intersection of all convex sets containing S, we have:

$$CoS \subset K(S)$$

To get the opposite inclusion, take a convex combination $\sum_{i=1}^{r} \lambda_i \mathbf{x}_i$ of elements $\mathbf{x}_i$ of S, and an arbitrary convex set T containing S. All we need to do is show that this convex combination is in T. Since the $\mathbf{x}_i$ are in S, they are in T, and Theorem 18.1.24 shows that all convex combinations of points of T are in T, so we are done. □

An immediate corollary of this theorem is that any point in the convex hull of a set S can be written as a finite convex combination of points in S. We will improve this in Carathéodory's Theorem 18.5.1, which says that if S is in $\mathbb{R}^n$, we only need convex combinations with at most $n + 1$ points. You can read the proof of that theorem now, if you want.

Since we no longer need to make the distinction between the convex hull and the set of all convex combinations, in both cases we write $K(S)$ and refer to it as the convex hull.

18.1.29 Theorem. *Let $T\colon V \to W$ be a linear transformation between two vector spaces V and W. Let S be a convex set in V. Then its image $T(S)$ under T is convex in W.*

Proof. Take any two points $\mathbf{p}$ and $\mathbf{q}$ in $T(S)$. We must show that for any λ, $0 < \lambda < 1$, $\lambda\mathbf{p} + (1 - \lambda)\mathbf{q}$ is in $T(S)$. By definition of $T(S)$, there is a $\mathbf{a} \in S$ such that $T(\mathbf{a}) = \mathbf{p}$ and a $\mathbf{b} \in S$ such that $T(\mathbf{b}) = \mathbf{q}$. Since S is convex, for our choice of λ, $\lambda\mathbf{a} + (1 - \lambda)\mathbf{b}$ is in S. By linearity of T,

$$T(\lambda\mathbf{a} + (1 - \lambda)\mathbf{b}) = \lambda T(\mathbf{a}) + (1 - \lambda)T(\mathbf{b}) = \lambda\mathbf{p} + (1 - \lambda)\mathbf{q},$$

which is therefore in $T(S)$, as required. □

18.1.30 Example. Ellipsoids are convex.

Proof. We can move the center of the ellipsoid to the origin, so it is written as the set of $\mathbf{x}$ satisfying $\mathbf{x}^T A^{-1}\mathbf{x} \leq 1$, where A is a (symmetric) $n \times n$ positive-definite matrix. A has a symmetric square root R, which is also an invertible $n \times n$ matrix. By Proposition 13.7.7, the ellipsoid is the image of the closed unit ball under the invertible linear transformation given by R. Since the ball is convex, its image under R is convex by Theorem 18.1.29, so the ellipsoid is convex. □

18.1.31 Definition. If S and T are non-empty subsets of $\mathbb{R}^n$, and a and b are fixed real numbers, then the *Minkowski sum* of S and T with coefficients a and b, written $aS + bT$, is

$$aS + bT := \{a\mathbf{s} + b\mathbf{t} \mid \forall\mathbf{s} \in S, \forall\mathbf{t} \in T\}.$$

If T is empty, then $aS + bT := aS$. Similarly, if S is empty, $aS + bT := bT$.

18.1.32 Proposition. *If S and T are convex, then so is the Minkowski sum $aS+bT$, for any choice of a and b.*

Proof. Pick two points $a\mathbf{s}_1 + b\mathbf{t}_1$ and $a\mathbf{s}_2 + b\mathbf{t}_2$ in $aS + bT$. We must show that for any λ, $0 < \lambda < 1$,

$$\lambda(a\mathbf{s}_1 + b\mathbf{t}_1) + (1 - \lambda)(a\mathbf{s}_2 + b\mathbf{t}_2)$$

is in $aS + bT$. This can be written

$$a(\lambda \mathbf{s}_1 + (1-\lambda)\mathbf{s}_2) + b(\lambda \mathbf{t}_1 + (1-\lambda)\mathbf{t}_2)$$

and since S and T are both convex, this is in $aS + bT$. □

18.1.33 Exercise. Let S be a convex set in the plane with coordinates x and y. Assume S contains an entire line L. For simplicity, and without loss of generality, let L be the line with equation $y = 0$, namely the x-axis. What are all the possibilities for S?

Hint: S could be just the line L, or the entire plane, or the upper half-plane $y \geq 0$, or the lower half-plane $y \leq 0$. In order to analyze the remaining cases, assume that S only contains points in the upper half-plane. Assume that it contains a point $\mathbf{p}$ with second coordinate $y = a$, for some $a > 0$. Then show, by connecting $\mathbf{p}$ to points on the lines with very large and very small x coordinates, that S contains the entire strip of points (x, y) with $0 \leq y < a$. Finally let b be the greatest lower bound of y-coordinates of points in the upper half-plane that are not in S. Note that b is greater than or equal to any a found previously. Then show that S is contained in the strip of points (x, y) with $0 \leq y \leq b$. Then what can you say?

18.1.34 Exercise. If S is the closed ball of radius r_1 centered at $\mathbf{c}_1$, and T the closed ball of radius r_2 centered at $\mathbf{c}_2$, then $S + T$ is the closed ball B of radius $r_1 + r_2$ centered at $\mathbf{c}_1 + \mathbf{c}_2$.

Hint: First show that $S + T \subset B$, because every point in $S + T$ is at most at distance $r_1 + r_2$ from $\mathbf{c}_1 + \mathbf{c}_2$. Then show the opposite inclusion, by writing every point of the boundary of B as the sum of points from S and T. Make a picture in $\mathbb{R}^2$.

18.2 Affine Geometry

An intermediate geometry between linear geometry and convex geometry is affine geometry. The language of affine geometry and affine sets clarifies the concepts of convex geometry. The exposition parallels that of convex sets, to emphasize the connection. To study convexity we consider collections of *positive* real numbers λ_i such that $\sum \lambda_i = 1$. To study affine geometry we just drop the hypothesis that the λ_i are positive. That is the only change. Later on, in §19.2 we will study a third geometry on the same model: conical geometry. The λ_i will be positive, but we drop the hypothesis that their sum is 1. Finally linear algebra can be viewed as the case where there is no requirement on the λ_i.

A more detailed exposition of affine geometry is given in [33], §2.4. In Roberts and Varberg [52], §3, the theory of convex and affine sets is developed in even closer parallel than here.

18.2.1 Definition. A set S in $\mathbb{R}^n$ is *affine* if for every pair of points $\mathbf{p}$ and $\mathbf{q}$ in S and every real number λ, the point $\lambda\mathbf{p} + (1-\lambda)\mathbf{q}$ is in S.

In other words, if two points are in S, then any point on the line joining the two points is in S.

18.2.2 Example (First Examples of Affine Sets)**.** Most of the convex sets in Definition 18.1.7 are affine. More generally let A be an $m \times n$ matrix, and let H be the set of solutions $\mathbf{x}$ of $A\mathbf{x} = \mathbf{b}$, where $\mathbf{b}$ is a fixed m-vector. Again, you should convince yourself that this space is affine. Note that H could be empty if $\mathbf{b}$ is not in the range of A.

As we will see soon, this is essentially the complete list of affine subspaces in $\mathbb{R}^n$.

18.2.3 Theorem. *The intersection of any collection of affine sets in $\mathbb{R}^n$ is affine.*

Proof. Use the proof of Theorem 18.1.15. □

18.2.4 Definition. The *affine hull* of a set $S \in \mathbb{R}^n$ is the intersection of all affine sets containing S.

18.2.5 Corollary. *The affine hull of a set S is affine.*

18.2.6 Definition. Let $\mathbf{x}_1, \dots, \mathbf{x}_r$ be a collection of r points in $\mathbb{R}^n$, where r is any positive integer. Then $\mathbf{x}$ is an *affine combination* of the points $\mathbf{x}_i$ if there exists real numbers λ_i, $\sum_{i=1}^r \lambda_i = 1$, such that

$$\mathbf{x} = \sum_{i=1}^{r} \lambda_i \mathbf{x}_i$$

18.2.7 Theorem (The Affine Combinations Theorem)**.** *A set S is affine if and only if all finite affine combinations of points of S are in S.*

Proof. Follow the proof of Theorem 18.1.24. □

18.2.8 Definition. For any set S, let $A(S)$ be the set of all finite affine combinations of points of S.

By taking the number of points r in the affine combination to be 1, so that $\lambda_1 = 1$, we have

$$S \subset A(S) \tag{18.2.9}$$

18.2.10 Theorem. *For any set S, $A(S)$ is an affine set.*

Proof. Follow the proof of Theorem 18.1.26. □

18.2.11 Theorem. *For any set S, the affine hull is equal to the set of affine combinations.*

Proof. Follow the proof of Theorem 18.1.28. □

18.2.12 Definition. We say that the points $\mathbf{x}^0$, ..., $\mathbf{x}^k$ are *affinely dependent* if there are real numbers a_i, such that

$$\sum_{i=0}^{k} a_i \mathbf{x}^i = \mathbf{0}, \text{ with } \sum_{i=0}^{k} a_i = 0 \text{ and not all } a_i = 0. \tag{18.2.13}$$

We say they are *affinely independent* otherwise.

18.2.14 Example. The points $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$ in $\mathbb{R}^3$ are affinely independent. Indeed, if you add the origin to this set of points, it is still affinely independent.

18.2.15 Exercise. Show that if there is repetition in the list of points $\mathbf{x}^0$, ..., $\mathbf{x}^k$, so for example if $\mathbf{x}^0 = \mathbf{x}^1$, the points are affinely dependent.

18.2.16 Proposition. *The points* $\mathbf{x}^0$, ..., $\mathbf{x}^k$ *are affinely dependent if and only if the vectors* $\mathbf{x}^i - \mathbf{x}^0$, $1 \leq i \leq k$, *are linearly dependent.*

Proof. Assume that $\mathbf{x}^0$, ..., $\mathbf{x}^k$ are affinely dependent, so there are real numbers a_i satisfying (18.2.13). Then

$$a_0 = -\sum_{i=1}^{k} a_i. \tag{18.2.17}$$

If $a_0 \neq 0$, substitute a_0 into the equation of affine dependence, getting

$$\sum_{i=1}^{k} a_i(\mathbf{x}^i - \mathbf{x}^0) = 0. \tag{18.2.18}$$

Not all the coefficients in this equation are zero by (18.2.13), so this is the required equation of linear dependence between the $\mathbf{x}^i - \mathbf{x}^0$.

To get the other implication, start from the equation of linear dependence (18.2.18) and define a_0 by (18.2.17). This gives (18.2.13), the required equation of affine dependence. □

An important special case occurs when $\mathbf{x}^0$ is the origin: starting from a basis of $\mathbb{R}^n$, and adding the origin, you get an affinely independent set of $n+1$ points.

18.2.19 Exercise. Prove that if $\mathbf{x}$ is an affine combination of $\mathbf{x}^0, \dots, \mathbf{x}^k$, and if the $\mathbf{x}^i$, $0 \le i \le k$, are affinely dependent, then $\mathbf{x}$ is an affine combination of a smaller number of the $\mathbf{x}^i$.

At this point the theory of affine sets and that of convex sets diverge: the theory of affine sets is much simpler. The next theorem shows that all non-empty affine sets are translates of linear spaces, a concept we will now make precise by using a special case of the Minkowski sum, where the first set is a point, and the second set a linear subspace of $\mathbb{R}^n$.

18.2.20 Definition. A set S of the form $\mathbf{s} + V$, where $\mathbf{s}$ is a point in S and V a linear subspace of $\mathbb{R}^n$ is called a *translate* of V, or a *flat*[3].

18.2.21 Theorem. *A non-empty subset S of $\mathbb{R}^n$ is affine if and only if it is a translate of a linear subspace V of $\mathbb{R}^n$.*

Proof. First we assume the set S is the Minkowski sum $\mathbf{s}+V$, and show it is affine. We take two points $\mathbf{s} + \mathbf{v}^1$ and $\mathbf{s} + \mathbf{v}^2$ in S, and show that for any $\lambda \in \mathbb{R}$,

$$\lambda(\mathbf{s} + \mathbf{v}^1) + (1 - \lambda)(\mathbf{s} + \mathbf{v}^2) = \mathbf{s} + \lambda\mathbf{v}^1 + (1 - \lambda)\mathbf{v}^2 \in S.$$

This is true because V is a subspace, so when $\mathbf{v}^1 \in V$ and $\mathbf{v}^2 \in V$, any linear combination of the two is in V. Thus S is affine. Finally, since the origin $\mathbf{0}$ is in V, $\mathbf{s} + \mathbf{0} = \mathbf{s}$ is in S.

Next we assume that S is affine. Pick any $\mathbf{s} \in S$. Consider the Minkowski sum $V := -\mathbf{s} + S$. It contains the origin. To show V is a subspace, we must show that it is closed under scalar multiplication and vector addition. So pick any $-\mathbf{s} + \mathbf{s}^1$ and $-\mathbf{s} + \mathbf{s}^2$ in V. We must show that for any λ_1 and λ_2 in $\mathbb{R}$,

$$\lambda_1(-\mathbf{s} + \mathbf{s}^1) + \lambda_2(-\mathbf{s} + \mathbf{s}^2) = -(\lambda_1 + \lambda_2)\mathbf{s} + \lambda_1\mathbf{s}^1 + \lambda_2\mathbf{s}^2 \text{ is in } V.$$

This element is in V if and only if when $\mathbf{s}$ is added to it, the new element is in S. This new element is written

$$(1 - \lambda_1 - \lambda_2)\mathbf{s} + \lambda_1\mathbf{s}^1 + \lambda_2\mathbf{s}^2.$$

This is an affine combination of $\mathbf{s}$, $\mathbf{s}^1$, and $\mathbf{s}^2$ which is therefore in S. □

18.2.22 Theorem. *Any non-empty affine set S in $\mathbb{R}^n$ can be written as $\mathbf{s}+V$, where $\mathbf{s}$ is an arbitrary point of S, and V is a uniquely determined linear subspace of $\mathbb{R}^n$. Indeed, V is the Minkowski sum $S - S$.*

[3] A term used in many books on affine geometry: see for example Lay [40], p. 12.

Proof. This all follows from Theorem 18.2.21 except the uniqueness of V, which we obtain by its description as a Minkowski sum. In the previous theorem, we constructed for each $\mathbf{s} \in S$ a linear space $V_{\mathbf{s}}$ such that $S = \mathbf{s} + V_{\mathbf{s}}$. Indeed, $V_{\mathbf{s}} = S - \mathbf{s}$. We will show that $V_{\mathbf{s}} \subset V_{\mathbf{t}}$, for any $\mathbf{t} \in S$. This implies that all the $V_{\mathbf{t}}$ are the same, finishing the proof. Let $\mathbf{s}^1 - \mathbf{s}$ be an arbitrary element of $V_{\mathbf{s}}$, so $\mathbf{s}^1 \in S$, Since $\mathbf{s}^2 := \mathbf{s}^1 - \mathbf{s} + \mathbf{t}$ is an affine combination of points of S, $\mathbf{s}^2$ is in S, so $\mathbf{s}^2 - \mathbf{t} = \mathbf{s}^1 - \mathbf{s}$ is in $V_{\mathbf{t}}$, as required. Thus all the $V_{\mathbf{s}}$ are the same, showing that each one is $S - S$. □

18.2.23 Definition. Let S be a non-empty affine set. Let V be the linear space associated to it by Theorem 18.2.22. Then the *dimension* of S is the vector space dimension of V. If V is empty, the dimension of S is -1.

18.2.24 Definition. The *dimension* of a convex set C is the dimension of the affine hull of C.

18.2.25 Example. As we will soon see, an m-simplex is the convex hull of $m + 1$ affinely independent points. So its dimension is m.

18.2.26 Example. Assume S is an affine space of dimension $n - 1$ in $\mathbb{R}^n$, and V its associated linear space. Then S is a hyperplane, and V a hyperplane through the origin. If $\mathbf{a}$ is a normal vector for V, so $V := H_{\mathbf{a},0}$, then S is the parallel (meaning it has the same normal vector) hyperplane $H_{\mathbf{a},c}$, where $c = \mathbf{a} \cdot \mathbf{s}$, for any $\mathbf{s} \in S$.

Since $\mathbf{a}$ is the normal vector for V, we have $\mathbf{a} \cdot \mathbf{v} = 0$, for all $\mathbf{v} \in V$. Since S is written $\mathbf{s} + V$, for any $\mathbf{s} \in S$, we see that $\mathbf{a} \cdot \mathbf{s}$ is constant on S. If we call this value c, this shows that S is $H_{\mathbf{a},c}$.

18.2.27 Example. Let's write the equations of the affine line in $\mathbb{R}^3$ through the points $(3, 0, 1)$ and $(1, 2, 0)$. That means we need to find all the solutions in (a, b, c, d) of the equations

$$ax + by + cz = d$$

that verify $3a + c = d$ and $a + 2b = d$. We have 2 equations in 4 unknowns, and we can easily solve in terms of a and b:

$$\begin{aligned} c &= -2a + 2b \\ d &= a + 2b \end{aligned}$$

Notice that there is only one *linear* equation through the points: indeed, when $d = 0$, we have , up to a scalar, $-2x + y + 6z = 0$. The two affine equations, are, for example, $x - 2z = 1$ and $y + 2z = 2$.

18.3 Polytopes and Polyhedra

18.3.1 Convex Independence

Just as we defined linear independence and affine independence for a set of points, we can do the same for convex independence.

18.3.1 Definition. A set S of two or more points is *convexly independent* if no point s_0 in S is in the convex hull of the remaining points. A single point is convexly independent.

18.3.2 Exercise. Show that if a (necessarily finite) set of points is linearly independent, then it is affinely independent. If a set is affinely independent, then it is convexly independent. Given an example of

1. An infinite set of points that is convexly independent. Because it is infinite, it cannot be affinely independent;
2. A finite set of points that is convexly independent, and not affinely independent.
3. An affinely independent set of points that is not linearly independent.

The following lemma will be used in the proof of Theorem 18.3.4. Its proof is a simple exercise, and is left to you.

18.3.3 Lemma. *Assume a set S is not convexly independent, so that there is a point $\mathbf{s}_0 \in S$ that is a convex combination of other points of S. Then $\mathbf{s}_0$ is not extreme for the convex hull of S.*

18.3.4 Theorem. *If S is a finite set of points, then the extreme points E of the convex hull of S form the unique convexly independent subset of S with convex hull equal to the convex hull $K(S)$ of S.*

Proof. If the set S is not convexly independent, then an $\mathbf{s}_0 \in S$ can be written as a convex combination of the remaining points of S. Then remove $\mathbf{s}_0$ from S: the remaining points have the same convex hull as S. Continue doing this one point at a time until you are left with a convexly independent subset S^0 with the same convex hull as S. None of the removed points is extreme by Lemma 18.3.3, and conversely it is easy to see that the extreme points are all contained in S^0. Write the points of S^0 as $\mathbf{a}^i$, $0 \le i \le m$. To conclude the proof we must show that all the $\mathbf{a}^i$ are extreme. We prove this by contradiction. Assume, without loss of generality, that $\mathbf{a}^m$ is not extreme. Then it can be written as a combination $\mathbf{a}^m = \lambda\mathbf{p} + (1-\lambda)\mathbf{q}$, with $0 < \lambda < 1$ and $\mathbf{p}$ and $\mathbf{q}$ in $K(S) = K(S^0)$, with $\mathbf{p} \neq \mathbf{a}^m \neq \mathbf{q}$. Since $\mathbf{p}$ and $\mathbf{q}$ are in the convex hull of the S^0, they can be written

$$\mathbf{p} = \sum_{i=0}^{m} \mu_i \mathbf{a}^i, \quad \sum_{i=0}^{m} \mu_i = 1, \mu_i \geq 0;$$
$$\mathbf{q} = \sum_{i=0}^{m} \nu_i \mathbf{a}^i, \quad \sum_{i=0}^{m} \nu_i = 1, \nu_i \geq 0;$$

so that

$$\mathbf{a}^m = \lambda \mathbf{p} + (1-\lambda)\mathbf{q} = \sum_{i=0}^{m} \big(\lambda\mu_i + (1-\lambda)\nu_i\big)\mathbf{a}^i.$$

For all i, $0 \leq i \leq m$, define

$$\pi_i = \lambda\mu_i + (1-\lambda)\nu_i. \tag{18.3.5}$$

Then $\pi_i \geq 0$, as you should check, and

$$\sum_{i=0}^{m} \pi_i = \lambda \sum_{i=0}^{m} \mu_i + (1-\lambda) \sum_{i=0}^{m} \nu_i = \lambda + (1-\lambda) = 1. \tag{18.3.6}$$

Moving the term in $\mathbf{a}^m$ to the left-hand side, we get:

$$(1-\pi_m)\mathbf{a}^m = \sum_{i=0}^{m-1} \pi_i \mathbf{a}^i$$

If $1 - \pi_m > 0$, divide by it to get

$$\mathbf{a}^m = \sum_{i=0}^{m-1} \frac{\pi_i}{1-\pi_m} \mathbf{a}^i.$$

Since all the coefficients in this sum are non-negative, and

$$\sum_{i=0}^{m-1} \frac{\pi_i}{1-\pi_m} = 1,$$

this expresses $\mathbf{a}^m$ as a convex combination of the remaining $\mathbf{a}^i$: a contradiction to the assumption of convex independence.

If $1 - \pi_m = 0$, the only other possibility, then all the other π_i are 0, since they are non-negative and (18.3.6) holds. By (18.3.5), since λ and $1-\lambda$ are both positive, this forces $\mu_i = \nu_i = 0$, for $0 \leq i \leq m-1$. This in turn says that $\mathbf{p} = \mathbf{q} = \mathbf{a}^m$, so that $\mathbf{a}^m$ is extreme. So all the points in S^0 are extreme. □

18.3.2 Polytopes

18.3.7 Definition. A *polytope* is the convex hull of a finite number of points.

18.3.8 Remark. By Theorem 18.3.4, it suffices to consider polytopes on convexly independent sets of points, which are then called the *vertices* of the polytope. This definition agrees with the more general definition of vertex since we are just talking about the extreme points of the convex set.

If the vertices are $\mathbf{a}^0, \dots, \mathbf{a}^m$, and we write A for the $(m+1) \times n$ matrix with rows $\mathbf{a}^i$, then we denote the polytope on these points by

$$P_A = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} = \sum_{i=0}^{m} \lambda_i \mathbf{a}^i, \text{for } 0 \leq \lambda_i \leq 1, 0 \leq i \leq m, \sum_{i=0}^{m} \lambda_i = 1\}. \tag{18.3.9}$$

A simplex is the special case of a polytope where the vertices are affinely independent:

18.3.10 Definition. A *m-simplex* in $\mathbb{R}^n$, for $m \leq n$ is the set of all convex combinations of $m+1$ affinely independent points $\mathbf{a}^0, \dots, \mathbf{a}^m$. The $\mathbf{a}^i$ are the vertices of the simplex, and the segments $[\mathbf{a}^i, \mathbf{a}^j]$ are the *edges*. We write the simplex:

$$H(\mathbf{a}^0, \dots, \mathbf{a}^m) = \{\mathbf{x} = \sum_{i=0}^{m} \lambda_i \mathbf{a}^i \mid \lambda_i \geq 0\,,\ \sum_{i=0}^{m} \lambda_i = 1\} \tag{18.3.11}$$

18.3.12 Example. A convenient n-simplex in $\mathbb{R}^n$ is the one where $\mathbf{a}^0$ is the origin, and then $\mathbf{a}^i$ is the i-th unit coordinate vector $\mathbf{e}^j$.

18.3.13 Definition. The λ_i in (18.3.11) are the *barycentric coordinates* of the point $\mathbf{x}$ in the m-simplex $H(\mathbf{a}_0, \dots, \mathbf{a}_m)$. The *barycenter* or *centroid* of the m-simplex is the point

$$\mathbf{c} := \frac{1}{n+1}(\mathbf{a}^0 + \mathbf{a}^1 + \dots + \mathbf{a}^n) \tag{18.3.14}$$

of the simplex.

The barycentric[4] coordinates are uniquely determined for every point in the affine hull of the vertices of the simplex. This is easily seen: if there were two sets of barycentric coordinates, they could be used to produce an equation of affine dependence between the vertices, a contradiction.

[4] Introduced by Möbius in 1827: see [33], p. 134

18.3.15 Example. The unit cube in $\mathbb{R}^n$ is the polytope with vertices all the points whose coordinates are either 0 or 1. Thus it has 2^n vertices. Its vertices are the 2^n points whose coordinates are either 0 or 1. So in $\mathbb{R}^2$, the vertices are $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$.

The *crosspolytope* is the polytope in $\mathbb{R}^n$ with vertices the $2n$ points whose coordinates are all 0 except in one position, where the coordinate is either 1 or -1. The crosspolytope in $\mathbb{R}^2$ has vertices $(-1,0)$, $(1,0)$, $(0,-1)$, $(0,1)$ and thus is just a rotated square. In $\mathbb{R}^3$ its vertices are the six points $(-1,0,0)$, $(1,0,0)$, $(0,-1,0)$, $(0,1,0)$, $(0,0,-1)$, $(0,0,1)$, so it is not a rotated cube: it does not have enough vertices.

For further examples see [40], chapter 8.

18.3.3 Polyhedra

A dual notion to polytope is that of a polyhedron.

18.3.16 Definition. A *polyhedron* P in $\mathbb{R}^n$ is the set of solutions $\mathbf{x}$ of a system of linear inequalities $A\mathbf{x} \le \mathbf{b}$, where A is an $m \times n$ matrix and $\mathbf{b}$ is a m-vector. If the i-th row of A is noted $\mathbf{a}^i$, then P is the intersection of the m half-spaces $H^-_{\mathbf{a}^i,b_i}$ (see Example 18.1.7) so it is closed and convex. We write $P(A,\mathbf{b})$ when we need to indicate the dependence on A and $\mathbf{b}$.

18.3.17 Definition. Each one of the inequalities $\langle \mathbf{a}^i, \mathbf{x}\rangle \le b_i$, $1 \le i \le m$, is called a *constraint*, and $\mathbf{a}^i$ is the normal vector to the constraint. By an *active constraint* at $\mathbf{p}$ we mean a constraint that gives an equality when evaluated at $\mathbf{p}$: $\langle \mathbf{a}^i, \mathbf{p}\rangle = b_i$.

We will use this terminology when we study constrained optimization. See for example §25.1, Definitions 25.3.2, and 31.1.5.

You should contrast this notation $P(A,\mathbf{b})$ for a polyhedron with the notation P_A given in (18.3.9) for a polytope. In both cases we have $m \times n$ matrix A. In the case of a polyhedron, the rows of A represent the coefficients of the inequalities that the points of the polyhedron must satisfy, while for a polytope, the rows of A are the vertices.

We will prove later that any bounded polyhedron is a polytope, Theorem 18.7.8, and then that any polytope is a polyhedron, the famous Weyl-Minkowski Theorem 20.2.14). Thus there must be a way of passing from the representation $P(A,\mathbf{b})$ for the bounded polyhedron to the representation P_A for the same set considered as a polytope: the matrix A will be different, of course. We now show how to do this for a simplex.

18.3.18 Theorem. *A simplex is a polyhedron.*

Proof. Suppose the simplex S is given by $n+1$ affinely independent vectors $\mathbf{a}^0$, $\mathbf{a}^1$, ..., $\mathbf{a}^n$ in $\mathbb{R}^n$. If S is in a bigger space, just take the affine hull of S to get to a space of the right dimension. To establish the result, we will write S as the intersection of $n+1$ half-spaces.

For any j, $0 \leq j \leq n$, let H'_j be the affine hyperplane that goes through $\mathbf{b}^i = \mathbf{a}^i - \mathbf{a}^j$, for all i except j. Since the $n+1$ points $\mathbf{a}^i$ are affinely independent, the n points $\mathbf{b}^i$ are linearly independent by Proposition 18.2.16, so H'_j is uniquely determined, as we show in the lemma below. Write the equation of H'_j as $c^j_1 x_1 + \dots + c^j_n x_n = d_j$. The equation for the hyperplane H_j passing through the $\mathbf{a}^i$, $i \neq j$, is $c^j_1 x_1 + \dots + c^j_n x_n = e_j$, where $e_j = d_j + f(\mathbf{a}^j)$, so only the right-hand side of the equation changes. If you substitute for the x_k the coordinates a^i_k of the i-th point $\mathbf{a}^i$, then the c^j_k and e_j must satisfy these n equations. Now let H^+_j be the half-space bounded by H_j that contains the last generator $\mathbf{a}^j$ of the simplex S. Clearly H^+_j contains S and is convex. So the intersection $C := \cap^n_{j=0} H^+_j$ contains S and is convex. Any point $\mathbf{p}$ in $\mathbb{R}^n$ is an affine combination of the $\mathbf{a}_i$, so it can be written

$$\mathbf{p} = \sum_{i=1}^{k} \lambda_i \mathbf{a}_i$$

with $\sum_{i=1}^{k} \lambda_i = 1$. Those that are convex combinations of the $\mathbf{a}_i$ also have all $\lambda_i \geq 0$. Suppose that there is a point $\mathbf{p} \in C$ that is not a convex combination of the $\mathbf{a}^i$. Then there is an i such that $\lambda_i < 0$. We evaluate H_i on $\mathbf{p}$. By linearity we see that its value is λ_i times the value at $\mathbf{a}^i$, since it vanishes at all the other $\mathbf{a}^j$. Since λ_i is negative, the point $\lambda_i \mathbf{a}^i$ is in the interior of the half-space H^-_j, so it is not in C, and we have our contradiction.

18.3.19 Lemma. *Take n linearly independent points $\mathbf{b}^i$, $1 \leq i \leq n$, in $\mathbb{R}^n$. Then there is a unique hyperplane H passing through these points.*

Proof. Write B for the $n \times n$ matrix whose i-th row are the coordinates of $\mathbf{b}^i$. Write the equation of H as

$$c_1 x_1 + c_2 x_2 + \dots c_n x_n = d,$$

so the vector $\mathbf{c}$ and the unknown number d satisfy the system of n equation $B\mathbf{c} = \mathbf{d}$, where $\mathbf{d} = (d, d, \dots, d)$. Linear independence of the points $\mathbf{b}^i$ is equivalent to saying that the matrix B is invertible, so there is a unique solution to the system of equations, up to scaling by a non-zero constant. Thus the hyperplane H is unique. □

This concludes the proof of the theorem. □

We only state the next theorem for the n-simplex in $\mathbb{R}^n$, so that its affine hull is all of $\mathbb{R}^n$.

18.3.20 Theorem. *The n-simplex in $\mathbb{R}^n$ contains an open set.*

Proof. Using the notation of the proof of Theorem 18.3.18, we see that the intersection of the open half-spaces $\mathring{H}_j^+$, which is open, is contained in S. □

This result is used in the proof of Theorem 18.4.2.

18.3.21 Example. The two polytopes in Example 18.3.15 are polyhedra. To prove this we need to exhibit the inequalities they satisfy.

The unit cube in $\mathbb{R}^n$ is given by the $2n$ inequalities $\{\mathbf{x} \in R^n \mid 0 \leq x_i \leq 1, \quad i = 1, \dots, n\}$.

The crosspolytope is given by the non-linear inequality $\{\mathbf{x} \in R^n \mid \sum_{i=1}^n |x_i| \leq 1\}$. We write this as a collection of linear equations by setting $\epsilon_i = \pm 1, 1 \leq i \leq n$. Thus there are 2^n possible vectors $\epsilon = (\epsilon_1, \dots \epsilon_n)$. We claim the crosspolytope is given by the inequalities

$$\{\mathbf{x} \in R^n \mid -1 \leq \sum_{i=1}^n \epsilon_i x_i \leq 1, \forall \epsilon\}.$$

Because $-\epsilon$ is an element of the set if ϵ is, all the inequalities on the left-hand side are unnecessary. To show that we get the desired vertices, we need to use Theorem 18.7.3. For example, let's show that the point $(1, 0, \dots, 0)$ is a vertex: we need to find n linearly independent defining inequalities that vanish at this point. Just consider those ϵ with $\epsilon_1 = 1$: by varying the remaining coefficients it is easy to construct a linearly independent set.

18.3.22 Exercise. Show that the representation of the crosspolytope as a polyhedron given in Example 18.3.21 gives the same set as the crosspolytope. For the crosspolytope in $\mathbb{R}^3$, write down explicitly a linearly independent set of active constraints at the point $(1, 0, 0)$.

18.3.4 Regular simplices

18.3.23 Definition. A simplex is *regular* if all its edges have the same length.

18.3.24 Example. Regular simplices exist in all dimensions. In dimension 1 we have the interval, in dimension 2 any equilateral triangle, in dimension 3 a pyramid (or tetrahedron) with base an equilateral triangle and sides an equilateral triangle of the same size. This construction can be pursued in all dimensions, for example by using barycentric coordinates. Assume you have a regular simplex in $\mathbb{R}^n$ with vertices $\mathbf{a}^0, \dots, \mathbf{a}^n$ and with edge length e. Let $\mathbf{c}$ be the barycenter of the simplex. Then for every i, $0 \le i \le n$, we have

$$\mathbf{a}^i - \mathbf{c} = \frac{1}{n+1} \sum_{j=0, j \neq i}^{n} (\mathbf{a}^i - \mathbf{a}^j) \tag{18.3.25}$$

By hypothesis, all the vectors $\mathbf{a}^i - \mathbf{a}^j$ in this sum have the same length e. Next work in the affine plane spanned by any three of the vertices, say $\mathbf{a}^i$, $\mathbf{a}^j$ and $\mathbf{a}^k$. Since they form an equilateral triangle, the angle between $\mathbf{a}^i - \mathbf{a}^j$ and $\mathbf{a}^i - \mathbf{a}^k$ is 60 degrees, so its cosine is $1/2$, so

$$\langle \mathbf{a}^i - \mathbf{a}^j, \mathbf{a}^i - \mathbf{a}^k \rangle = e^2/2.$$

This allows us to compute the length of $\mathbf{a}^i - \mathbf{c}$. Indeed, by (18.3.25) we can get the dot product

$$\begin{aligned}
\langle \mathbf{a}^i - \mathbf{c}, \mathbf{a}^i - \mathbf{c} \rangle &= \frac{1}{(n+1)^2} \langle \sum_{j \neq i}^{n} (\mathbf{a}^i - \mathbf{a}^j), \sum_{k \neq i}^{n} (\mathbf{a}^i - \mathbf{a}^k) \rangle \\
&= \frac{1}{(n+1)^2} \sum_{j \neq i}^{n} \big(\sum_{k \neq i}^{n} \langle \mathbf{a}^i - \mathbf{a}^j, \mathbf{a}^i - \mathbf{a}^k \rangle \big) \\
&= \frac{e^2}{(n+1)^2} \sum_{j \neq i}^{n} (1 + (n-1)/2) \\
&= \frac{e^2}{(n+1)^2} \big(\frac{n(n+1)}{2} \big) = \frac{n}{2(n+1)} e^2.
\end{aligned}$$

so the vertices are all on a sphere of radius

$$r_n = e \sqrt{\frac{n}{2(n+1)}}$$

centered at $\mathbf{c}$.

Observe that as n increases, r_n increases: see Exercise 18.3.26. The computation confirms the easy facts from elementary geometry: $r_1 = e/2$ and $r_2 = e/\sqrt{3}$.

This computation gives us conditions a regular simplex must satisfy. Now we construct the regular simplex of side length e in $\mathbb{R}^n$ by induction on n. Start at $n = 1$, with a line segment of length e. Assuming the regular simplex S_n of dimension n, with vertices $\mathbf{a}^0, \dots, \mathbf{a}^n$ and side length e has been constructed, we construct the one in dimension $n + 1$ by taking a line L perpendicular to the plane of S_n passing through the barycenter $\mathbf{c}^n$ of S_n. The last vertex a_{n+1} of a regular simplex S_{n+1} extending the simplex S_n can be found on L at distance $r_{n+1} + \sqrt{r_{n+1}^2 - r_n^2}$ from the plane of S_n. The barycenter $\mathbf{c}^{n+1}$ of S_{n+1} is the point on L at distance $\sqrt{r_{n+1}^2 - r_n^2}$ from $\mathbf{c}^n$, on the same side as a_{n+1}. Note how we used the fact the r_n increase with n.

18.3.26 Exercise. Prove that the function

$$f(x) = \Bigl(\frac{x}{x+1}\Bigr)^{1/2}$$

is increasing. Hint: do a derivative computation.

18.3.27 Exercise. When $n = 3$, we see that the regular tetrahedron of side length e is inscribed in a sphere of radius $r_3 = e\sqrt{3/8}$. Confirm this using space geometry as follows. We can set $e = 1$ by changing our unit of length. Center the sphere of radius $\sqrt{3/8}$ at the origin, so its equation is

$$x^2 + y^2 + z^2 = 3/8.$$

Pick a point $\mathbf{a}^3$ on this sphere, say $(0, 0, \sqrt{3/8})$, and take the sphere of radius 1 centered at $\mathbf{a}^3$. The points at the right distance from $\mathbf{a}^3$ must lie on this sphere, which has equation

$$x^2 + y^2 + (z - \sqrt{3/8})^2 = 1.$$

Find the set of common solutions of these two equations by subtracting the second from the first. You get a linear equation in z. Plug the solution of this equation into either of the two equations, and show you get the equation of a circle of radius $1/\sqrt{3}$. Explain why you are done.

18.4 The Topology of Convex Sets

We first look at convex sets, and show that both their closure and their interior is convex. Then we look at open sets and show that their convex hull is open. We also show that the convex hull of a compact set is compact. Finally an example shows that the convex hull of a closed set is not necessarily closed.

18.4.1 Theorem. *If C is a convex set in $\mathbb{R}^n$, then its closure $\bar{C}$ is also convex.*

Proof. We must show that if $\mathbf{x}$ and $\mathbf{y}$ are in $\bar{C}$, then the segment $(\mathbf{x}, \mathbf{y})$ of points between $\mathbf{x}$ and $\mathbf{y}$ is in $\bar{C}$. A point $\mathbf{z}$ in $(\mathbf{x}, \mathbf{y})$ can be written $\mathbf{z} = \lambda\mathbf{x} + (1-\lambda)\mathbf{y}$, with $0 < \lambda < 1$. Because $\mathbf{x}$ is in the closure of C, it can be written as the limit of a sequence $\{\mathbf{x}_i\}$ of points of C, and similarly $\mathbf{y}$ can be written as the limit of a sequence $\{\mathbf{y}_i\}$ of points of C. Because C is convex, the segment $(\mathbf{x}_i, \mathbf{y}_i)$ is in C. In particular the point $\mathbf{z}_i = \lambda\mathbf{x}_i + (1-\lambda)\mathbf{y}_i$ is in C. The sequence $\{\mathbf{z}_i\}$ converges to $\mathbf{z}$, so $\mathbf{z}$ is in the closure of C, as required. □

18.4.2 Theorem. *For a convex set C in $\mathbb{R}^n$, the set of its interior points is non-empty if and only if the dimension of C is n.*

Proof. We first prove: if C has a non-empty interior, then it has dimension n. Indeed, if it has a non-empty interior, just pick a point in the interior: a small open ball around that point is in C, so an n-simplex is in C. The affine hull of an n-simplex has dimension n, so we are done.

In the other direction, if the dimension of C is n, then one can find $n+1$ affinely independent points in C. But then the simplex on these $n+1$ points is in C, so we conclude with Theorem 18.3.20. □

For a convex set of dimension n we have:

18.4.3 Lemma. *If C is a convex set, then the set of its interior points, known as its interior $\mathring{C}$, is convex.*

Proof. We must show that if $\mathbf{x}$ and $\mathbf{y}$ are in $\mathring{C}$, then the segment $[\mathbf{x}, \mathbf{y}]$ is in $\mathring{C}$. So pick a point $\mathbf{z}$ in $[\mathbf{x}, \mathbf{y}]$. It can be written $\mathbf{z} = \lambda\mathbf{x} + (1-\lambda)\mathbf{y}$, with $0 \le \lambda \le 1$. and since C is convex we have $\mathbf{z} \in C$. Pick an arbitrary $\mathbf{u}$ in a ϵ-neighborhood of $\mathbf{0}$, so that $\mathbf{z}+\mathbf{u}$ is within ϵ of $\mathbf{z}$. Then

$$\mathbf{z} + \mathbf{u} = \lambda(\mathbf{x} + \frac{\mathbf{u}}{\lambda}) + (1-\lambda)\mathbf{y}. \tag{18.4.4}$$

Since $\mathbf{x}$ is in $\mathring{C}$, $\mathbf{x} + \mathbf{u}/\lambda$ is in C, if ϵ is small enough. Then the convexity of C and (18.4.4) show that $\mathbf{z}+\mathbf{u}$ is in C. Since this is true for any $\mathbf{u}$ in an ϵ-neighborhood of $\mathbf{z}$, $\mathbf{z}$ is an interior point of C. □

In the same way we can prove:

18.4.5 Corollary. *If C is a convex set and if M is the affine hull of C, then the set* $\operatorname{relint} C$*, the interior of C relative to M, is also convex.*

18.4.6 Theorem. *The convex hull of an open set S is open.*

Proof. Given any point $\mathbf{x}$ in the convex hull $K(S)$ of S, we must show that a small ball centered at $\mathbf{x}$ is in $K(S)$. Let $\mathbf{u}$ stand for an arbitrary element of $\mathbb{R}^n$ of length less than some $\epsilon > 0$. Then we must show that $\mathbf{x}+\mathbf{u}$ is in $K(S)$. Note that ϵ will depend on $\mathbf{x}$. By definition we can write $\mathbf{x}$ of a convex combination of some finite number r of points $\mathbf{x}_i$ of S:

$$\mathbf{x} = \sum_{i=1}^{r} \lambda_i \mathbf{x}_i \quad \text{where } \sum_{i=1}^{r} \lambda_i = 1 \text{ and } \lambda_i \geq 0, \forall i,$$

Then

$$\mathbf{x} + \mathbf{u} = \sum_{i=1}^{r} \lambda_i (\mathbf{x}_i + \mathbf{u}). \tag{18.4.7}$$

An $\epsilon_i > 0$ neighborhood of each $\mathbf{x}_i$ is contained in S. By taking ϵ smaller than the smallest of the ϵ_i, we see that for any $\mathbf{u}$ of length less than ϵ, each $\mathbf{x}_i + \mathbf{u}$ is in S, so by (18.4.7) $\mathbf{x} + \mathbf{u}$ is in $K(S)$, and we are done. □

18.4.8 Example. The following example in $\mathbb{R}^2$ shows that the convex hull of a closed set need not be closed. Let S be the set of points (x, y) in the plane such that $y \geq 1/(1 + x^2)$. It is easy to see that S is closed. The convex hull of S is the set of points (x, y), $y > 0$, so it is not closed.

First we show that any point (a, b) with $b > 0$ is in the convex hull. If $b \geq 1$, it is already in S, so we may assume $b < 1$. Then take the horizontal line $y = b$. It intersects the curve $y = 1/(1 + x^2)$ in two points, showing that the point is in the convex hull. Finally, we take any point (a, b), with $b \leq 0$. We examine how a line through (a, b) could meet S in two points on either side of (a, b). This cannot happen since there is no point in S with second coordinate less than or equal to 0.

On the other hand we have:

18.4.9 Theorem. *The convex hull of a compact set S in $\mathbf{R}^n$ is compact.*

Proof. We first show that the convex hull $K(S)$ of a bounded set S is bounded. If S is bounded, then there is a R such that $\|\mathbf{x}\| \leq R$ for all $\mathbf{x} \in S$. Any element in the convex hull of S can be written as a convex combination of some number $k+1$ elements of S, so by the triangle inequality

$$\|\sum_{i=0}^{k} \lambda_i \mathbf{x}_i\| \leq \sum_{i=0}^{k} (\lambda_i R) = R,$$

since $\sum_{i=0}^{k} \lambda_i = 1$. Thus the elements in the convex hull are bounded by the same R.

To finish the proof, we show that $K(S)$ is closed. We need Carathéodory's Theorem 18.5.1. Take any $\mathbf{x}$ in the closure of $K(S)$. We must show (by Definition 14.4.7) that $\mathbf{x}$ is in $K(S)$. We know that there is a sequence of points $\mathbf{x}_j$ in $K(S)$

approaching $\mathbf{x}$. Each one of the $\mathbf{x}_j$ can be written, using Carathéodory's Theorem, as

$$\mathbf{x}_j = \sum_{i=0}^{n} \lambda_{ij}\mathbf{y}_{ij},$$

where the numbers λ_{ij} are all bounded, since between 0 and 1, and the points $\mathbf{y}_{ij} \in S$ are all bounded, since within a ball of radius R around the origin. We now use Theorem 15.3.3, which tells us that we can extract a convergent subsequence from any bounded sequence. We set the index i to 0. Then we can find a subsequence of the j so that both the subsequence of the λ_{0j} and the $\mathbf{y}_{0j}$ converge on the subsequence. Then we take a subsequence of the subsequence that makes the λ_{1j} and the $\mathbf{y}_{1j}$ converge. We can repeat the process a finite number of times, and we end up with a subsequence of the $\{j\}$ that we write $\{j_k\}$ so that all the subsequences converge:

$$\lim_{k\to\infty} \lambda_{ij_k} = \lambda_i \text{ and } \lim_{k\to\infty} \mathbf{y}_{ij_k} = \mathbf{y}_i.$$

Because S is closed and $\{\mathbf{y}_{ij_k}\}$ is in S for all k, the limit $\mathbf{y}_i$ is in S. Now

$$\lim_{k\to\infty} \mathbf{x}_{j_k} = \lim_{k\to\infty} \sum_{i=0}^{n} \lambda_{ij_k}\mathbf{y}_{ij_k} = \sum_{i=0}^{n} \lambda_i \mathbf{y}_i$$

and by construction this is $\mathbf{x}$. So we have written $\mathbf{x}$ as a convex combination of elements of S, as required.[5] □

As a special case, we get

18.4.10 Corollary. *A polytope is a compact set.*

Proof. This is immediate, since a polytope is the convex hull of a finite number of points. □

[5] This proof can be found in [52] p.78. Another proof is given in [33], §5.3. Here is a different proof from [22], p.15, using another key theorem of Lecture 16.

Let Δ_n be the collection of $\lambda \in \mathbb{R}^{n+1}$ of points with coordinates λ_i with $\lambda_i \geq 0$ and $\sum_{i=0}^{n} \lambda_i = 1$. Δ_n is closed and bounded and therefore compact. We define a map:

$$(\lambda, \mathbf{x}^0, \dots, \mathbf{x}^n) \mapsto \sum_{i=0}^{n} \lambda_i \mathbf{x}^i$$

on $\Delta_n \times S \times \dots \times S$, which is a compact set in $\mathbb{R}^{n^2+2n+1}$. Clearly the image is contained in the convex hull of S. By Carathéodory's Theorem it actually contains the convex hull. The map is continuous and the domain compact, so by Theorem 16.2.1 the image is compact, which is what we needed.

18.4.11 Exercise. A *convex body* is a compact convex set of dimension n in $\mathbb{R}^n$. Its interior is non-empty by Theorem 18.4.2. Let C be a convex body with boundary B. Pick a point $\mathbf{p}$ in the interior of C. Now consider the sphere $S_r(\mathbf{p})$ of radius r centered at $\mathbf{p}$, and Show there there is a one-to-one map $\mathbf{f}\colon S_r(\mathbf{p}) \to B$.

Hint: Recall Definition 9.1.4 for the concept of a ray r emanating from the point $\mathbf{p}$. Such a ray r meets any sphere centered at $\mathbf{p}$ in a unique point. Because C is compact and convex, r also meets B in a unique point. Explain why. This establishes the map. Once you have studied convex functions, you should prove that this map is continuous. This is used in Theorem 26.7.10. See Exercise 18.7.10 for a similar construction.

18.5 Carathéodory's Theorem

Next a beautiful and important theorem that tells us how many terms r we need in the convex combination of the convex hull of any set in $\mathbb{R}^n$. We have already used it in the proof of Theorem 18.4.9. The important fact is not so much the bound itself, but that there is a uniform bound for all points.

18.5.1 Theorem (Carathéodory's Theorem)**.** *If S is a set in $\mathbb{R}^n$ and $\mathbf{x}$ a point in the convex hull of S, then $\mathbf{x}$ can be written as a convex combination of at most $n+1$ points in S.*

Proof. Theorem 18.1.28 says any $\mathbf{x} \in K(S)$ can be written as a convex combination of points in S, but it does not give us a bound. We find a bound by arguing by contradiction. Assume there is a point $\mathbf{x} \in K(S)$ for which the shortest representation as a convex combination of points of S required N points, $N > n+1$, so

$$\mathbf{x} = \sum_{i=1}^{N} \lambda_i \mathbf{x}_i \text{ where all } \mathbf{x}_i \in S \text{ , and } \lambda_i > 0 \text{ , and } \sum_{i=1}^{N} \lambda_i = 1$$

Consider the $N-1$ points $(\mathbf{x}_i - \mathbf{x}_N)$, $1 \le i \le N-1$. Since $N-1 > n$, these points are linearly dependent in $\mathbb{R}^n$, so we write an equation of linear dependence (so not all the coefficients are 0)

$$\sum_{i=1}^{N-1} \gamma_i (\mathbf{x}_i - \mathbf{x}_N) = 0$$

or

$$\sum_{i=1}^{N} \gamma_i \mathbf{x}_i = 0$$

where we have set $\gamma_N = -\sum_i^{N-1} \gamma_i$.

The following argument will be used many times in this set of lectures so is well worth remembering. Let t be a real variable. For every $t \in \mathbb{R}$ we can write

$$\mathbf{x} = \sum_{i=1}^{N} (\lambda_i - t\gamma_i)\mathbf{x}_i$$

Setting $\eta_i(t) = \lambda_i - t\gamma_i$, and recalling that the λ_i are all positive, our goal is to find a value of t so that all the η_i are non-negative, and at least one is 0. For such value of t the $\eta_i(t)$ give a representation of $\mathbf{x}$ as a convex combination of at most $N - 1$ points of S, the desired contradiction. So look at the set I_+ of indices i where γ_i is positive. Since the sum of all the γ is 0, this set is non-empty. Consider the set of ratios λ_i/γ_i, $i \in I_+$. Pick an index i_0 for which this ratio is minimal, and let $t_0 = \lambda_{i_0}/\gamma_{i_0}$, so that $\eta_{i_0}(t_0) = 0$ and all the other η are non-negative. Then $\mathbf{x}$ is a convex combination of fewer than N of the $\mathbf{x}_i$, the desired contradiction. □

18.5.2 Corollary. *If the dimension of the convex hull of S is $d < n$ then the estimate in Carathéodory's Theorem improves to $d + 1$.*

18.5.3 Exercise. According to Definition 18.3.10, the n-simplex in $\mathbb{R}^n$ is the convex combination of its $n + 1$ vertices S spanning $\mathbb{R}^n$. Show that there are points in the simplex that are not a convex combination of fewer than $n + 1$ points, showing that Carathéodory's Theorem gives the best general bound for the number of points needed.

18.5.4 Definition. Let the polytope P (see Definition 18.3.7) in $\mathbb{R}^n$ be the convex hull of its m extreme points $\mathbf{a}^1, \dots, \mathbf{a}^m$. Without loss of generality we can assume that the dimension of P is n, so $m \geq n+1$. To each set of $n+1$ affinely independent subsets of the m points $\mathbf{a}^i$ we can associate a simplex S_j with that set of points as vertices. These simplices are called the *subsimplices* of P.[6]

Note that a simplex has only one subsimplex: itself.

18.5.5 Corollary. *A polytope is the union of its subsimplices.*

Proof. Just use the main argument in the proof of the theorem. □

18.5.6 Example. Find the subsimplices of the cube and the crosspolytope (see 18.3.15) in $\mathbb{R}^3$.

[6] Definition 19.3.10 does something similar in the context of cones.

18.6 Separation Theorems

A hyperplane $H_{\mathbf{a},c}$ divides $\mathbb{R}^n$ into two half-spaces $H^+_{\mathbf{a},c}$ and $H^-_{\mathbf{a},c}$. Our goal is to show that any two disjoint convex sets in $\mathbb{R}^n$ can be separated by a hyperplane. We first state carefully what we mean by *separation*.

Let S and T be two sets in $\mathbb{R}^n$.

18.6.1 Definition. S and T are *separated* by $H_{\mathbf{a},c}$ if

$$\begin{aligned}\langle \mathbf{a}, \mathbf{s}\rangle &\geq c, \text{ for all } s \in S,\\ \langle \mathbf{a}, \mathbf{t}\rangle &\leq c, \text{ for all } t \in T.\end{aligned}$$

In other words, S is contained in the closed half-space $H^+_{\mathbf{a},c}$ and T in the open half-space $H^-_{\mathbf{a},c}$. In particular the two closed half-spaces $H^+_{\mathbf{a},c}$ and $H^-_{\mathbf{a},c}$ are separated by $H_{\mathbf{a},c}$, even though their intersection is precisely $H_{\mathbf{a},c}$. Even worse, the hyperplane $H_{\mathbf{a},c}$ is separated from itself.

Here is a separation definition with a stronger requirement:

18.6.2 Definition. S and T are *strictly separated* by $H_{\mathbf{a},c}$ if

$$\begin{aligned}\langle \mathbf{a}, \mathbf{s}\rangle &> c, \text{ for all } s \in S,\\ \langle \mathbf{a}, \mathbf{t}\rangle &< c, \text{ for all } t \in T.\end{aligned}$$

This definition implies that the sets S and T are disjoint. The open half-spaces $\mathring{H}^+_{\mathbf{a},c}$ and $\mathring{H}^-_{\mathbf{a},c}$ are strictly separated by $H_{\mathbf{a},c}$. Strict separation obviously implies separation.

18.6.3 Proposition. *Let C be a non-empty closed set, and let $\mathbf{b}$ be a point not in C. Then there is a point $\mathbf{c}_0 \in C$ that minimizes the distance $d(\mathbf{c}, \mathbf{b})$ between a point $\mathbf{c} \in C$ and $\mathbf{b}$. The minimum distance d is positive. If C is convex, the distance minimizing point $\mathbf{c}_0$ is unique.*

Proof. The function $D(\mathbf{x})$ giving the distance between $\mathbf{x}$ and $\mathbf{b}$ is continuous, as we showed in Example 11.1.3. Pick any point $\mathbf{c}$ in C, and let $D_0 = D(\mathbf{c})$ be its distance to $\mathbf{b}$. Now consider the set C_0 that is the intersection of C with the closed ball of radius D_0 centered at $\mathbf{b}$. Obviously the point that minimizes the distance between C and $\mathbf{b}$ is in C_0. The set C_0 is closed and bounded, therefore compact, so by the Weierstrass Theorem 16.2.2 there is a point $\mathbf{c}_0 \in C_0$ where the distance is minimum. This distance cannot be 0, as that would imply that $\mathbf{b}$ is in the closure of C: but C is closed and $\mathbf{b}$ is not in C, so that is impossible.

Now assume further that C is convex. If there are two distinct distance minimizing points $\mathbf{c}_0$ and $\mathbf{c}_1$ in C, then by convexity any point in the line segment

$[\mathbf{c}_0, \mathbf{c}_1]$ is in C. But since $\mathbf{c}_0$ and $\mathbf{c}_1$ are at the same distance from $\mathbf{b}$, the points in between are closer, which is a contradiction. This shows there is a unique minimizer. □

18.6.4 Corollary. *Let C be a non-empty closed and convex set, and let $\mathbf{b}$ be a point not in C. The point $\mathbf{c}_0$ found in Proposition 18.6.3 that minimizes the distance from $\mathbf{b}$ to C is the only point $\mathbf{c}$ of C satisfying*

$$\langle \mathbf{b} - \mathbf{c}, \mathbf{x} - \mathbf{c} \rangle \leq 0, \textit{ for all } \mathbf{x} \in C. \tag{18.6.5}$$

This says that for all $\mathbf{x} \in C$, the angle of vertex $\mathbf{c}_0$ and sides through $\mathbf{b}$ and $\mathbf{x}$ is at least a right angle.

Proof. First we show that the equation is satisfied for $\mathbf{c}_0$. Assume there is a point $\mathbf{x} \in C$ such that (18.6.5) fails, so the angle is acute. Then points on the open segment $(\mathbf{c}_0, \mathbf{x})$ close enough to $\mathbf{c}_0$ would be closer to $\mathbf{b}$ than $\mathbf{c}_0$. Since these points are in C by convexity, this is a contradiction.

To show that $\mathbf{c}_0$ is the only such point of C, suppose there is a second one, call it $\mathbf{c}$. The triangle with vertices $\mathbf{b}$, $\mathbf{c}_0$ and $\mathbf{c}$ would have two angles of at least $\pi/2$, an impossibility. □

18.6.6 Theorem. *Let C be a closed convex set, and $\mathbf{b}$ a point not in C. Then there is a hyperplane $H_{\mathbf{a},c}$ that strictly separates C and $\mathbf{b}$.*

Proof. Let $\mathbf{c}_0$ be the unique point in C that realizes the minimum distance d between $\mathbf{b}$ and a point of C, found in Proposition 18.6.3. Let $\mathbf{a} = \mathbf{c}_0 - \mathbf{b}$, and let $\mathbf{m}$ be the midpoint of the segment $[\mathbf{c}_0, \mathbf{b}]$, so $\mathbf{m} = (\mathbf{c}_0 + \mathbf{b})/2$. The equation of the hyperplane with normal vector $\mathbf{a}$ and passing through $\mathbf{m}$ is $\langle \mathbf{a}, \mathbf{x} \rangle = \langle \mathbf{a}, \mathbf{m} \rangle$. Letting $c = \langle \mathbf{a}, \mathbf{m} \rangle$, we claim that this hyperplane $H_{\mathbf{a},c}$ strictly separates C and $\mathbf{b}$.

It is clear by construction that b is in the open half-space $\mathring{H}^-_{\mathbf{a},c}$, and Corollary 18.6.4 shows that C is in $\mathring{H}^+_{\mathbf{a},c}$: indeed the hyperplane $H_{\mathbf{a},c}$ is perpendicular to $\mathbf{a} = \mathbf{c}_0 - \mathbf{b}$. □

18.6.7 Theorem. *Let C be convex, and $\mathbf{b}$ a point in $\mathbb{R}^n$ that is not in C. Then there is a hyperplane $H_{\mathbf{a},c}$ with $\mathbf{b}$ on the hyperplane, and C in one of the closed half-spaces bounded by $H_{\mathbf{a},c}$.*

Proof. The result follows immediately from Theorem 18.6.6 except when $\mathbf{b}$ is in the closure $\bar{C}$ of C. In that case we can find a sequence of points $\{\mathbf{b}_i\}$ in the complement of $\bar{C}$ converging to $\mathbf{b}$. By Lemma 18.4.1, $\bar{C}$ is convex.

We apply Theorem 18.6.6 to the point $\mathbf{b}_i$ and the closed convex set $\bar{C}$, getting a hyperplane $H_{\mathbf{a}_i,c_i}$ with $\bar{C}$ in the positive open half-space and b_i in the negative open half-space. As already noted, we may assume that $\mathbf{a}_i$ has length 1. Thus we can view each $\mathbf{a}_i$ as a point on the unit sphere, which is compact. By the fundamental Theorem 15.3.3 on sequences in compact sets, out of the sequence $\{\mathbf{a}_i\}$ we can extract a convergent subsequence which converges to a point $\mathbf{a}$ on the unit sphere. The sequence $\{c_i\}$ converges to $c = \langle \mathbf{a}, \mathbf{b} \rangle$.

As noted, for each index i in the subsequence and every $\mathbf{x}$ in C, $\langle \mathbf{a}_i, \mathbf{x} \rangle > c_i$. Thus in the limit we get $\langle \mathbf{a}, \mathbf{x} \rangle \geq c$, as desired. □

18.6.8 Theorem (The separating hyperplane theorem for disjoint convex sets)**.** *Let B and D be disjoint convex sets in $\mathbb{R}^n$. Then there is a hyperplane $H_{\mathbf{a},c}$ so that*

$$B \subset H^+_{\mathbf{a},c} \textit{ and } D \subset H^-_{\mathbf{a},c}.$$

B and D are separated by $H_{\mathbf{a},c}$. For all $\mathbf{b} \in B$ and $\mathbf{d} \in D$ we have

$$\langle \mathbf{a}, \mathbf{b} \rangle \geq c \geq \langle \mathbf{a}, \mathbf{d} \rangle.$$

Proof. We consider the set $S = B - D$ of Definition 18.1.31: the case $a = 1$ and $b = -1$. S is convex by Proposition 18.1.32, and does not contain the origin since B and D are disjoint. So apply Theorem 18.6.7 to S and the origin: there is a hyperplane $H_{\mathbf{a},c}$ containing the origin (which implies that $c = 0$) such that S is in the half-space $H^+_{\mathbf{a},0}$, so that

$$\langle \mathbf{a}, \mathbf{s} \rangle \geq 0, \text{ for any } \mathbf{s} \in S. \tag{18.6.9}$$

Recall that $\mathbf{s} = \mathbf{b} - \mathbf{d}$, for $\mathbf{b} \in B$ and $\mathbf{d} \in D$. If m is the greatest lower bound of $\langle \mathbf{a}, \mathbf{b} \rangle$, $\forall \mathbf{b} \in B$, and M is the least upper bound for all $\langle \mathbf{a}, \mathbf{d} \rangle$, $\forall \mathbf{d} \in D$, then (18.6.9) tells us that $m \geq M$. Then the hyperplane $H_{\mathbf{a},c}$, where c takes any value between m and M separates B and D. □

In Theorem 18.6.7 we required that the point $\mathbf{b}$ not be in the convex set C. However, notice that the only property of $\mathbf{b}$ we used is that there is a sequence of points $\{\mathbf{b}_i\}$ in the complement of the closure $\bar{C}$ of C converging to $\mathbf{b}$. This says that $\mathbf{b}$ is a boundary point of C.

18.6.10 Definition. Let C be a set, $\mathbf{x}$ a boundary point of C. Then a hyperplane $H_{\mathbf{a},c}$ that passes through $\mathbf{x}$ and has C in one of its half-spaces is called a *supporting hyperplane* to C at $\mathbf{x}$. Given a supporting hyperplane $H_{\mathbf{a},c}$ to C, the closed half-space (see Definition 18.1.7)$H^\pm_{\mathbf{a},c}$ containing C is called the *supporting half-space*.

In this new language, Theorem 18.6.7 says

18.6.11 Theorem. *A convex set C has at least one supporting hyperplane at each one of its boundary points.*

Theorem 22.1.2 will tell us that when the boundary of C is given by the graph of a differentiable function, the supporting hyperplane is unique.

18.6.12 Corollary. *A closed set X in $\mathbb{R}^n$ is convex if and only if it is the intersection of all its supporting half-spaces.*

Proof. If X is convex, for any point $\mathbf{b}$ not in X, take the separating hyperplane through the unique point of X at minimum distance from $\mathbf{b}$: see Corollary 18.6.4. The corresponding half-space will not contain $\mathbf{b}$. If X is not convex, there exists a point $\mathbf{b}$ not in X but in the convex hull of X. It is not possible to separate $\mathbf{b}$ from X. □

18.6.13 Definition. A half-space $H^-_{\mathbf{a},c}$ is called a *support* for a set S if $S \subset H^-_{\mathbf{a},c}$.

This definition allows us to state the following

18.6.14 Corollary. *The convex hull of a set S is the intersection of the supports of S.*

The proof is an exercise for the reader.

18.7 Minkowski's Theorem

18.7.1 Theorem (Minkowski's Theorem)**.** *Let C be a compact convex set, and let E be the set of extreme points of C. Then E is non-empty and C is the convex hull of E.*

Proof. We prove this by induction on the dimension of the convex set C. The result is clear if C has dimension 1 - and therefore is a closed interval: every point in C is in the convex hull of the two end points of the interval. Assume the result true for dimension $n - 1$. Let C be a compact convex set of dimension n.

First let $\mathbf{x}$ be a boundary point of C, and H a supporting hyperplane of C through $\mathbf{x}$. $C \cap H$ is a compact convex set of dimension at most $n - 1$. Thus by induction, $\mathbf{x}$ can be written as a convex combination of extreme points of $C \cap H$. But an extreme point of $C \cap H$ is an extreme point of C: $\mathbf{x}$ is not an interior point of a segment $[\mathbf{a}, \mathbf{b}] \in H$, because $\mathbf{x}$ is extreme in $C \cap H$. On the other hand $\mathbf{x}$ is not an interior point of a segment $[\mathbf{a}, \mathbf{b}]$ transverse to H, thus meeting H just in the point $\mathbf{x}$, since H is a supporting hyperplane of C, so that C is contained in one of the closed half-planes delimited by H.

Now assume $\mathbf{x}$ is not a boundary point of C: take any line ℓ through $\mathbf{x}$. Because C is compact, ℓ intersects C in two points $\mathbf{x}_1$ and $\mathbf{x}_2$ in the boundary of C, with $\mathbf{x} \in [\mathbf{x}_1, \mathbf{x}_2]$. By the previous argument, $\mathbf{x}_1$ and $\mathbf{x}_2$ are in the convex hull of extreme points, so is $\mathbf{x}$. □

As Example 18.1.13 shows, the number of extreme points of a compact convex set need not be finite. By Definition 18.3.7 a polytope has only a finite number of extreme points, and Corollary 18.7.5 shows the same is true for polyhedra.

18.7.2 Definition. If the compact convex set C has a finite number of extreme points, each extreme point of C is called a *vertex*.

We will use the word vertex and extreme point interchangeably.

18.7.3 Theorem. *Let $\mathbf{p}$ be a boundary point of the polyhedron $P = P(A, \mathbf{b})$. Then $\mathbf{p}$ is an extreme point of the polyhedron if and only if the normal vectors of the constraints that are active at $\mathbf{p}$ span $\mathbb{R}^n$. In particular there must be at least n active constraints at $\mathbf{p}$ for it to be an extreme point.*

Proof. First assume that the active normal vectors do not span. Then the intersection of the hyperplanes $H_{\mathbf{a}^i, b_i}$ is a positive dimensional linear space containing $\mathbf{p}$. So we can find a line segment $\mathbf{p} + t\mathbf{u}$, $-\epsilon \leq t \leq \epsilon$, $\epsilon > 0$ in P so $\mathbf{p}$ is not extreme.

Next assume that the active $\mathbf{a}^i$ at $\mathbf{p}$ span. Assume that $\mathbf{p}$ is not extreme, and derive a contradiction. If $\mathbf{p}$ is not extreme, we can find $\mathbf{q}$ and $\mathbf{r}$ different from $\mathbf{p}$ in P with

$$\mathbf{p} = \frac{\mathbf{q}}{2} + \frac{\mathbf{r}}{2}. \tag{18.7.4}$$

For each active i, we have

$$\begin{aligned}
\langle \mathbf{a}^i, \mathbf{q} \rangle &\leq b_i, && \text{because } \mathbf{q} \in P;\\
\langle \mathbf{a}^i, \mathbf{r} \rangle &\leq b_i, && \text{because } \mathbf{r} \in P;\\
\langle \mathbf{a}^i, \mathbf{p} \rangle &= b_i, && \text{because } i \text{ is active at } \mathbf{p};\\
\frac{1}{2}\langle \mathbf{a}^i, \mathbf{q} \rangle + \frac{1}{2}\langle \mathbf{a}^i, \mathbf{r} \rangle &= b_i, && \text{by (18.7.4).}
\end{aligned}$$

Thus, for each active constraint, we have

$$\langle \mathbf{a}^i, \mathbf{p} \rangle = \langle \mathbf{a}^i, \mathbf{q} \rangle = \langle \mathbf{a}^i, \mathbf{r} \rangle = \mathbf{b}_i.$$

Since the active $\mathbf{a}^i$ span, the system of all $\langle \mathbf{a}^i, \mathbf{x} \rangle = \mathbf{b}_i$ has only one solution, so $\mathbf{q}$ and $\mathbf{r}$ are not distinct from $\mathbf{p}$. Thus $\mathbf{p}$ is extreme. □

18.7.5 Corollary. *A polyhedron has either no extreme points, or a finite number of extreme points.*

Proof. Theorem 18.7.3 tells us that the extreme points are the points where any set of at least n linear equations with linearly independent left-hand sides meet. For each set of n such equations there is at most one solution, so all in all there are only a finite number of solutions and therefore only a finite number of vertices. Indeed, the number of solutions is at most $\binom{m}{n}$. In particular if $m < n$ there are no extreme points. □

The corollary does not exclude the possibility that a polyhedron has no extreme points. Indeed, any polyhedron defined by fewer than n half-spaces has no extreme points.

18.7.6 Example. Consider the polyhedron P in $\mathbb{R}^3$ given by the inequalities $x \geq 0$, $y \geq 0$, $z \geq 0$ and $x + y \leq 3$, $-1 \leq z - x \leq 2$, and $y + z \leq 4$. We want to find the vertices of P, by considering the points in P where three inequalities with linearly independent directions vanish. Clearly the origin $\mathbf{0}$ is a vertex: it satisfies all the constraints and the three positivity constraints are active there. The next easiest vertices to find are those that are the intersection of two positivity constraints and one other equation. An easy computation gives the vertices $(1, 0, 0)$, $(0, 3, 0)$ and $(0, 0, 2)$. Next we find those where only one coordinate vanishes. Checking cases, we get $(1, 2, 0)$, $(2, 0, 4)$, $(3, 0, 4)$, $(3, 0, 2)$, $(0, 3, 1)$, $(0, 2, 2)$. There are no vertices where all three coordinates are non-zero: this is because the directions of the constraints (other than the positivity constraints) only span a 2-dimensional vector space. We end up with a compact polyhedron with 10 vertices: so it is the convex hull of these vertices.

The following corollary will be useful.

18.7.7 Corollary. *Let P be a non-empty polyhedron in $\mathbb{R}^n$ given as the intersection of m half-spaces $H^-_{\mathbf{a}^i, b_i}$. Assume that the normal vectors $\mathbf{a}^i$ span $\mathbb{R}^n$, so that $m \geq n$. Then P has at least one extreme point.*

Proof. Pick a collection of n normal vectors that form a basis of $\mathbb{R}^n$. By reordering the half-spaces, we can assume they are $\mathbf{a}^i$, $1 \leq i \leq n$. The polyhedron P_0 which is the intersection of these n half-spaces clearly has a unique extreme point: the intersection $\mathbf{p}$ of the n linearly independent hyperplanes $H_{\mathbf{a}^i, b_i}$, $1 \leq i \leq n$. Next define the polyhedron P_1 to be the intersection of P_0 with $H_{\mathbf{a}^{n+1}, b_{n+1}}$. If $\mathbf{p}$ is in $H_{\mathbf{a}^{n+1}, b_{n+1}}$, it is an extreme point of P_1. Otherwise linear algebra tells us that we can find a subset of $n - 1$ of the n half-spaces defining P_0, such that their normal vectors and $\mathbf{a}^{n+1}$ form a basis of $\mathbb{R}^n$. The intersection point $\mathbf{p}_1$ of the corresponding hyperplanes is an extreme point of P_1. Continuing in this way, adding one half-space at a time, gives the result. □

We will use this result in Theorem 19.6.13. Compare it to Exercise 18.7.10.

18.7.8 Theorem. *A bounded polyhedron P is a polytope.*

Proof. Our goal is to apply Minkowski's Theorem 18.7.1. Since P is the intersection of m half-spaces given by $\mathbf{a}^i \cdot \mathbf{x} \le b_i$, P is closed. Since it is bounded, it is compact. Since it is a polyhedron, it is convex. Minkowski's Theorem tells us that P is the convex hull of its extreme points, which are finite in number by Corollary 18.7.5. Thus P is a polytope. □

We will prove the converse later: Corollary 20.2.14. We already proved the result for simplices in Example 18.3.18.

18.7.9 Exercise. Prove the following result. If C is a compact convex set, then a point $\mathbf{p} \in C$ at maximum distance from the origin is an extreme point of C. There is nothing special about the origin in this statement: any point will do.

Hint: If $\mathbf{p}$ is not extreme, then it is *between* two points $\mathbf{q}$ and $\mathbf{r}$ in C. A little geometry in the plane spanned by the three points $\mathbf{q}$, $\mathbf{r}$ and $\mathbf{0}$ gives the result.

18.7.10 Exercise. Prove that a closed convex set C has an extreme point if and only if it does not contain a line.

Hint: First assume C contains a line L. A point on the line clearly is not an extreme point. Pick a point $\mathbf{p}$ off the line that is extreme. Then Exercise 18.1.33 shows that in the plane spanned by L and $\mathbf{p}$, C contains a strip bounded by L on one side, and by the line L' parallel to L through $\mathbf{p}$ on the other. Because C is closed, L' is in C, and $\mathbf{p}$ is not extreme. This contradiction proves the result.

Now assume that C does not contain a line. Pick a point $\mathbf{q}$ in C. We now construct a function whose domain is the set of lines ℓ through $\mathbf{q}$. This set of lines is compact, by an argument similar to the one used in the proof of Theorem 9.1.2. Consider the function $d(\ell)$ that associates to ℓ the distance from $\mathbf{q}$ to the closest point where ℓ intersects the boundary of C. Since C contains no lines, this distance is finite. Show $d(\ell)$ is continuous, so it has a maximum. Conclude using Exercise 18.7.9.

18.8 Convexity Applied to Permutation Matrices

Consider the vector space of all real matrices of size $n \times n$. It has dimension n^2, with coordinates the entries x_{ij} of the general $n \times n$ matrix X . Inside this space, we can look at the following set:

18.8.1 Definition. A *doubly stochastic* matrix is a square matrix P with non-negative real entries such that the sum of the elements in each row and in each column add up to 1.

Thus the doubly stochastic $n \times n$ are defined by the $2n$ linear equations: $R_i \colon \{\sum_{j=1}^n x_{ij} = 1\}$, and $C_j \colon \{\sum_{i=1}^n x_{ij} = 1\}$, and the n^2 inequalities $x_{ij} \geq 0$.

We need to study the system of linear equations R_i and C_j. In the 2×2 case, where we have 4 variables and 4 equations: we list the variables in the order x_{11}, x_{12}, x_{21}, x_{22}, and list the row equations first, and then the column equations. A moment's thought will convince you that we get the matrix (1.4.14) from the transportation problem.

The first question to ask is: what is the rank of the system of equations R_i and C_j? It is clearly at most $2n-1$, because we have the relation $\sum_{i=1}^n R_i = \sum_{j=1}^n C_j$. The proof of Theorem 26.2.1 will show that the rank is exactly $2n-1$. Thus the doubly stochastic matrices live in an affine subspace L of dimension $n^2 - 2n + 1 = (n-1)^2$, and are defined by n^2 inequalities inside L.

18.8.2 Theorem. *Doubly stochastic matrices form a convex and compact subspace D in the space of all $n \times n$ matrices.*

Proof. This is a good exercise in the techniques we have developed. First we show they form a convex set. Let A and B be two doubly stochastic matrices. We must show that for every t, $0 < t < 1$, the matrix $C = tA + (1-t)B$ is doubly stochastic. This is obvious since the ij-th entry c_{ij} of this matrix is $ta_{ij} + (1-t)b_{ij}$, and is therefore a convex combination of a_{ij} and b_{ij}, and thus $0 \leq c_{ij} \leq 1$. Furthermore

$$\sum_{i=1}^n c_{ij} = t \sum_{i=1}^n a_{ij} + (1-t) \sum_{i=1}^n b_{ij} = t + (1-t) = 1$$

so the column sums are equal to 1. An obvious modification of this argument shows the row sums are also 1, so the set is convex.

The doubly stochastic matrices are clearly bounded (since all entries are between 0 and 1) so we need to prove that they are closed. Let $A(n)$ be a convergent sequence of doubly stochastic matrices. We must show that the limit is doubly stochastic. This is true because equalities are preserved in the limit, and inequalities $a \leq b$ are too. □

Then by Minkowski's Theorem 18.7.1 the set of doubly stochastic matrices D is the convex hull of its extreme points. Furthermore, since D is a bounded polyhedron in L, it is a polytope by Theorem 18.7.8, so it has a finite number of extreme points. What are they?

Permutation matrices, which we looked at in §6.4, are obviously doubly stochastic. They are extreme points of D. Indeed, by Theorem 18.7.3, all we need to show is that the normal vectors to active constraints span the ambient space. Now we have $n^2 - n$ active constraints (all the zeroes of a permutation matrix) in a $(n-1)^2$-dimensional affine space L, and they clearly span. Thus the permutation matrices are extreme points.

The converse, showing that all the extreme points are permutation matrices, is harder. It is proved below.

18.8.3 Theorem. *The permutation matrices are the extreme points of the doubly stochastic matrices.*[7]

Proof. The proof is by induction on n. The case $n = 1$ is clear. Now consider the case n. Select an extreme point of the doubly stochastic matrices. By Theorem 18.7.3 the corresponding matrix X satisfies $(n-1)^2$ active constraints, so the matrix must have at least $(n-1)^2$ entries that are 0. By the pigeon-hole principle applied to the n rows of the matrix, because $(n-1)^2 > n(n-2)$, at least one row (say the i_0-th row) must contain at least $n-1$ zeroes. That is the maximum number of zeroes in a given row, since the sum of the entries in the row must be 1. This show that for some j_0, the entry $x_{i_0 j_0} = 1$. All the other entries in the i_0-th row and the j_0-th column must be 0, since we are dealing with a doubly stochastic matrix. Therefore, if you remove the i_0-th row and the j_0-th column from X, you get a $(n-1) \times (n-1)$ doubly stochastic matrix. It has at least $(n-1)^2 - 2(n-1) - 1 = (n-2)^2$ zeroes, so it is an extreme point of the set of $n-1 \times n-1$ doubly stochastic matrices. By induction, it is a permutation matrix of size $n-1$. Putting the missing row and column back in, we get a permutation matrix of size n, which completes the proof. □

Thus we have constructed a beautiful polytope with $n!$ vertices. When $n = 2$, the polytope consists of the matrices

$$\begin{bmatrix} a & 1-a \\ 1-1 & a \end{bmatrix}, 0 \le a \le 1.$$

This is just the unit interval in $L = \mathbb{R}$, with the vertices $a = 0$ and $a = 1$.

For $n = 3$, we get the matrices

$$\begin{bmatrix} a & b & 1-a-b \\ c & d & 1-c-d \\ 1-a-c & 1-b-d & a+b+c+d-1 \end{bmatrix},$$

[7] This theorem due to Dénes König and Garrett Birkhoff. It is called the Birkhoff-von Neumann Theorem in [4], Theorem II.5.2. Other references are [39], p.164, [9], Exercise 4.22, p.74. The proof here follows Barvinok [4].

where all entries of the matrix are non-negative, a polyhedron in $L = \mathbb{R}^4$.

Carathéodory's Theorem 18.5.1 implies that any $n \times n$ doubly stochastic matrix is a convex combination of at most $(n-1)^2 + 1$ permutation matrices.

Lecture 19

Convex Cones and the Farkas Alternative

As a first step to doing linear optimization, we study the feasible set of a linear optimization problem. We start with the set F of solutions $\mathbf{x}$ of the system of equations $A\mathbf{x} = \mathbf{b}$, $\mathbf{x} \succeq \mathbf{0}$, where A is an $m \times n$ matrix and all the variables x_i are constrained to be non-negative. Later in the lecture we consider more general F. Our final results on the structure of F is given in §19.6. In order to understand F, we introduce a new type of geometry called conical geometry in §19.2. It is closely related to convex geometry. The key to understanding F is the finite cone C_A associated to the matrix A, introduced in §19.3. Finite cones are the conical analogs of polytopes in convex geometry. First we prove that finite cones are convex in Proposition 19.3.4 and then the harder result that finite cones are closed in Corollary 19.4.5.

The main result of this lecture is the Farkas Alternative 19.5.1, a crucial ingredient in the proof of the Duality Theorem 25.5.1 of linear optimization. The two tools used in the proof of the Farkas Alternative are the separation theorem for convex sets, which we studied in §18.6, and the fact that finite cones are closed, mentioned above. We give three versions of the Farkas Alternative. Theorem 19.5.1 has the most important case. Corollary 19.7.9 contains a version with inequalities that can be reduced to the first version by introducing slack variables: see Definition 19.7.2. The most general version is discussed in §19.8: in it we have a mixture of constraint equalities and inequalities, and we only require that some of the coordinates of $\mathbf{x}$ be non-negative. To handle this last issue, we use another standard device: the replacement of a variable x_i that is not constrained to be non-negative by two non-negative variables u_i and v_i with $x_i = u_i - v_i$. This last section should be skipped on first reading.

19.1 Introduction

The linear algebra notational conventions of Appendix A.5 remain in force. We also use conventions from Appendix A.3: For example, we say a vector $\mathbf{x}$ is *positive* (resp. *non-negative*), and write $\mathbf{x} \succ \mathbf{0}$ (resp. $\mathbf{x} \succeq \mathbf{0}$) , if all its coordinates $x_j > 0$ (resp. $x_j \geq 0$).

We also need some notation and results from Lecture 18 on Convex Sets. The hyperplane in $\mathbb{R}^n$ with equation $\sum_{j=1}^n a_j x_j = c$ is written $H_{\mathbf{a},c}$ as noted in (18.1.5). Recall the Separation Theorem 18.6.8, the definition of a supporting hyperplane in Definition 18.6.10 and of an extreme point in Definition 18.1.10.

In linear algebra we solve the equation

$$A\mathbf{x} = \mathbf{b} \tag{19.1.1}$$

Here A is an $m \times n$ matrix, so $\mathbf{x}$ is an n-vector and $\mathbf{b}$ an m-vector.

As usual we write $\mathbf{a}_j = (a_{1j}, a_{2j}, \dots, a_{mj})$ for the j-th column of A. Then (19.1.1) can be written

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = \mathbf{b} \tag{19.1.2}$$

If $\mathbf{b}$ is not the zero vector, this can be solved if and only if $\mathbf{b}$ is a linear combination of the columns of A. If $\mathbf{b}$ is the zero vector, then the solutions $\mathbf{x}$ belong to the nullspace of the linear transformation associated to A. In this lecture we study the following

19.1.3 Problem. We require that the solutions $\mathbf{x}$ of the system (19.1.1) be non-negative. In other words, we study the system of equations

$$A\mathbf{x} = \mathbf{b}, \text{ with } \mathbf{x} \succeq \mathbf{0}.$$

When is the set of solutions F of this system not empty? How can it be described?

An important criterion is given by the Farkas Alternative 19.5.1, the main result of this lecture.

The geometric object in $\mathbb{R}^m$ represented by non-negative linear combinations of the columns $\mathbf{a}_j$ of A:

$$\{\mathbf{x} \in \mathbb{R}^n | x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n, x_i \geq 0 \text{ for all } i\}, \tag{19.1.4}$$

is called the finite cone on A, written C_A. We study it in §19.3.

The key, albeit elementary, remark is that Equation 19.1.2 has a solution $\mathbf{x} \succeq \mathbf{0}$ if and only $\mathbf{b}$ is in C_A. Finite cones are analogs of the polytopes studied in convex geometry: see Definition 18.3.7. Just as polytopes are the convex hull of a finite number of vectors, finite cones are the conical hull of a finite number of vectors: in our example the columns $\mathbf{a}_i$, $1 \leq i \leq n$ of A. We will see that this conical hull is a convex set in Proposition 19.3.4, so we will be able to use the results of Lecture 18.

19.2 Conical Geometry and Polyhedral Cones

In these lectures we reviewed linear algebra, and then introduced other related geometries: affine geometry and convex geometry. In this lecture we introduce one last geometry on the same model: conical geometry, the geometry of cones.

We will study conical analogs of polytopes and polyhedra. Because we will be working in the column space of the $m \times n$ matrix A, our cones will live in $\mathbb{R}^m$, rather than the usual $\mathbb{R}^n$.

19.2.1 Definition. A set $C \subset \mathbb{R}^m$ is a *cone* with vertex $\mathbf{0}$ if whenever $\mathbf{x}$ is in C, then $\lambda\mathbf{x}$ is also in C, for all real numbers $\lambda \geq 0$.

We will be especially interested in cones that are convex sets, which we call convex cones.

19.2.2 Proposition. *A set $C \subset \mathbb{R}^m$ is a* convex cone *if and only if*

1. *it is a cone*
2. *and*

$$\mathbf{x}_1 + \mathbf{x}_2 \in C \textit{ whenever } \mathbf{x}_1, \mathbf{x}_2 \in C. \tag{19.2.3}$$

Proof. A convex set obviously satisfies (19.2.3), so we only need to show that (19.2.3) implies that the cone C is a convex set: in other words we must show that for all $\mathbf{x}_1, \mathbf{x}_2 \in C$ and any λ, $0 < \lambda < 1$, $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ is in C. This is clear since the cone property implies that $\lambda\mathbf{x}_1$ and $(1-\lambda)\mathbf{x}_2$ are in C. □

19.2.4 Example. The simplest cone is a *ray*, defined in §9.1 as follows. Pick a non-zero vector $\mathbf{a}$. The *ray* $r_{\mathbf{a}}$ on $\mathbf{a}$ is the set of elements $t\mathbf{a}$, for $t \geq 0$.

If a cone contains a non-zero point $\mathbf{a}$, it contains the ray $r_{\mathbf{a}}$. Just as we defined extreme points in convex geometry, we define extreme rays in conical geometry.

19.2.5 Definition. A ray $r_{\mathbf{a}}$ in the cone C is an *extreme ray* if for any $\mathbf{x}$ and $\mathbf{y}$ in C, if $\mathbf{x} + \mathbf{y}$ is in $r_{\mathbf{a}}$, then both $\mathbf{x}$ and $\mathbf{y}$ are in $r_{\mathbf{a}}$.

19.2.6 Example. Consider C, the closed first octant $\{\mathbf{x} \in \mathbb{R}^3 \mid \mathbf{x} \succeq 0\}$. C is a convex cone. Its extreme rays are the coordinate axes. For example take the ray r on $(1,0,0)$. Can we write it as a sum of elements $\mathbf{x} = (x_1, x_2, x_3)$ and $\mathbf{y} = (y_1, y_2, y_3)$ in the first octant? This would force $x_i = y_i = 0$ for $2 \leq i \leq 3$, meaning that both $\mathbf{x}$ and $\mathbf{y}$ are on the ray r. Finally convince yourself that C has no other extremal rays.

In what follows, B is a $n \times m$ matrix, the transpose of the $m \times n$ matrix A we have been considering. We introduce B because our cones lie in $\mathbb{R}^m$, rather than $\mathbb{R}^n$, so the roles of m and n are interchanged. Write $\mathbf{b}^1, \dots, \mathbf{b}^n$ for the rows of B. They are m-vectors.

19.2.7 Theorem. *The polyhedron $P(B, \mathbf{0})$ (see Definition 18.3.16) is a closed, convex cone in $\mathbb{R}^m$ defined by n inequalities: $B\mathbf{x} \le \mathbf{0}$.*

Proof. Because we take the right-hand side of the defining equation of the polyhedron, to be $\mathbf{0}$, we get a cone, as the defining equations are homogeneous of degree 1. As noted in Definition 18.3.16, it is convex and closed, so we are done. □

Because a polyhedral cone is convex, it has a dimension d as a convex set. Of course $d \le m$, but we could have $d < m$ if the inequalities defining the polyhedron imply both $b_1x_1 + \cdots + b_mx_m \le 0$ and $-b_1x_1 - \cdots - b_mx_m \le 0$ for some combinations of the rows of B. We will usually, without warning, restrict ourself to the affine space of dimension d containing the polyhedral cone: the affine hull of the cone.

This allows us to make a definition:

19.2.8 Definition. Any cone that can be written as $P(B, \mathbf{0})$, for a matrix B with m columns, is a *polyhedral cone* in $\mathbb{R}^m$.

19.2.9 Theorem. *The extreme rays of the polyhedral cone $P(B, \mathbf{0})$ are the intersection of $m-1$ hyperplanes of the form $\mathbf{b}^j \cdot \mathbf{x} = 0$, where the vectors $\mathbf{b}^j$ are rows of B that are linearly independent in $\mathbb{R}^m$. In particular the matrix B must have at least $m-1$ columns for $P(B, \mathbf{0})$ to have an extreme ray.*

Proof. We modify the proof of Theorem 18.7.3. We see that $r_{\mathbf{p}}$ is an extreme ray of $P(B, \mathbf{0})$ if and only if the normal vectors $\mathbf{b}^j$ of the hyperplanes that are active at $\mathbf{p}$, meaning that $\mathbf{b}^j \cdot \mathbf{p} = 0$, span the orthogonal complement of $r_{\mathbf{p}}$ in $\mathbb{R}^m$. So there must be at least $m-1$ linearly independent active constraints at $r_{\mathbf{p}}$.

This is because any collection of hyperplanes of the form $\mathbf{b}^j \cdot \mathbf{x} = 0$ intersect at the origin. So if they also intersect at $\mathbf{p}$, then they intersect along the ray $r_{\mathbf{p}}$. So we proceed as in the proof of Theorem 18.7.3, but working in the affine hyperplane passing through $\mathbf{p}$ and perpendicular to $r_{\mathbf{p}}$. □

19.3 Finite Cones

Finite cones are our main concern in this lecture. They are the analog in conical geometry of polytopes (see Definition 18.3.7) in affine geometry.

19.3.1 Definition. A set C in $\mathbb{R}^m$ is a *finite* cone, if there is a finite set of m-vectors $\mathbf{a}_1, \dots, \mathbf{a}_n$ in $\mathbb{R}^m$ such that C can be written

$$C = \{\mathbf{x} \mid \mathbf{x} = \sum_{j=1}^{n} \lambda_j \mathbf{a}_j, \text{for } \lambda_j \geq 0\}.$$

Finite cones are sometimes called finitely generated cones.

Just as in Remark 18.3.8 we may assume that we have selected a minimal set of $\mathbf{a}_j$ generating the cone, meaning that if we remove one of them, the resulting cone is strictly smaller. Then the $\mathbf{a}_j$ are called the *generators* of the finite cone. Let A be the $m \times n$ matrix whose columns are the $\mathbf{a}_j$, and let λ be the n-vector $(\lambda_1, \dots, \lambda_n)$. Then

$$C = \{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{x} = A\lambda, \text{ for } \lambda \succeq \mathbf{0}\} \tag{19.3.2}$$

We call the cone C_A when we want to emphasize the dependence on A.

The generators of a finite cone are not uniquely determined, since one can obviously replace any generator by a positive multiple. One might hope that the rays supported by the generators are uniquely determined. Indeed one might hope that the generators of a finite cone are supported by the extreme rays of the cone: see Definition 19.2.5. That is not the case, as shown by Example 19.3.6. In §20.5, we will show when finite cones behave like polytopes.

19.3.3 Exercise. Define the *conical hull* of a finite collection of vectors, modeling your definition on that of the convex hull: see Definition 18.1.17.

19.3.4 Proposition. *Finite cones are convex.*

Proof. Let C be the finite cone C_A, so C is the set of points $A\mathbf{z}$, for all vectors $\mathbf{z} \succeq \mathbf{0}$, and the generators of C are the columns of A. Pick two arbitrary points $A\mathbf{z}^1$ and $A\mathbf{z}^2$ in C, so $\mathbf{z}^1 \succeq \mathbf{0}$ and $\mathbf{z}^2 \succeq \mathbf{0}$. We must show that the segment $(A\mathbf{z}^1, A\mathbf{z}^2) \subset C$. We parametrize the segment by λ, $0 < \lambda < 1$. Now

$$\lambda A\mathbf{z}^1 + (1-\lambda)A\mathbf{z}^2 = A(\lambda\mathbf{z}^1 + (1-\lambda)\mathbf{z}^2)$$

by linearity. Furthermore, since all the coefficients of $\mathbf{z}^1$ and $\mathbf{z}^2$ are non-negative, $\lambda\mathbf{z}^1 + (1-\lambda)\mathbf{z}^2 \succeq 0$, so this point is in the cone. □

19.3.5 Definition. The *dimension* of a finite cone C is the dimension of C considered as a convex set: see Definition 18.2.24.

19.3.6 Example. Here are some examples of cones generated by the coordinate vectors $\mathbf{e}^1, \dots, \mathbf{e}^m$, the standard basis of $\mathbb{R}^m$.

- The cone generated by the $\mathbf{e}^i$, $1 \le i \le m$, is the first octant. We determined its extreme rays in Example 19.2.6.

- The cone generated by $\mathbf{e}^1$, …, $\mathbf{e}^k$, $k < m$, has dimension k. Its extremal rays are the $\mathbf{e}^i$, $1 \le i \le k$.

- The cone generated by $\pm\mathbf{e}^i$, $1 \le i \le k$, is a linear subspace of dimension k of $\mathbb{R}^m$. The next example shows that this is a minimal set of generators of the cone, since removal of any of the generators results in a smaller cone. There are no extreme rays when $k \ge 2$. If $k = 1$ we have a line, and both rays of this line are extreme.

- The cone generated by $\mathbf{e}^{i_0}$ and $\pm\mathbf{e}^i$, for $i \ne i_0$, is the closed half-space $H^+_{\mathbf{e}^{i_0},0}$. If $m > 2$, there are no extreme rays; if $m = 2$, so we are looking at a half-plane, there are two extreme rays forming a line.

We are mainly interested in finite cones, but here is an example of a cone that is not finite.

19.3.7 Exercise. Consider the set C in $\mathbb{R}^3$ defined by $x_1^2 + x_2^2 \le x_3^2$ and $x_i \ge 0$ for $1 \le i \le 3$. Show that C is a cone. Draw it by considering its intersection with the plane $x_3 = r$, where $r \ge 0$. Is it convex? Show that it is not finite.

19.3.8 Example. The set C_1 in $\mathbb{R}^3$ with coordinates (x, y, z) given by

$$xy \ge z^2, x \ge 0, y \ge 0, z \ge 0$$

is a closed convex cone. Similarly the set C_2 in $\mathbb{R}^3$ with coordinates (x, y, z) given by

$$xy \ge -z^2, x \le 0, y \ge 0, z \ge 0$$

is a closed convex cone. Consider the Minkowski sum (see Definition 18.1.31) $M = C_1 + C_2$. M is convex by Proposition 18.1.32, and is a cone because both C_1 and C_2 are cones. It is clear that M is contained in the cone given by $z \ge 0$ and $y \ge 0$, since these inequalities are satisfied for all points in C_1 and in C_2. This example is closely related to Example 18.4.8, and can be worked out in the same way.

We now define the analog of a r-simplex: see Definition 18.3.10.

19.3.9 Definition. A finite cone C in $\mathbb{R}^m$ is *basic* if its generators are linearly independent vectors.

The generators $\mathbf{a}_j$, $1 \leq j \leq r$ of a basic cone C span a linear subspace V of dimension r in $\mathbb{R}^m$, and $C \subset V$. If we add the vertex $\mathbf{a}^0$ of the cone (the 0 vector), then the $r+1$ vectors $\mathbf{a}^0, \mathbf{a}^1, \dots \mathbf{a}^r$ are affinely independent by Proposition 18.2.16.

Let C be the finite cone C_A associated to the $m \times n$ matrix A whose columns are the generators of C. If C is basic, then A has rank n. In general, let $r \leq n$ be the rank of A. Now consider a subset $S = \{\mathbf{a}_{j_1}, \mathbf{a}_{j_2}, \dots, \mathbf{a}_{j_r}\}$ of r generators of C_A. Such a subset is called *basic* if the r vectors are linearly independent. Note that this implies that any remaining generator of C is linearly dependent on the ones in S.

With this notation, we have

19.3.10 Definition (Basic Subcones)**.** Let S_k be a basic subset of generator of C, and A_k the $m \times r$ submatrix of A whose columns are the elements in S. Then the finite cone C_k generated by the elements of S_k, namely the columns of A_k, is called a *basic subcone of* C.

19.3.11 Remark. Each basic C_k is a subset of C, and has dimension r. Furthermore C has a finite number of basic subcones: at most $\binom{n}{r}$.

Here are some examples:

19.3.12 Example. Let C be the finite cone in $\mathbb{R}^2$ generated by the three vectors $(1,0)$, $(0,1)$ and $(-1,0)$, which we call $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$. There are two pairs of linearly independent generators:

- $(\mathbf{a}_1, \mathbf{a}_2)$ generating the basic cone C_1.
- $(\mathbf{a}_2, \mathbf{a}_3)$ generating the basic cone C_2.

The remaining pair is not linearly independent: indeed the two rays $r_{\mathbf{a}_1}$ and $r_{\mathbf{a}_3}$ form a line.

As the next example 19.3.13 shows, there can be inclusion relations among the C_j.

19.3.13 Example. Let C be the finite cone in $\mathbb{R}^2$ generated by the three vectors $\mathbf{a}_1 = (1,0)$, $\mathbf{a}_2 = (0,1)$ and $\mathbf{a}_3 = (-1,1)$. The corresponding basic cones are called C_1, C_2 and C_3. In this example $C_2 = C$, so C itself can be generated by two vectors and is itself basic. Clearly $C_1 \subset C_2$ and $C_3 \subset C_2$.

Things get more complicated in $\mathbb{R}^3$.

19.3.14 Example. Consider the finite cone C_A in $\mathbb{R}^3$ generated by the four columns of the matrix

$$A = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

Note that any three generators are linearly independent. Show that the cone C is not equal to any one of its basic cones, and that C itself is not all of $\mathbb{R}^3$. Intersect C and its basic cones with the unit sphere S centered at the origin. We get a map (called *central projection*) from $C \smallsetminus \mathbf{0}$ to S by associating to each non-zero vector $\mathbf{a} \in C$ the vector $\mathbf{a}/\|\mathbf{a}\|$ in S. You can understand the cone C by understanding its image under central projection to S.

19.3.15 Exercise. What are the basic subcones of $\mathbb{R}^m$, considered as a cone, using only the generators $\pm\mathbf{e}^i$, $1 \le i \le m$, considered in Example 19.3.6? How many basic subcones are there?

19.4 Finite Cones Are Closed

The main result of this section is Corollary 19.4.5, the key tool for our proof of the Farkas Alternative. First, Examples 19.3.12 and 19.3.14 suggest the following theorem concerning the basis subcones (see Definition 19.3.10) of C.

19.4.1 Theorem. *Let C be a finite cone in $\mathbb{R}^m$, with associated basic cones C_k, $1 \le k \le q$. Then C is the union of its basic subcones, which are finite in number. In other words:*

$$C = C_1 \cup C_2 \cup \cdots \cup C_q.$$

Before proving this, we rephrase the theorem's statement. By Definition 19.3.1, if $\mathbf{b}$ is in C, it can be written as a non-negative linear combination of the generators of C. The theorem says that $\mathbf{b}$ can be written as a non-negative linear combination of a linearly independent subset of the generators.

Proof. We reduce this to the key argument of the proof of Carathéodory's Theorem 18.5.1. Pick a vector $\mathbf{b}$ in the cone C. Then we can write

$$\mathbf{b} = \sum_{j=1}^{n} \lambda_j \mathbf{a}_j, \qquad \text{for } \lambda_j > 0. \tag{19.4.2}$$

where possibly only a subset of the original generators appears in the sum. Further assume that this is the shortest representation of $\mathbf{b}$ as a sum of the $\mathbf{a}_j$ with positive coefficients. Finally, we assume that the $\mathbf{a}_j$ appearing are not linearly independent, and derive a contradiction. Write an equation of linear dependence

$$\sum_{j=1}^{n} \mu_j \mathbf{a}_j = 0, \tag{19.4.3}$$

where not all the μ_j are 0. Multiplying this equation by -1, if necessary, we may assume that at least one of the μ_j is positive. For a small enough positive real number t, subtract t times (19.4.3) from (19.4.2) to get

$$\mathbf{b} = \sum_{j=1}^{n} (\lambda_j - t\mu_j)\mathbf{a}_j$$

Just as in the proof of Carathéodory's Theorem, by a suitable choice of t we get a representation of $\mathbf{b}$ as a linear combination with positive coefficients of a smaller number of $\mathbf{a}_j$, a contradiction that proves the result. □

19.4.4 Theorem. *Basic cones are closed.*

Proof. We know that polyhedral cones are closed (see Theorem 19.2.7), so it is enough to show that a basic cone is a polyhedral cone. A basic cone is the analog of a simplex, so we will imitate the proof of Theorem 18.3.18 that shows that a simplex is a polyhedron. We may without loss of generality assume that the dimension of the cone is m: otherwise restrict to the smallest linear subspace containing the cone. Consider the $m + 1$ affinely independent points **0**, $\mathbf{a}_1, \dots, \mathbf{a}_m$, and take the m hyperplanes that pass through **0** (so that they are linear hyperplanes, and not only affine ones) and $m - 1$ of the remaining m points. These hyperplanes are supporting hyperplanes for the convex cone, so that by taking the intersection of the corresponding half-spaces, we see that the cone is contained in their intersection. By the same argument as in the proof of Theorem 18.3.18, any point in the intersection is in the cone, so we are done. □

19.4.5 Corollary. *Finite cones are closed.*

Proof. A finite cone is a finite union of its basic subcones by Theorem 19.4.1 and a finite union of closed sets is closed by Theorem 14.4.15, so we are done. □

19.5 The Farkas Alternative

We now answer our original question: when does (19.1.2) have a non-negative solution? The answer is given by the Farkas Alternative, the inequality form of

Corollary 7.2.4. The result follows easily from the Separation Theorems in 18.6 and Corollary 19.4.5.

Note the use of the 'exclusive' *or* in the statement: it means that either one or the other of the assertions of the alternative is true, but not both. See §2.1.2 for details. For a restatement in terms of necessary and sufficient conditions see Corollary 19.5.3.

19.5.1 Theorem (The Farkas Alternative)**.** *Let A be an $m \times n$ matrix and* $\mathbf{b}$ *an m-vector. Then either*

1. *The system of equations $A\mathbf{x} = \mathbf{b}$ has a solution $\mathbf{x} \succeq \mathbf{0}$,*

   or (exclusive)

2. *there is a vector $\mathbf{y}$ in $\mathbb{R}^m$ with*

$$\mathbf{y}^T A \succeq \mathbf{0} \qquad \textit{and } \mathbf{y}^T \mathbf{b} < 0.$$

Proof. We first translate the statement of the theorem into geometry: Let C_A be the finite cone in $\mathbb{R}^m$ generated by the columns $\mathbf{a}_j$ of A, as in Definition 19.3.1. Then either

1. $\mathbf{b}$ is in C_A. Indeed, $\mathbf{b} = A\mathbf{x} = \sum_{j=1}^n x_j \mathbf{a}_j$, so if all the x_j are non-negative, $\mathbf{b}$ is in the cone generated by the $\mathbf{a}_j$, and that is C_A, by definition.

   Or (exclusive)

2. there exists a hyperplane $H_{\mathbf{y},0}$ in $\mathbb{R}^m$ through the origin that separates C_A and the point $\mathbf{b}$. Indeed we require that C_A only be in the closed positive half-space $H^+_{\mathbf{y},0}$, while $\mathbf{b}$ must be in the open negative half-space $\overset{\circ}{H}{}^-_{\mathbf{y},0}$). We do not have strict separation because the origin, which is in C_A, is also in $H_{\mathbf{y},0}$.

Now we prove the result. By Proposition 19.3.4 C_A is convex, and by Theorem 19.4.5 C_A is closed.

If case 1 is false, we must show that case 2 is satisfied. Now the point $\mathbf{b}$ is not in C_A. By Corollary 18.6.4, there is a unique point $\mathbf{b}_m$ in C_A minimizing the distance of $\mathbf{b}$ to a point of C_A. The existence of $\mathbf{b}_m$ follows from the fact that C_A is closed, and the uniqueness from the fact that it is convex. The same corollary says that the hyperplane H through the point $\mathbf{b}_m$ and perpendicular to $\mathbf{b} - \mathbf{b}_m$ is a supporting hyperplane to C_A at $\mathbf{b}_m$. By construction $\mathbf{b}$ is in an open half-space bounded by H, since its distance to the closed set C_A is positive. To finish the proof, we need to show that the origin is on H. If the closest point $\mathbf{b}_m$ is the origin, there is nothing to prove. Otherwise the closest point is on a ray r of C_A, and is not the origin. Because $\mathbf{b}_m$ is the closest point, the entire ray is in H, so the origin is in H. So we have found a hyperplane H meeting the requirements of case 2.

Now we have to prove the other implication: we assume that case 2 is false, and show that case 1 is satisfied. This is easy, since the failure of case 2 means that $\mathbf{b}$ is not at positive distance from C_A. That just means, since C_A is closed, that $\mathbf{b}$ is in C_A, and that is case 1, so we are done. □

19.5.2 Remark. The entries of $\mathbf{y}$ are just the coordinates of a hyperplane in $\mathbb{R}^m$. We can replace $\mathbf{y}$ by $-\mathbf{y}$ without changing the hyperplane, which implies that the alternative should be valid even after this replacement. Thus the alternative in Farkas can be written

There is a vector $\mathbf{y}$ in $\mathbb{R}^m$ with $\mathbf{y}^T A \preceq \mathbf{0}$ and $\mathbf{y}^T\mathbf{b} > 0$.

Convince yourself that the Farkas alternative can be reformulated in the following way, which translates the 'exclusive or' into the more familiar language of a necessary and sufficient condition.

19.5.3 Corollary. *Let $\mathbf{a}_1$, $\mathbf{a}_2$, …, $\mathbf{a}_n$ be a collection of n vectors in $\mathbb{R}^m$. Let $\mathbf{b}$ be another vector in $\mathbb{R}^m$. A necessary and sufficient condition for writing $\mathbf{b}$ as a non-negative linear combination*

$$\mathbf{b} = \sum_{j=1}^{n} x_j \mathbf{a}_j \, , \, x_j \geq 0 \, , \forall j \tag{19.5.4}$$

is that for every vector $\mathbf{y}$ such that $\mathbf{y}^T\mathbf{a}_j \geq 0$ for $1 \leq j \leq n$, then $\mathbf{y}^T\mathbf{b} \geq 0$.

19.6 The Polyhedron Associated to the Farkas Theorem

The Farkas Alternative 19.5.1 tells us when the system

$$A\mathbf{x} = \mathbf{b} \, , \mathbf{x} \succeq \mathbf{0}, \tag{19.6.1}$$

has a solution in $\mathbf{x}$ for a given value $\mathbf{b}$. This system is often called the canonical system. In this section we fix a $\mathbf{b}$ for which there is a solution, and describe the full set of solutions F of (19.6.1) for that $\mathbf{b}$.

Recall that the columns of the $m \times n$ matrix A are the generators $\mathbf{a}_j$ of the cone C_A (see Definition 19.3.1) in $\mathbb{R}^m$. By definition, the system (19.6.1) has a solution if and only if $\mathbf{b} \in C_A$.

It could happen that the linear span W of the $\mathbf{a}_j$ is strictly smaller than $\mathbb{R}^m$, so its dimension is $p < m$. Since (19.6.1) has a solution by assumption, this implies $\mathbf{b} \in W$. Choose a basis for W and then extend it to a basis for $\mathbb{R}^m$. In this basis the last $m - p$ coordinates of the $\mathbf{a}_j$ and of $\mathbf{b}$ are 0, so we could replace $\mathbb{R}^m$ by W.

So without loss of generality, we may assume:

19.6.2 Assumption (Rank Assumption)**.** The $m \times n$ matrix A has rank exactly m. Thus the m rows of A are linearly independent, so $m \leq n$.

The rank assumption is harmless: we have simply written an equivalent set of constraints in a more concise way, by removing dependent rows of A. The feasible set is unchanged, and the new problem has the same solution as the old one.

The affine subspace V given by the solutions of the equations $A\mathbf{x} = \mathbf{b}$ has dimension $n - m$: each equation reduces the dimension by 1 by the rank assumption. The set F of all solutions of (19.6.1) is the intersection of V with the convex cone given by the *positive octant* $\mathbb{R}^n_+$ of vectors $\mathbf{x} \succeq \mathbf{0}$. F is therefore a polyhedron in V, so by Definition 18.3.16 it is closed and convex. It is non-empty if and only if V meets $\mathbb{R}^n_+$.

19.6.3 Example. Here is the most elementary example of this set-up: let A be the matrix written in block notation as $\begin{bmatrix} I & 0 \end{bmatrix}$, where I is the $m \times m$ identity matrix and 0 the $m \times (n - m)$ zero matrix. Then V is just the subspace of $\mathbf{x}$ such that $x_i = b_i$, for $1 \leq i \leq m$. The intersection of V with the positive octant is non-empty if and only if all the b_i are non-negative. If all b_i are non-negative, and k of them are positive, then the intersection has dimension k. So all possible dimensions between 0 and $n - m$ may occur.

Returning to the general case, we use linear algebra to study the polyhedron F. By the rank assumption we can use the m equations $A\mathbf{x} = \mathbf{b}$ to solve for m of the x_j in terms of the remaining $n - m$ variables. Computationally we do this by Gaussian elimination. We start with a variable x_{j_1} that appears (meaning its coefficient a_{1,j_1} is non-zero) in the first row of A. We divide this equation by a_{1,j_1}, and then use this new equation to eliminate x_{j_1} from the remaining equations. Then we repeat the process: find a variable x_{j_2} that appears in the second equation and continue, getting $x_{i_2}, \dots, x_{i_m}$. By backsubstitution we can then eliminate all but x_{i_k} from the k-th equation.

19.6.4 Definition. Any collection of m variables for which this elimination process is possible is called a set of *basic variables* for the system of equations. The remaining variables are called *free variables*, since for any values of the free variables one gets a solution of the system.

19.6.5 Exercise. Convince yourself that this use of the word basic agrees with Definitions 19.3.9 and 19.3.10. In Example 19.6.3, the only collection of basic variables are the first m variables.

The rank assumption guarantees that a set of basic variables exists. From elementary linear algebra we have:

19.6.6 Theorem. *The variables x_{j_1}, x_{j_2}, ..., x_{j_m} form a set of basic variables if and only if the submatrix of A formed by the columns of index j_1, ..., j_m, is invertible.*

By changing the indexing, we may assume that the basic variables are x_1, x_2, ..., x_m. After this reindexing, we get a system of equations $\begin{bmatrix} I & A' \end{bmatrix} \mathbf{x} = b'$ with the same solutions as (19.6.1). We have n inequalities in the remaining x_{m+1}, x_{m+2}, ..., x_n: first the $n - m$ obvious ones $x_j \geq 0$, $m + 1 \leq j \leq n$. The others come from the inequality $x_j \geq 0$ after solving for the basic variables x_j, $1 \leq j \leq m$. So F is a polyhedron defined by n inequalities inside the affine space V of dimension $n - m$.

If we assume instead that the basic variables are $x_{n-m+1}, x_{n-m+2}, \dots, x_n$, we get a system $\begin{bmatrix} A'' & I \end{bmatrix} \mathbf{x} = b''$ equivalent to (19.6.1).

19.6.7 Exercise. Study (19.6.1) when A is the matrix

$$\begin{bmatrix} a_{11} & a_{12} & 1 & 0 \\ a_{21} & a_{22} & 0 & 1 \end{bmatrix} \tag{19.6.8}$$

Solve for x_3 and x_4 in terms of x_1 and x_2 and find all the inequalities satisfied by x_1 and x_2. Is it possible to choose values for the a_{ij} so that the system does not satisfy the Rank Assumption 19.6.2?

Choose values so that the polyhedron F is empty; non-empty and compact; non-empty and unbounded.

Given a set of basic variables, we can study V and F by projection to the subspace of non-basic variables. By reordering the variables, we assume that the basic variables are the last m variables $x_{n-m+1}, \dots, x_n$. Consider the linear map from $\mathbb{R}^n$ to $\mathbb{R}^{n-m}$ obtained just by forgetting the last m variables. So the matrix of this linear transformation is the $(n - m) \times n$ matrix $\begin{bmatrix} I_{n-m} & 0 \end{bmatrix}$, where I_{n-m} is the identity matrix of size $n - m$ and 0 is the zero matrix of size $(n - m) \times m$. This is called a projection, and is a minor variant of the orthogonal projection[1] in Definition 7.6.1.

[1] In terms of the notation there, the subspace K is the span of the basic variables, and R the space of free variables. Instead of mapping to the entire space, we only map to the subspace R, which means that we omit the bottom submatrix $\begin{bmatrix} 0_{kr} & 0_k \end{bmatrix}$ of zeroes.

Here is why this projection map is useful. Rewrite system (19.6.1) as the equivalent system $[A, I_m]\mathbf{x} = \mathbf{b}$, where A is a new matrix of size $m \times (n-m)$ obtained by Gaussian elimination. Thus the last m variables are basic.

19.6.9 Theorem. *Let P be the projection from $\mathbb{R}^n$ to $\mathbb{R}^{n-m}$ obtained by omitting the last m variables from the system $\begin{bmatrix} A & I_m \end{bmatrix} \mathbf{x} = b$, where A is an $m \times (n-m)$ matrix. Then to each point $\mathbf{q}$ in $\mathbb{R}^{n-m}$, there is a unique point $\mathbf{p}$ in V whose projection $P(\mathbf{p})$ is $\mathbf{q}$. The image of F under projection is a polyhedron F' defined by the inequalities*

$$\sum_{j=1}^{n-m} a_{ij}x_j \leq b_i, \quad \textit{for } 1 \leq i \leq m;$$

$$x_j \geq 0, \quad \textit{for } 1 \leq j \leq n-m.$$

Furthermore the polyhedra F and F' have the same structure as convex sets.

Proof. The first statement just reiterates the fact that given any value for the free variables, there is a unique solution for the basic variables. The fact that the projection of F is a polyhedron follows from Theorem 18.1.29, as does the last statement, which simply means that if $\mathbf{p}_1$ and $\mathbf{p}_2$ are in F, and $\mathbf{q}_1$ and $\mathbf{q}_2$ are their images in F', then $\lambda\mathbf{q}_1 + (1-\lambda)\mathbf{q}_2$ is the image of $\lambda\mathbf{p}_1 + (1-\lambda)\mathbf{p}_2$, for any $0 \leq \lambda \leq 1$. The determination of the inequalities defining F' is left to you. □

19.6.10 Example. Here is a simple example where the projection can be visualized. Take $n = 3$ and $m = 1$, so that we have a single constraint, say, $x_1 + 2x_2 + 3x_3 = 6$, and of course the positivity constraints $\mathbf{x} \succeq \mathbf{0}$.

Then F is the intersection of the affine hyperplane H determined by the constraint with the first octant. Note that H intersects the x_1-axis at the point $(6, 0, 0)$, the x_2-axis at $(0, 3, 0)$ and the x_3-axis at $(0, 0, 2)$. All three points are in the first octant, since all their coordinates are non-negative. A moment's thought should convince you that F is the convex hull of the triangle in space bounded by the three points.

A basic submatrix is just a 1×1 submatrix that is non-zero. In our example all possible submatrices are basic, so we have 3 of them. We choose the third coordinate as basic submatrix. Thus to get to the inequality problem, we project by forgetting the x_3 coordinate. What is the projection F' of F? You should check that it is the convex hull of the three points $(6, 0)$, $(0, 3)$ and the origin in the plane. That is the simplex determined by three inequalities $x_1 \geq 0$, $x_2 \geq 0$, and $x_1 + 2x_2 \leq 6$. You should make a graph of F, F' and the projection.

Under the Rank Assumption 19.6.2, we make the definition:

19.6.11 Definition. Let $\mathbf{x}$ be a solution of (19.6.1). Then $\mathbf{x}$ is *basic* if it has at most m non-zero (and therefore positive) coordinates x_i corresponding to linearly independent columns $\mathbf{a}_i$ of A.

In terms of Definition 19.6.4, this says that there is a collection of basic variables so that the only non-zero coordinates of our basic solution correspond to these basic variables. It does not say that to each basic variable the coefficient of the basic solution is non-zero: in particular a basic solution could have more than $n - m$ zeroes.

In Example 19.6.3, a basic solution exists if $b_i \geq 0$ for $1 \leq i \leq m$. The solution is $x_i = b_i$, $1 \leq i \leq m$, and $x_j = 0$, $j > m$. If one of the $b_i = 0$, we get a basic solution with fewer than m non-zero entries.

In general, reordering the columns of A if necessary, so that the basic variables come first, we may assume that a basic solution is written $\mathbf{x} = [\mathbf{x}_B, \mathbf{0}]$ where $\mathbf{x}_B$ is the vector of first m entries of $\mathbf{x}$. By hypothesis the columns $\mathbf{a}_1, \dots, \mathbf{a}_m$ of A are then linearly independent.

So far we have only used linear algebra. Now, from our study of finite cones we get the important

19.6.12 Theorem. *If* (19.6.1) *has a solution, it has a basic solution.*

Proof. Indeed, by Theorem 19.4.1, if there is a solution, the vector $\mathbf{b}$ lies in at least one basic subcone, which is therefore generated by linearly independent columns of A. Writing $\mathbf{b}$ is terms of these generators, we get a solution with non-zero coordinates only along the generators of the basic subcone: therefore it is basic. □

Then, recalling the Definition 18.1.10 of an extreme point of a convex set, we get the equivalence:

19.6.13 Theorem. *Assume that the set of solutions F of the system of equations* (19.6.1) *is non-empty. Then the extreme points of the convex set F are precisely the basic solutions of the system. In particular F always has extreme points.*

Proof. We start with a basic solution $\mathbf{x}$ of the system of equations (19.6.1). Thus $\mathbf{x}$ has $r \leq m$ non-zero entries, and the corresponding r columns of the matrix A are linearly independent. Then, using the fact that A has rank m, to those r columns can be added $m - r$ additional columns, so that the m columns form a basis for the column space of A. This gives a set of basic variables, generally not uniquely determined if $r < m$. Then we can project V to the $\mathbb{R}^{n-m}$ of non-basic variables and F to F' as described above. Under this projection, the basic solution

$\mathbf{x}$ gets mapped to the origin in $\mathbb{R}^{n-m}$, which is clearly an extreme point of F', and therefore $\mathbf{x}$ is an extreme point of F by Theorem 19.6.9.

To go the other way, we start with an extreme point $\mathbf{p}$ of F, and must show that it corresponds to a basic solution. We could invoke Theorem 18.7.3, but it is easier to proceed directly. Consider the submatrix A' of A corresponding to the columns where $\mathbf{p}$ has non-zero (and therefore positive) entries. A' is a $m \times r$ matrix for some $r > 0$, and if $\mathbf{p}'$ is the corresponding subvector of $\mathbf{p}$, we have $A'\mathbf{p}' = \mathbf{b}$.

A' has trivial nullspace, as we now show: otherwise we can find a non-zero $\mathbf{q} \in \mathbb{R}^r$ with $A'\mathbf{q} = \mathbf{0}$. Then for small enough ϵ, the vector $\mathbf{p} + \epsilon\mathbf{q}$ has only positive entries, and is a solution of $A'\mathbf{x}' = \mathbf{b}$. Thus, extending the r-vector $\mathbf{p} + \epsilon\mathbf{q}$ to an n-vector by putting zeroes in the appropriate places, we get a solution to (19.6.1) for all small enough ϵ. This contradicts the assumption that $\mathbf{p}$ is extreme. So A' has trivial nullspace. This shows that $r \leq m$. If $r = m$, we are done. If $r < m$, complete the collection of linear independent columns of A already selected by adding $m - r$ additional columns in order to get a basis of the columns space. The corresponding variables are basic, and therefore the extreme point is a basic solution. □

If F is compact, then it is the convex hull of its extreme points by Minkowski's Theorem 18.7.1.

19.6.14 Example. We continue Example 19.6.10. F is compact, and is the convex hull of its three extreme points $(6, 0, 0)$, $(0, 3, 0)$ and $(0, 0, 2)$. If we project F to F' by forgetting the third coordinate, as above, the extreme point $(0, 0, 2)$ is projected to the origin, which is extreme in F', as claimed.

19.7 Slack Variables and a Generalization of the Farkas Alternative

In this section we examine the system of equations:

$$A\mathbf{x} \succeq \mathbf{b}, \text{ and } \mathbf{x} \succeq \mathbf{0}, \tag{19.7.1}$$

often called the standard system, to distinguish it from the canonical system 19.6.1: We prove a version of the Farkas Theorem for it, and we also show that the polyhedron of solutions always has extreme points.

Later in this lecture we will generalize both systems in §19.8. The standard and canonical cases are done separately because of their intrinsic importance and because the notation of the general case is forbidding.

We reduce the system of equations (19.7.1) to the canonical case by introducing new variables, called slack variables, into each inequality to get back to the case of equality.

19.7.2 Definition. Consider the inequality $\sum_{j=1}^{n} a_j x_j \geq b$. Introduce a new variable z, and replace the inequality by the system

$$\sum_{j=1}^{n} a_j x_j - z = b \quad \text{with} \quad z \geq 0 \tag{19.7.3}$$

The two systems have the same solutions in $\mathbf{x}$. The new variable z, which takes up the slack between the inequality and the equality, is called a *slack* variable.

We rewrite the canonical system as the $n+m$ inequalities

$$\begin{aligned} \sum_{j=1}^{n} a_{ij} x_j &\geq b_i, \quad \text{for } 1 \leq i \leq m; \\ x_j &\geq 0, \quad \text{for } 1 \leq j \leq n. \end{aligned} \tag{19.7.4}$$

We introduce m new slack variables z_i, $1 \leq i \leq m$, so that z_i is associated with the i-th constraint equation, replacing (19.7.4) by

$$\begin{aligned} \sum_{j=1}^{n} a_{ij} x_j - z_i &= b_i, \quad \text{for } 1 \leq i \leq m; \\ x_j &\geq 0, \quad \text{for } 1 \leq j \leq n; \\ z_i &\geq 0, \quad \text{for } 1 \leq i \leq m. \end{aligned}$$

We can solve for the slack variables in terms of the x_j since the matrix of coefficients of the system is, in block notation, $[A, -I]$, where I is the $m \times m$ identity matrix, so that it is already diagonalized in the slack variables.

We easily see:

19.7.5 Proposition. *The system $A\mathbf{x} \succeq \mathbf{b}$, and $\mathbf{x} \succeq \mathbf{0}$ is equivalent to the canonical system*

$$A\mathbf{x} - I\mathbf{z} = \mathbf{b}, \ \textit{and}\ \mathbf{x} \succeq \mathbf{0}, \mathbf{z} \succeq \mathbf{0}.$$

By this we mean that if $\mathbf{x}$ is a solution of the first system, then there exist a unique $\mathbf{z}$ so that the pair $(\mathbf{x}, \mathbf{z})$ is a solution of the second system. Conversely, if $(\mathbf{x}, \mathbf{z})$ is a solution of the second system, then $\mathbf{x}$ is a solution of the first system.

Geometrically, this says that the feasible set F for the canonical problem, which is in $\mathbb{R}^{n+m}$, projects to the feasible set F' of the standard problem (in $\mathbb{R}^n$) under the map that forgets the last m coordinates. This is the same projection map we studied in Theorem 19.6.9.

19.7.6 Example. We look for the solutions of $x_1+2x_2 \geq 4$, $x_1 \geq 0$, $x_2 \geq 0$. Thus $n=2$, $m=1$, and $b=4$. The associated system with equality is $x_1+2x_2-z=4$, $x_1 \geq 0$, $x_2 \geq 0$, $z \geq 0$. To go from the inequality $x_1+2x_2 \geq 4$ to the equality, we just solve for $z = x_1+2x_2-4$. Clearly there is always a unique solution.

19.7.7 Remark. This pairs each one of our $n+m$ variables with one of our $n+m$ constraints.

- The original variable x_j, $1 \leq j \leq n$, is paired with the j-th constraint $x_j \geq 0$ in 19.7.4.
- The i-th slack variable z_i is paired with the i-th equation in (19.7.4), which just says $z_i \geq 0$.

19.7.8 Remark. The $m \times m$ submatrix of the new constraint matrix given by the slack variable columns is $-I$. So the constraint matrix has maximal rank m, an observation that will be useful later.

The corresponding version of the Farkas Theorem is:

19.7.9 Corollary. *Let A be an $m \times n$ matrix and $\mathbf{b}$ an m-vector. Then either*

1. $A\mathbf{x} \succeq \mathbf{b}$ *has a solution* $\mathbf{x} \succeq \mathbf{0}$

   or (exclusive)

2. *there is a vector* $\mathbf{y} \in \mathbb{R}^m$*,* $\mathbf{y} \succeq \mathbf{0}$*,* $\mathbf{y}^T A \preceq \mathbf{0}$ *and* $\mathbf{y}^T\mathbf{b} > 0$.

Proof. We apply Theorem 19.5.1, substituting the block matrix $\begin{bmatrix} A - I \end{bmatrix}$ for A and the vector $[\mathbf{x}, \mathbf{z}]$ for $\mathbf{x}$, and $n+m$ for n . Then the alternative to the existence of a solution is the existence of a $\mathbf{y}$ satisfying

$$\mathbf{y}^T \begin{bmatrix} A & -I \end{bmatrix} \preceq \mathbf{0}, \qquad \text{and} \qquad \mathbf{y}^T\mathbf{b} > 0.$$

The inequality on the left breaks up as two sets of inequalities : $\mathbf{y}^T A \preceq \mathbf{0}$, and $\mathbf{y}^T I \succeq \mathbf{0}$. The first forces $\mathbf{y}^T\mathbf{a}_j \preceq \mathbf{0}$. The last one says that $\mathbf{y} \succeq \mathbf{0}$. These inequalities are equivalent to the alternative given in the corollary. □

As for the earlier Farkas alternative (see 19.5.3), we will occasionally use it in the equivalent form:

19.7.10 Corollary. *Let $\mathbf{a}_1$, $\mathbf{a}_2$, …, $\mathbf{a}_n$ be a collection of n vectors in $\mathbb{R}^m$, viewed as the columns of a $m \times n$ matrix A. Let $\mathbf{b}$ be another vector in $\mathbb{R}^m$. A necessary and sufficient condition for the equation $A\mathbf{x} \succeq \mathbf{b}$ to have a solution $\mathbf{x} \succeq 0$ is that every vector $\mathbf{y} \succeq \mathbf{0}$ such that $\mathbf{y}^T A \preceq \mathbf{0}$, also satisfies $\mathbf{y}^T\mathbf{b} \le 0$.*

Theorem 19.6.9 tells us that the polyhedron defined by (19.7.1) always has extreme points. We will not make much use of this fact, because to compute with (19.7.1) we always use slack variables to transform it to a system of type 19.6.1, and then use Theorem 19.6.13.

19.7.11 Example. Suppose that $m = n = 2$, both variables x_1 and x_2 are non-negative, and our other constraints are

$$\begin{aligned} x_1 + 2x_2 &\ge 4 \\ 3x_1 + \ x_2 &\ge 3 \end{aligned}$$

Draw the feasible set, meaning the set in the plane satisfying the four inequalities. Its boundary is polygonal with vertices the origin, $\mathbf{a} = (0,3)$, $\mathbf{b} = (2/5, 9/5)$ and $\mathbf{c} = (4,0)$. The sides of the polygon are the vertical (x_2) axis, the segment $[\mathbf{a},\mathbf{b}]$, the segment $[\mathbf{b},\mathbf{c}]$ and the horizontal axis. Adding slack variables asks for the solution of

$$\begin{aligned} x_1 + 2x_2 - z_1 \qquad &= 4 \\ 3x_1 + \ x_2 \qquad - z_2 &= 3 \end{aligned}$$

in the first octant (where all variables are non-negative) in $\mathbb{R}^4$. In this case it is easier to see, without slack variables, that the feasible set is non-empty. Thus the alternative in Farkas must be false. So there is no non-negative vector $\mathbf{y} \in \mathbb{R}^2$ with

$$\begin{aligned} y_1 + 3y_2 &\le 0 \\ 2y_1 + \ y_2 &\le 0 \end{aligned}$$

and

$$4y_1 + 3y_2 > 0$$

This is easily seen by graphing the lines $y_2 = -\frac{1}{3}y_1$, $y_2 = -2y_1$, and $y_2 = -\frac{4}{3}y_1$. Question: how can you modify the vector $\mathbf{b} = (4,3)$ to make the alternative of Farkas true?

19.8 The General Farkas Alternative

Here is the most general system of equations for the feasible set of a linear optimization problem. The idea is simple: was allow a mixture of equalities and inequalities, and only force some of the variables to be constrained to be non-negative.

19.8.1 Definition. A is an $m \times n$ matrix, $\mathbf{b}$ a m-vector, and our variable $\mathbf{x}$ is an n-vector. The coordinates of $\mathbf{x}$ are indexed by the running variable j, $1 \leq j \leq n$. We pick an arbitrary subset $\mathcal{J}$ of the index set $\{1, 2, \dots, n\}$ and let $\mathcal{J}'$ be the complement of $\mathcal{J}$ in the index set, so $\mathcal{J} \cup \mathcal{J}' = \{1, 2, \dots, n\}$.

We also pick an arbitrary subset $\mathcal{I}$ of the index set $\{1, 2, \dots, m\}$ for the rows of A, using as running variable i, $1 \leq i \leq m$.

With this notation we write a set of equations derived from A and $\mathbf{b}$ as

- $\sum a_{ij}x_j \geq b_i$ when $i \in \mathcal{I}$,
- $\sum a_{ij}x_j = b_i$ when $i \notin \mathcal{I}$,
- $x_j \geq 0$ when $j \in \mathcal{J}$.

Let us call this set of equations the $(\mathcal{I}, \mathcal{J})$-system associated to A and $\mathbf{b}$. We denote it $(A, \mathbf{b}, \mathcal{I}, \mathcal{J})$.

This general form is given for completeness: we will almost always stick to the forms given in (19.6.1) and (19.7.1). In this section you will find a version of the Farkas Alternative for the general system.

19.8.2 Example. Here is how to describe the two key systems in this new notation.

- The canonical case treated in (19.6.1) corresponds to

  $\mathcal{I}$ empty, since there are only equalities, and

  $\mathcal{J} = (1, \dots, n)$, since all the variables x_j are required to be non-negative.

- The standard case treated in (19.7.1) corresponds to

  $\mathcal{I} = (1, \dots, m)$, since there are only inequalities, and

  $\mathcal{J} = (1, \dots, n)$, since all the variables x_j are required to be non-negative.

The goal is to reduce the system of equations $(A, \mathbf{b}, \mathcal{I}, \mathcal{J})$ to the canonical system. To do this we use two tools. The first one we have already met: slack variables: see Definition 19.7.2. The second tool, new to this section, is to replace a variable x, which is not constrained to be non-negative, as are the variables in our canonical equation, by two non-negative variables u and v, writing $x = u - v$.

19.8.3 Proposition. *Consider the system of equations $(A, \mathbf{b}, \mathcal{I}, \mathcal{J})$. Let m^* be the number of elements of $\mathcal{I}$, and n^* that of the complement $\mathcal{J}'$ of $\mathcal{J}$. For each $i \in \mathcal{I}$,* *we introduce a new slack variables z_i, which we view as the coordinates of an m^*-vector $\mathbf{z}$. For each $j \in \mathcal{J}'$ we introduce a new variable w_j which we view as the coordinates of an n^*-vector $\mathbf{w}$. Let $A_{\mathcal{J}'}$ denote the matrix formed by the columns of A corresponding to elements $j \in \mathcal{J}'$. So $A_{\mathcal{J}'}$ is an $m \times n^*$ matrix. Finally let $I_{\mathcal{I}}$ denote the $m \times m^*$ submatrix of the $m \times m$ identity matrix formed by only taking the columns of index $i \in \mathcal{I}$.*

Then the system $(A, \mathbf{b}, \mathcal{I}, \mathcal{J})$ is equivalent to the canonical system

$$A\mathbf{x} - A_{\mathcal{J}'}\mathbf{w} - I_{\mathcal{I}}\mathbf{z} = \mathbf{b} \textit{ and } \mathbf{x} \succeq \mathbf{0}, \mathbf{w} \succeq \mathbf{0}, \mathbf{z} \succeq \mathbf{0}. \tag{19.8.4}$$

If we rewrite the left-hand side of (19.8.4) *as block matrices, we get*

$$\begin{bmatrix} A & -A_{\mathcal{J}'} & -I_{\mathcal{I}} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{w} \\ \mathbf{z} \end{bmatrix} = \mathbf{b}, \textit{ and } \mathbf{x} \succeq \mathbf{0}, \mathbf{w} \succeq \mathbf{0}, \mathbf{z} \succeq \mathbf{0}.$$

Proof. The proof is just a matter of keeping track of indices. Notice that a variable x_j, not constrained to be non-negative (so $j \in \mathcal{J}'$), gets replaced by $x_j - w_j$. where both x_j and w_j are non-negative. □

Thus, at the cost of introducing a number of extra variables, we can reduce to a system of the form (19.6.1).

We now establish a generalization of the Farkas Alternative for our general system.

19.8.5 Theorem (The General Farkas Alternative)**.** *Let A be an $m \times n$ matrix,* $\mathbf{b}$ *an m-vector, $\mathcal{I}$ a subset of the index set $1 \le i \le m$, $\mathcal{J}$ a subset of the index set $1 \le j \le n$ and $\mathcal{J}'$ its complement. Then either*

1. *the $(\mathcal{I}, \mathcal{J})$-system $(A, \mathbf{b}, \mathcal{I}, \mathcal{J})$ has a solution* $\mathbf{x}$*,*

   or (exclusive)

2. *there is a vector $\mathbf{y} \in \mathbb{R}^m$, with (here $\mathbf{a}_j$ is the j-th column of A):*

   - $y_i \ge 0$*, if $i \in \mathcal{I}$;*
   - $\mathbf{y}^T\mathbf{a}_j \le 0$*, if $j \in \mathcal{J}$;*
   - $\mathbf{y}^T\mathbf{a}_j = 0$*, if $j \notin \mathcal{J}$; and*
   - $\mathbf{y}^T\mathbf{b} > 0$.

Proof. We apply Theorem 19.5.1, substituting $[A, -A_{\mathcal{J}'}, -I_{\mathcal{I}}]$ for A and $[\mathbf{x}, \mathbf{w}, \mathbf{z}]$ for $\mathbf{x}$, and $n + n^* + m^*$ for n . Then the alternative to the existence of a solution is the existence of a $\mathbf{y}$ satisfying

$$\mathbf{y}^T \begin{bmatrix} A & -A_{\mathcal{J}'} & -I_{\mathcal{I}} \end{bmatrix} \preceq \mathbf{0}, \qquad \text{and} \qquad \mathbf{y}^T\mathbf{b} > 0.$$

The inequality on the left breaks up as three sets of inequalities :

$$\mathbf{y}^T A \preceq \mathbf{0}, -\mathbf{y}^T A_{\mathcal{J}'} \preceq \mathbf{0}, \text{ and } \mathbf{y}^T I_{\mathcal{I}} \succeq \mathbf{0}.$$

The first two together force $\mathbf{y}^T\mathbf{a}_j = 0$ if $j \in \mathcal{J}'$. The last one says that $y_i \geq 0$ when $i \in \mathcal{I}$. These inequalities are equivalent to the alternative given in the corollary. □

19.8.6 Example. Let

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

Let $\mathcal{I} = (1, 2)$ and $\mathcal{J} = (3)$. So we are looking for the Farkas alternative to the statement

$A\mathbf{x} \geq \mathbf{b}$ has a solution $\mathbf{x}$ with $x_3 \geq 0$.

We form the matrix

$$\tilde{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & -a_{11} & -a_{12} & -1 & 0 \\ a_{21} & a_{22} & a_{23} & -a_{21} & -a_{22} & 0 & -1 \end{bmatrix}$$

and apply the usual Farkas alternative 19.5.1 to $\tilde{A}$ and $\mathbf{b}$, using $\mathbf{z}$ for the variable, so that $\mathbf{z} \geq \mathbf{0}$. Then, taking for example the first row $\tilde{\mathbf{a}}^1$ of $\tilde{A}$, we get

$$\langle \tilde{\mathbf{a}}^1, \mathbf{z} \rangle = a_{11}(z_1 - z_4) + a_{12}(z_2 - z_5) + a_{13}(z_3) - z_7$$

Let $x_1 = z_1 - z_4$, $x_2 = z_2 - z_5$ and $x_3 = z_3$. So $x_3 \geq 0$, while x_1 and x_2 have arbitrary sign.

By Theorem 19.8.5 the alternative is that there exists a $\mathbf{y} = (y_1, y_2)$ with $\mathbf{y}^T\tilde{\mathbf{a}}_j \leq 0$ and $\mathbf{y}^T\mathbf{a}_1 = 0$ and $\mathbf{y}^T\mathbf{a}_2 = 0$. Furthermore we require $\mathbf{y}^T\mathbf{b} > 0$. If the first two columns of A are linearly independent, then the only solution is $(0, 0)$.

Lecture 20

Polarity

In §20.1 we define support functions. In §20.2 the polar set is introduced and used to prove the Bipolar Theorem 20.2.13. Its Corollary 20.2.14 shows that polytopes are polyhedra.

In §20.3 we develop the notion of the conjugate of a function, which is the precise analog of the polar of a set, studied in §20.2. The transformation from a function to its conjugate is often called the Fenchel transform. It is a generalization of the better known Legendre transform used in physics. This will be useful when we study duality for convex functions later.

The material in §20.4 on polar cones is parallel to the material on convex sets in §20.2 and that on convex functions in §20.3. Finally, in §20.5, we associate to many finite cones a compact polytope. Alternate references for this material are [23], [52] and [7].

20.1 Support Functions

Expanding on Definition 18.6.13, here is a way of encoding the description of a closed convex set C as an intersection of hyperplanes. Another way will be discussed in §20.2. This material can be skipped on first reading.

Fix a vector $\mathbf{a}$, and look at all $c \in \mathbb{R}$ such that C is contained in the half-space $H^-_{\mathbf{a},c}$. In the terminology of Defintion 18.6.13, $H^-_{\mathbf{a},c}$ is a support for C. For some vectors $\mathbf{a}$, there may be no such c, in which case we set $c = \infty$. Otherwise, because there will be a best - namely smallest - c that works: the c that makes $H_{\mathbf{a},c}$ a supporting hyperplane with $C \in H^-_{\mathbf{a},c}$. The traditional way of writing this c is

$$c = \sup_{\mathbf{x} \in C} \langle \mathbf{a}, \mathbf{x} \rangle.$$

This allows us to make a

20.1.1 Definition. The *support function* of C is the function $s(\mathbf{y})$ on $\mathbb{R}^n$ given by

$$s(\mathbf{y}) = \begin{cases} \sup_{\mathbf{x}\in C}\langle \mathbf{y}, \mathbf{x}\rangle & \text{if the } \sup \text{ exists;}\\ \infty & \text{otherwise.}\end{cases}$$

The domain D of s is the set of $\mathbf{y}$ where s takes on finite values.

Support functions will be used in §20.3.

Note that D is a cone: if $\mathbf{y} \in D$ then $\lambda\mathbf{y} \in D$ for every positive λ. We will study cones in detail in Lecture 19.

20.1.2 Example. The domain of C is empty if $C = \mathbb{R}^n$, and a single point if C is a half-space. Indeed, if $C = H^-_{\mathbf{a},c}$, then the domain of its support function is the point $\mathbf{a}$ and its value at that point is c.

If C is the single point $\mathbf{x}$, then there is no need to take the $\sup$, and the support function is defined for all $\mathbf{y}$, with $s(\mathbf{y}) = \langle \mathbf{y}, \mathbf{x}\rangle$.

Note that the support function of a bounded set is defined for all $\mathbf{y}$.

20.1.3 Exercise. Show that the support function of the sphere of radius r centered at the origin, which is defined for all $\mathbf{y}$, is $s(\mathbf{y}) = r\|\mathbf{y}\|$.

20.1.4 Exercise. Find the support function of the segment in the plane bounded by the points $(\mathbf{a}_1, \mathbf{a}_2)$ and $(\mathbf{b}_1, \mathbf{b}_2)$.

Find the support function of the region above the parabola $y = x^2$ in the plane.

20.1.5 Exercise. Show that the support function is *positively homogeneous*, namely that $s(\lambda\mathbf{y}) = \lambda s(\mathbf{y})$ for every positive λ and any $\mathbf{y}$ in its domain. Note first that this is verified in all the previous examples.

20.1.6 Proposition. *The support function of a convex set is sublinear:*

$$s(\mathbf{a} + \mathbf{b}) \le s(\mathbf{a}) + s(\mathbf{b}).$$

Proof. Here is a sketch. The only content is when both $s(\mathbf{a})$ and $s(\mathbf{b})$ are finite. We need to show that

$$\sup_{\mathbf{x}\in C}\langle \mathbf{a} + \mathbf{b}, \mathbf{x}\rangle \le \sup_{\mathbf{x}\in C}\langle \mathbf{a}, \mathbf{x}\rangle + \sup_{\mathbf{x}\in C}\langle \mathbf{b}, \mathbf{x}\rangle.$$

Obviously

$$\langle \mathbf{a} + \mathbf{b}, \mathbf{x}\rangle = \langle \mathbf{a}, \mathbf{x}\rangle + \langle \mathbf{b}, \mathbf{x}\rangle \text{ , for any } \mathbf{x},$$

so since the $\sup$ is taken separately on the right-hand side, we get the desired inequality. □

Finally, the result that motivates the introduction of the support function:

20.1.7 Theorem. *If C is a closed convex set in $\mathbb{R}^n$ with support function $s(\mathbf{y})$, then*

$$C = \bigcap_{\mathbf{a} \in D} H^-_{\mathbf{a},s(\mathbf{a})}.$$

To handle the case where D is empty, the empty intersection is defined to be $\mathbb{R}^n$.

This is just a restatement of Corollary 18.6.12.

While we are mainly interested in the support function of a convex set, the definition makes sense for any set whatsoever. The support function will always be positively homogeneous and sublinear: indeed, convexity of the set was not used in establishing either property.

20.1.8 Example. The support function 20.1.1 of a convex set is a convex function.

Indeed, by definition, it is the $\sup$ of a collection of linear, hence convex functions, so this is a special case of Example 22.3.11.

20.1.9 Exercise. Let $s(\mathbf{y})$ be the support function of an arbitrary set $S \subset \mathbb{R}^n$. What is the set

$$\bigcap_{\mathbf{a} \in D} H^-_{\mathbf{a},s(\mathbf{a})}?$$

First work out what happens when S is just two distinct point $(\mathbf{a}_1, \mathbf{a}_2)$ and $(\mathbf{b}_1, \mathbf{b}_2)$ in the plane.

Hint: This set is an intersection of closed half-spaces, so it is both closed and convex. This is just a reformulation of Corollary 18.6.14.

20.2 Polarity for Convex Sets

The goal of this section is to build on Corollary 18.6.12, which says that a closed set S is convex if and only if it is the intersection of all the half-spaces containing it. How can we describe these half-spaces? We have already done this using support functions in §20.1. Here is a second approach that should be compared to the support function approach. This material should be skipped on first reading.

In Example 18.1.7 we wrote $H^-_{\mathbf{a},c} = \{\mathbf{x} \mid \langle \mathbf{a}, \mathbf{x} \rangle \leq c\}$. As long as $c \neq 0$, we get the same half-space by dividing the equation by c, so we look at: $H^-_{\mathbf{a},1} = \{\mathbf{x} \mid \langle \mathbf{a}, \mathbf{x} \rangle \leq 1\}$. This suggests that to the set S we associate all the vectors $\mathbf{a}$ so that S is contained in $H^-_{\mathbf{a},1}$. We make this into a definition:

20.2.1 Definition. Let S by a non-empty set in $\mathbb{R}^n$. Then the *polar set* S^* of S is given by

$$S^* = \{\mathbf{y} \in \mathbb{R}^n \mid \langle \mathbf{y}, \mathbf{x} \rangle \leq 1 \text{ for all } \mathbf{x} \in S\}. \tag{20.2.2}$$

Thus S lies in the intersection of the half-spaces $H^-_{\mathbf{y},1}$, for all $\mathbf{y} \in S^*$. Dually, S^* is the intersection of the half-spaces:

$$S^* = \bigcap_{\mathbf{x} \in S} H^-_{\mathbf{x},1}.$$

20.2.3 Example. If the set S contains a single point $\mathbf{a}$ other than the origin, then S^* is the closed half-space bounded by the hyperplane $H_{\mathbf{a},1}$ with equation $\langle \mathbf{a}, \mathbf{x} \rangle = 1$, that contains the origin.

If S only contains the origin, then S^* is all of $\mathbb{R}^n$.

20.2.4 Proposition. *If $S = \overline{N}_r(\mathbf{0})$, the closed ball of radius r centered at the origin, then $S^* = \overline{N}_{1/r}(\mathbf{0})$*

Proof. This follows from the Cauchy-Schwarz inequality 5.4.6. To test if a non-zero element $\mathbf{y}$ is in S^*, dot it with the unique element $\mathbf{x}$ on the same ray through the origin and on the boundary of S. Then $\|x\| = r$ and

$$\langle \mathbf{y}, \mathbf{x} \rangle = \|\mathbf{x}\| \|\mathbf{y}\| = r\|\mathbf{y}\| \leq 1$$

so $\|\mathbf{y}\| \leq 1/r$. If this is true, then the Cauchy-Schwarz inequality shows us that for any $\mathbf{x} \in \overline{N}_r(\mathbf{0})$,

$$\langle \mathbf{y}, \mathbf{x} \rangle \leq \|\mathbf{x}\| \|\mathbf{y}\| \leq 1,$$

as required. □

We have the elementary

20.2.5 Theorem. *If $\{S_\alpha\}$ is an arbitrary collection of sets indexed by α, then the polar of the union of the $\{S_\alpha\}$ is the intersection of the polars of the S_α.*

From this we deduce the useful:

20.2.6 Theorem. *The polar of an arbitrary set S is a closed and convex set containing the origin.*

Proof. Write S as the union of its points, and notice from Example 20.2.3 that the polar of a point is convex, closed and contains the origin. By Theorem 18.1.15, any intersection of convex sets is convex, and by Theorem 14.4.15, any intersection of closed sets is closed, so we are done. □

Another elementary consequence of Theorem 20.2.5 is

20.2.7 Theorem. *If $S \subset T$, then $T^* \subset S^*$.*

Proof. Because $S \subset T$, S^* is the intersection of a smaller number of half-spaces than T^*, so certainly $T^* \subset S^*$. □

20.2.8 Theorem. *Assume that the polytope P has the points $\mathbf{a}^0, \dots, \mathbf{a}^m$ as vertices. Then*

$$P^* = \{\mathbf{y} \mid \langle \mathbf{a}^i, \mathbf{y} \rangle \leq 1 \textit{ for all } i = 0, \dots, m\}. \tag{20.2.9}$$

Proof. This is easy. The right-hand side of (20.2.9) contains the left-hand side by the definition of the polar, so all we need is the opposite inclusion. So take any $\mathbf{y}$ satisfying the right-hand side inequalities. An arbitrary point in the polytope is given by (18.3.9). Dot this expression with $\mathbf{y}$ to get

$$\sum_{i=0}^{m} \lambda_i \langle \mathbf{a}^i, \mathbf{y} \rangle \leq \sum_{i=0}^{m} \lambda_i = 1,$$

since the λ_i are non-negative, and $\sum_{i=0}^{m} \lambda_i = 1$. Thus $\mathbf{y}$ is in P^*. □

Thus the polar of a polytope P_A is the polyhedron $P(A, \mathbf{1})$, using the notation of (18.3.9) and Definition 18.3.16.

The first important result of the section is

20.2.10 Theorem. *Let S be a compact and convex set of dimension n in $\mathbb{R}^n$ that contains the origin in its interior. Then S^* is a compact convex set of dimension n containing the origin in its interior.*

Proof. Theorem 20.2.6 tells us that S^* is closed, convex and contains the origin. Thus we need only prove that S^* is bounded and that the origin is an interior point. Because the origin is an interior point of S, for some small radius r, the ball $\overline{N}_r(\mathbf{0})$ is contained in S. But then S^* is contained in the ball $\overline{N}_{1/r}(\mathbf{0})$ by Proposition 20.2.4 and Theorem 20.2.7, which shows that S^* is bounded. Because S is compact it is bounded, so is a subset of $\overline{N}_R(\mathbf{0})$ for a large enough R. Proceeding as before, this shows that the ball $\overline{N}_{1/R}(\mathbf{0})$ is contained in S^*, showing that the origin is an interior point. □

The next step is to apply polarity twice.

20.2.11 Definition. The *bipolar* S^{**} of a set S as the polar of the polar of S, $S^{**} = (S^*)^*$.

Then in complete generality we have $S \subset S^{**}$. Indeed, rewrite (20.2.2) for S^*:

$$S^{**} = \{\mathbf{x} \in \mathbb{R}^n \mid \langle \mathbf{y}, \mathbf{x} \rangle \leq 1 \text{ for all } \mathbf{y} \in S^*\}. \tag{20.2.12}$$

Comparing this to (20.2.2) shows that if $\mathbf{x}$ is in S, then it is in S^{**}, so $S \subset S^{**}$.

Now the main result of this section.

20.2.13 Theorem (The Bipolar Theorem)**.** *Let S be a closed convex set containing the origin. Then the bipolar S^{**} of S is equal to S.*

Proof. We have just established the inclusion $S \subset S^{**}$. To get the opposite inclusion, pick a point $\mathbf{b}$ not in S. We must show it is not in S^{**}. Since S is convex and closed, by the Separation Theorem 18.6.6, we can find a hyperplane $H = \{\mathbf{x} \mid \langle \mathbf{a}, \mathbf{x} \rangle = 1\}$ strictly separating S and $\mathbf{b}$. Because $\mathbf{0} \in S$, we have $\langle \mathbf{a}, \mathbf{x} \rangle < 1$ for all $\mathbf{x} \in S$, and $\langle \mathbf{a}, \mathbf{b} \rangle > 1$. The first inequality says that $\mathbf{a}$ is in S^*, from which the second inequality say that $\mathbf{b}$ is not in S^{**}, and we are done. □

By this result and Theorem 20.2.6 we see that $S = S^{**}$ if and only if S is a closed convex set containing the origin.

20.2.14 Corollary. *A polytope is a polyhedron, and a bounded polyhedron is a polytope.*

Proof. The last statement is Theorem 18.7.8, so we need only prove the first one. By restricting to the affine hull of the polytope P, we can assume it has maximum dimension, so that it has a non-empty interior. Then by translating it, we can make the origin an interior point. Then by Theorem 20.2.10, P^* is compact, and by Theorem 20.2.8 it is a polyhedron, therefore a bounded polyhedron. So by Theorem 18.7.8, P^* is a polytope, so its polar $(P^*)^*$ is a polyhedron. By the Bipolar Theorem, $(P^*)^* = P$, so P is a polyhedron as claimed. □

We now see that bounded polyhedra and polytopes are the same. This result is known as the Weyl-Minkowski Theorem: see [73].

We could pursue this line of inquiry by determining the polar of a given polytope. Example 18.3.18 shows that the polar polytope of a simplex is again a simplex. This is investigated in [40], chapter 9, which is a good reference for the material in this section. A more advanced reference is [4], chapter IV.

20.2.15 Exercise. Show that the polar polytope of the cube is the crosspolytope. from which is follows that the polar polytope of the crosspolytope is the cube. First work this out in $\mathbb{R}^2$ and $\mathbb{R}^3$.

20.3 Conjugate Functions and Duality

This section is harder than the previous ones, and should be skipped on first reading. It is the analog for convex functions to §20.2 for convex sets.

We apply Example 22.3.11 to generate a new convex function from any function $f(\mathbf{x})$: the new function is called the conjugate, or the polar, of the original function. The conjugate of the conjugate is closely related to the original function, as always happens in duality.

Assume that $f(\mathbf{x})$ is defined on the set $D \subset \mathbb{R}^n$. We do not assume that f is convex. For each fixed vector $\mathbf{a} \in \mathbb{R}^n$, we ask: for which numbers c is the affine function $\langle \mathbf{a}, \mathbf{x} \rangle - c$ no bigger than $f(\mathbf{x})$:

$$\langle \mathbf{a}, \mathbf{x} \rangle - c \leq f(\mathbf{x}), \quad \forall \mathbf{x} \in D,$$

or, rearranging,

$$c \geq \langle \mathbf{a}, \mathbf{x} \rangle - f(\mathbf{x}), \quad \forall \mathbf{x} \in D.$$

By definition the best possible c is $\sup_{\mathbf{x} \in D}\{\langle \mathbf{a}, \mathbf{x} \rangle - f(\mathbf{x})\}$.

So to each vector $\mathbf{a}$ we can associate this number, which allows us to define a new function, called the *conjugate* of f,

$$f^*(\mathbf{y}) := \sup_{\mathbf{x} \in D}\{\langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x})\}. \tag{20.3.1}$$

The conjugate f^* is often called the Fenchel transform of f. It generalized the Legendre transformation for differentiable functions: see Rockafellar [53], Section 26, and [54], p. 16. Here is the Legendre transformation in its simplest form.

20.3.2 Definition (The Legendre Transform)**.** Assume that $f(\mathbf{x})$ is strictly convex and $\mathcal{C}^2$ on $\mathbb{R}^n$. In fact we will assume that the hessian of f is positive definite at every point. Since f is strictly convex, by Corollary 22.1.10 the map $\nabla f \colon R^n \to R^n$ that associates to $\mathbf{x} \in \mathbb{R}^n$ the gradient ∇f of f evaluated at $\mathbf{x}$ is one-to-one onto its image that we call D. Because f is $\mathcal{C}^2$, ∇f is $\mathcal{C}^1$, and its gradient is the $n \times n$ Hessian of f. Because we assume the Hessian is positive definite, the map ∇f satisfies the hypotheses of the inverse function theorem.

To compute the value of $f^*(\mathbf{y})$ at any point $\mathbf{y}$, we must maximize the function

$$\langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x}),$$

holding $\mathbf{y}$ fixed. This is an unconstrained maximization problem, whose solutions (by Theorem 13.1.2 are given by the solutions of the equations

$$\mathbf{y} - \nabla f(\mathbf{x}) = \mathbf{0}.$$

If $\mathbf{y}$ is in D, there is exactly one solution $\mathbf{x}_0$, otherwise there is none. When there is a solution, it is necessarily a maximum because the function $\langle \mathbf{y}, \mathbf{x}\rangle - f(\mathbf{x})$ is concave, as it is the sum of a linear function, which is both concave and convex, and the function $-f$ which is concave because f is convex. This is the content of Theorem 22.4.1.

In other words the mapping $\nabla f(\mathbf{x})$ has an inverse on D, which we write $\mathbf{s}(\mathbf{y})$. Thus the conjugate of f can be written:

$$f^*(\mathbf{y}) = \langle \mathbf{y}, \mathbf{s}(\mathbf{y})\rangle - f(\mathbf{s}(\mathbf{y})). \tag{20.3.3}$$

This is the *Legendre transform* of f. It shows that the conjugate function of a strictly convex differentiable function can be found by solving an optimization problem: we do this in Examples 20.3.7 and 20.3.8 below. We can also compute the gradient of $f^*(\mathbf{y})$ using the product rule and the chain rule. Because $\mathbf{s}$ is a vector function, its gradient $\nabla\mathbf{s}$ is actually a $n \times n$ matrix, which accounts for what happens to the dot product.

$$\nabla f^*(\mathbf{y}) = \mathbf{s}(\mathbf{y}) + \mathbf{y}^T \nabla\mathbf{s}(\mathbf{y}) - \nabla f(\mathbf{s}(\mathbf{y}))\nabla\mathbf{s}(\mathbf{y}) \tag{20.3.4}$$

You should check that all the terms in this equation have the right dimension. Since $\mathbf{s}(\mathbf{y})$ is the inverse function to $\nabla f(x)$, we have $\nabla\mathbf{s}(\mathbf{y}) = \mathbf{y}$, so the last two terms in (20.3.4) cancel, and we are left with

$$\nabla f^*(\mathbf{y}) = \mathbf{s}(\mathbf{y}).$$

Taking one more derivative, we see that the Hessian of $f^*(\mathbf{y})$ is the gradient of $\mathbf{s}(\mathbf{y})$. Since $\mathbf{s}$ is the inverse of ∇f, its gradient is the inverse of the Hessian of f: this is where we use our assumption that the Hessian of f is positive definite. But the inverse of a positive definite matrix is positive definite, so we saw in Corollary 9.2.10, so the Hessian of f^* is positive definite and f^* is strictly convex. Thus f^* satisfies the same hypotheses as f, so that we can compute its conjugate in the same way we computed that of f. So if we write $\mathbf{h}(\mathbf{x})$ for the inverse of $\nabla f^*(\mathbf{y})$, we have

$$f^{**}(\mathbf{x}) = \langle \mathbf{x}, \mathbf{h}(\mathbf{x})\rangle - f^*(\mathbf{h}(\mathbf{x})).$$

so plugging in the value of f^* from (20.3.3) we get

$$f^{**}(\mathbf{x}) = \langle \mathbf{x}, \mathbf{h}(\mathbf{x})\rangle - \langle \mathbf{y}, \mathbf{s}(\mathbf{y})\rangle + f(\mathbf{s}(\mathbf{y})).$$

Since $\mathbf{x} = \mathbf{s}(\mathbf{y})$ and $\mathbf{y} = \mathbf{h}(\mathbf{x})$, we get $f^{**}(\mathbf{x}) = f(\mathbf{x})$, so that applying the conjugate operation twice gives us the original function.

We want to generalize this result to the case where the function f is not differentiable. We also want to start with a In this more general situation, we motivated the definition of the conjugate in terms of the subgradient in (21.3.14), which we now rewrite as a lemma:

20.3.5 Lemma. *If* $\mathbf{a}$ *is a subgradient to the function* $f(\mathbf{x})$ *at the point* $\mathbf{x}_0$*, then*

$$f^*(\mathbf{a}) = \langle \mathbf{a}, \mathbf{x}_0 \rangle - f(\mathbf{x}_0).$$

For some $\mathbf{y}$, $f^*(\mathbf{y})$ may take the value $+\infty$. Let $D^* \in \mathbb{R}^n$ be the locus where $f^*(\mathbf{y})$ takes finite values, and call D^* the proper domain of f^*.

20.3.6 Theorem. D^* *is a convex set, and* $f^*(\mathbf{x})$ *is a convex function on* D^**.*

Proof. First we show that D^* is a convex set. Take any two points $\mathbf{y}_0$ and $\mathbf{y}_1$ in D^*: we need to show that for any λ, $0 < \lambda < 1$, the convex combination $\lambda\mathbf{y}_0 + (1-\lambda)\mathbf{y}_1$ is in D^*. Because $\mathbf{y}_0$ and $\mathbf{y}_1$ are in D^* there exist finite c_0 and c_1 such that, for all $\mathbf{x} \in D$,

$$\begin{aligned} c_0 &\geq \langle \mathbf{y}_0, \mathbf{x} \rangle - f(\mathbf{x}), \\ c_1 &\geq \langle \mathbf{y}_1, \mathbf{x} \rangle - f(\mathbf{x}). \end{aligned}$$

Multiply the first equation by λ and the second by $1-\lambda$, and add. You get

$$\lambda c_0 + (1-\lambda)c_1 \geq \langle \lambda\mathbf{y}_0 + (1-\lambda)\mathbf{y}_1, \mathbf{x} \rangle - f(\mathbf{x}),$$

so that $\lambda c_0 + (1-\lambda)c_1$ is a bound for $f^*(\lambda\mathbf{y}_0 + (1-\lambda)\mathbf{y}_1)$, which establishes the convexity of D^*.

Now we can establish the convexity of the function f^* on D^*. For a fixed $\mathbf{x}$, $\langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x})$ is an affine, and therefore convex, function of $\mathbf{y}$. So $f^*(\mathbf{y})$ is convex, since it is the pointwise least upper bound of a family of convex functions: Example 22.3.11. □

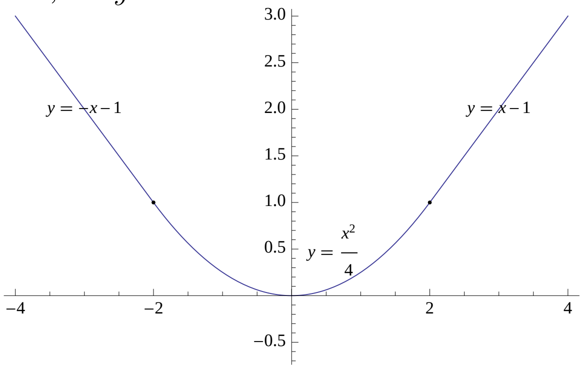
20.3.7 Example. Here is a concrete example of a conjugate function in one variable. Let $f(x) = x^2$, restricted to $-1 \leq x \leq 1$. Then

$$f^*(y) = \max_{-1 \leq x \leq 1} (yx - x^2).$$

So for each fixed y, we want to maximize the function $g(x) = -x^2 + xy$, $-1 \leq x \leq 1$. Clearly the quadratic $-x^2 + xy$ has unique maximizer at $x = y/2$. So, when $|y| \leq 2$, the maximizer is in the interval $-1 \leq x \leq 1$, so the maximum value is $g(y/2) = y^2/4$. When $y > 2$, the maximizer is at the end point is $x = 1$, so the maximum value is $g(1) = y - 1$. When $y < -2$, the maximizer occurs at $y = -1$, so the maximum value is $g(1) = -y - 1$. So

$$\mathbf{f}^*(y) = \begin{cases} -y-1, & \text{if } y \leq -2; \\ y^2/4, & \text{if } -2 \leq y \leq 2; \\ y-1, & \text{if } y \geq 2. \end{cases}$$

Here is a graph of the function, showing it is convex. It is even differentiable at $y = -2$, and $y = 2$.



20.3.8 Example. Now for the same example in one extra dimension. Let $f(x_1, x_2) = x_1^2 + x_2^2$, restricted to the closed unit disk D. Then

$$f^*(\mathbf{y}) = \max_{\mathbf{x} \in D} \left(\langle \mathbf{x}, \mathbf{y} \rangle - x_1^2 - x_2^2 \right).$$

We compute this function in Example 31.4.5, using the techniques of constrained optimization, getting:

$$f^*(\mathbf{y}) = \begin{cases} \frac{y_1^2+y_2^2}{4}, & \text{if } y_1^2 + y_2^2 \leq 4; \\ \sqrt{y_1^2 + y_2^2} - 1, & \text{if } y_1^2 + y_2^2 > 4. \end{cases}$$

Each part of the function is convex, and they agree where they meet: the circle of radius 2, confirming what Theorem 20.3.6 tells us: $f^*(\mathbf{y})$ is convex.

20.3.9 Example. Pick an arbitrary set D, and let $\psi(\mathbf{x})$ be the *indicator function* of D: the function that takes the value 0 on D, and the value ∞ everywhere else. Note that $\psi(\mathbf{x})$ is a convex function if and only if D is a convex set. Then by definition its conjugate function $\psi^*(\mathbf{y}) = \sup_{\mathbf{x} \in D} \langle \mathbf{x}, \mathbf{y} \rangle$. Thus ψ^* is the support function of D, that we defined and studied in §20.1. Its importance stems the fact that D is contained in the half-space $H^-_{\mathbf{y},c}$ if and only if $\psi^*(\mathbf{y}) \leq c$.

For a concrete example, let D by the interval $[-2, 1]$. Then since $\psi^*(y) = \sup_{x \in D} yx$, we see that if $y \geq 0$, we have $\psi^*(y) = y$, while if $y \leq 0$, $\psi^*(y) = -2y$. Thus $\psi^*(y)$ is a convex function.

Analogously, working backwards, we see that the function $|y|$ is the conjugate of the indicator function of the interval $[-1, 1]$, so $|y|$ is its support function.

20.3.10 Exercise. Show that the distance function $\|\mathbf{y}\|$ in $\mathbb{R}^n$ is the conjugate of the indicator function of the closed unit ball in $\mathbb{R}^n$.

20.3.11 Exercise. Show that $f^*(\mathbf{0}) = -\inf_{\mathbf{x} \in D} f(\mathbf{x})$.

Also show that if f and g have the same domain D, and $f(\mathbf{x}) \leq g(\mathbf{x})$ for all $\mathbf{x} \in D$, then $g^*(\mathbf{y}) \leq f^*(\mathbf{y})$.

Recalling Definition 21.3.12, we have:

20.3.12 Theorem. *If the function f has a subgradient $\mathbf{y}_0$ at $\mathbf{x}_0$, then f^* has subgradient $\mathbf{x}_0$ at $\mathbf{y}_0$.*

Proof. Lemma 20.3.5 establishes the formula

$$f(\mathbf{x}_0) + f^*(\mathbf{y}_0) = \langle \mathbf{y}_0, \mathbf{x}_0 \rangle,$$

when $\mathbf{y}_0$ is a subgradient of f at $\mathbf{x}_0$. Rearranging, we have

$$f(\mathbf{x}_0) = \langle \mathbf{y}_0, \mathbf{x}_0 \rangle - f^*(\mathbf{y}_0).$$

□

20.3.13 Theorem. *If the function f has a subgradient $\mathbf{y}$ at $\mathbf{x}$, and if the subgradient hyperplane corresponding to this subgradient only meets the graph of f at the point $(\mathbf{x}, f(\mathbf{x})$, then f^* is differentiable at $\mathbf{y}$, and $\nabla f^*(\mathbf{y}) = \mathbf{x}$.*

Proof. The proof has two parts: first an analysis of what happens near the unique point of contact of the subgradient hyperplane with the graph, and second a global analysis. The function is strictly convex near the point of interest. □

Next we take the conjugate $(f^*)^*$ of the conjugate f^*, which we of course write f^{**}. So

$$f^{**}(\mathbf{x}) = \sup_{\mathbf{y}} \{\langle \mathbf{x}, \mathbf{y} \rangle - f^*(\mathbf{y})\}. \tag{20.3.14}$$

Because $f^{**}(\mathbf{x})$ is a conjugate, it is a convex function. Entering the definition of $f^*(\mathbf{y})$ in (20.3.14), we get

$$f^{**}(\mathbf{x}) = \sup_{\mathbf{y}} \{\langle \mathbf{x}, \mathbf{y} \rangle - \sup_{\zeta \in D} \{\langle \mathbf{y}, \zeta \rangle - f(\zeta)\}\}.$$

When $\mathbf{x} \in D$, this implies that $f^{**}(\mathbf{x}) \leq f(\mathbf{x})$. Indeed, we have the important theorem

20.3.15 Theorem. *The biconjugate $f^{**}(\mathbf{x})$ is the largest convex and closed function less than or equal to $f(\mathbf{x})$ at each point $\mathbf{x} \in D$.*

20.3.16 Theorem. *$f^{**}(\mathbf{x}) = f(\mathbf{x})$ if and only if f has a supporting hyperplane at $\mathbf{x}$.*

20.3.17 Theorem. *If f^* is differentiable at $\mathbf{y}$, then letting $\mathbf{x} = \nabla f^*(\mathbf{y})$, then $f(\mathbf{x}) = f^{**}(\mathbf{x})$.*

20.3.18 Example. The biconjugate of the indicator function ψ of an arbitrary set D is the indicator function of the closure of the convex hull of D.

Indeed, the domain of definition of the biconjugate must be at least the convex hull of D, since it must be defined on a convex set containing D. The indicator function of the convex hull of D is the largest convex function less than or equal of f. Finally since ψ^{**} is closed, it must extend to 0 at every point in the closure of the convex hull.

20.4 Polar Cones

We considered the polar set of a convex set in §20.2. As we will now see, the polar set of a cone has a special form: in particular it is a cone. For that reason it is sometimes called the dual cone, rather than the polar cone, but we will just call it the polar cone. This section can be skipped on first reading.

20.4.1 Proposition. *Let C be a cone in $\mathbb{R}^m$, and let C^* be the polar set of C. Then*

$$C^* = \{\mathbf{y} \in \mathbb{R}^m \mid \langle \mathbf{y}, \mathbf{x} \rangle \leq 0 \textit{ for all } \mathbf{x} \in C\}. \tag{20.4.2}$$

In particular C^* is a cone.

Proof. Start from the defining equation 20.2.2 of the polar set: the set of $\mathbf{y}$ such that $\langle \mathbf{y}, \mathbf{x} \rangle \leq 1$ for all $\mathbf{x} \in C$. We need to show that any $\mathbf{y}$ such that $\langle \mathbf{y}, \mathbf{x}_0 \rangle > 0$ for an $\mathbf{x}_0 \in C$, is not in C^*. Assume by contradiction that $\langle \mathbf{y}, \mathbf{x}_0 \rangle = a > 0$. Then test $\mathbf{y}$ on the point $\lambda \mathbf{x}_0$, for $\lambda > 0$, which is in C because C is a cone. Then $\langle \mathbf{y}, \lambda \mathbf{x}_0 \rangle = \lambda a$. So by taking $\lambda > 1/a$, we see that $\mathbf{y}$ is not in the polar set of C. □

20.4.3 Corollary. *For any cone, $C \subset C^{**}$.*

Proof. Since C^* is a cone, we have

$$C^{**} = \{\mathbf{x} \in \mathbb{R}^m \mid \langle \mathbf{y}, \mathbf{x} \rangle \leq 0 \text{ for all } \mathbf{y} \in C^*\}, \tag{20.4.4}$$

so that by (20.4.2) all the $\mathbf{x}$ in C belong to C^{**}. □

20.4.5 Example. The polar of the ray $r_{\mathbf{a}}$ defined in 19.2.4 is the closed half-space

$$H^{-}_{\mathbf{a},0} = \{\mathbf{x} \mid \langle \mathbf{a}, \mathbf{x} \rangle \leq 0\}$$

defined in Example 18.1.7. As per Theorem 18.1.7, it is closed and convex and contains the origin.

20.4.6 Example. The polar of a linear space is the linear space orthogonal to it.

Proof. First consider the line L given parametrically as $t\mathbf{a}$, where $\mathbf{a}$ is a non-zero vector and $t \in \mathbb{R}$. We just saw that the polar of the ray $r_{\mathbf{a}}$ is the half-space $H^{-}_{\mathbf{a},0}$; the polar of the opposite ray $r_{-\mathbf{a}}$ is $H^{+}_{\mathbf{a},0}$. Thus the polar of L is the intersection of the two, namely the hyperplane $H_{\mathbf{a},0}$.

Now suppose L is a linear space of dimension k. Let $\mathbf{a}_1, \dots, \mathbf{a}_k$ be a basis for L. As noted in Example 19.3.6, L is a finite cone with minimal set of generators $\pm\mathbf{a}_1, \dots, \pm\mathbf{a}_k$. Thus the polar of L is the set of points $\mathbf{y}$ satisfying $\langle \mathbf{a}_i, \mathbf{y} \rangle = 0$, for $1 \leq i \leq k$, which is the orthogonal complement.

Finally, if $k = m$, so L is the entire $\mathbb{R}^m$, then the polar is the trivial cone reduced to the origin: no rays at all. □

More generally, we see:

20.4.7 Example. The polar of the finite cone C_A generated by vectors $\mathbf{a}_1, \dots, \mathbf{a}_n$ is the polyhedral cone (see Definition 18.3.16) $P(A^T, \mathbf{0})$.

Since the finite cone consists of all elements $\sum_{j=1}^{n} \lambda_j \mathbf{a}_j$, for all $\lambda_j \geq 0$, we see that an element $\mathbf{x}$ of the polar must satisfy $\langle \mathbf{a}_j, \mathbf{x} \rangle \leq 0$ for all j. Conversely every element $\mathbf{x}$ satisfying these inequalities is in the polar. To conclude, just note that this defines $P(A^T, \mathbf{0})$.

20.4.8 Theorem. *Let C_1 and C_2 be convex cones in $\mathbb{R}^m$. We can build new convex cones in the following three ways:*

1. *The intersection $C_1 \cap C_2$ is a convex cone.*
2. *The Minkowski sum (see Definition 18.1.31) $C_1 + C_2$ is a convex cone.*
3. *The polar C_1^* of C_1 is a closed convex cone.*

Proof. The intersection of cones is a cone, so just use Theorem 18.1.15 to get the first statement.

We already proved the convexity of the Minkowski sum $C_1 + C_2$ in Example 19.3.8.

For the last statement, using Theorem 20.2.6 again, all we need is to show that the polar C_1^* is a cone, and that is Proposition 20.4.1. □

20.4.9 Proposition. *Let C_1 and C_2 be convex cones in $\mathbb{R}^m$. If $C_1 \subset C_2$, then $C_2^* \subset C_1^*$ and $(C_1^*)^* \subset (C_2^*)^*$.*

Proof. This is a special case of Theorem 20.2.7. □

20.4.10 Theorem. *If C is a closed convex cone, then $C = C^{**}$.*

Proof. This is a special case of the Bipolar Theorem 20.2.13, since cones always contain the origin. □

20.4.11 Exercise. In Corollary 20.4.3 we noticed that $C \subset (C^*)^*$. Use the Farkas alternative to show that for a finite cone C, $C = (C^*)^*$. Hint: Take an element $\mathbf{b}$ in $(C^*)^*$ and show it is in C. This proves Theorem 20.4.10 in the special case of finite cones.

Finally just as in the case of bounded polyhedra and polytopes covered by Corollary 20.2.14, we have:

20.4.12 Theorem (Weyl's Theorem)**.** *Every finite cone is a polyhedral cone.*

This is due to Hermann Weyl [73]. We follow his proof.

20.4.13 Corollary. *The polar of a polyhedral cone is a finite cone. Every polyhedral cone is a finite cone.*

Proof. Write the polyhedral cone as $P(A^T, \mathbf{0})$, so it is the polar of the finite cone C_A by Example 20.4.7. By Theorem 20.4.10 its polar, which is the bipolar of C_A, is also C_A. That establishes the first point.

The last point follows by an argument similar to the proof of Corollary 20.2.14. Next start with a polyhedral cone $P(A^T, \mathbf{0})$, the polar of C_A. By Weyl's theorem the finite cone C_A is a polyhedral cone we can write $P(D^T, \mathbf{0})$, for some $m \times k$ matrix D. The polar of $P(D^T, \mathbf{0})$ is the finite cone C_D, so $P(A^T, \mathbf{0})$ is the same set as C_D, and we are done. □

20.5 From Finite Cones to Polytopes

In Definition 19.3.5 we defined the dimension of a finite cone as its dimension as a convex set: see Definition 18.2.24). By restricting to the affine hull of the cone, we can assume that we are working with a cone of maximum dimension.

One difficulty is working with cones is the fact that there are never compact (because not bounded), and the best results we have found for convex sets concern compact ones. To remedy this problem we define the base of a cone. The best cones will have a base that is compact. That allows us to get interesting results.

This section can be skipped on first reading.

20.5.1 Definition. Let C be a cone of dimension m in $\mathbb{R}^m$, and let V be an affine hyperplane not containing the origin. Then V is called a *base* for C if V intersects every ray in C.[1]

Since a ray of C contains the origin which is not in V, it follows from linearity that a ray and V can meet in at most one point. This simple remark shows:

20.5.2 Proposition. *If the cone C contains a line L, then it does not have a base.*

Proof. Indeed, the line meets V in at most one point, and yet it corresponds to two rays (19.2.4) of C. □

20.5.3 Proposition. *Let C be a convex cone of dimension m in $\mathbb{R}^m$. Assume V is a base for C, and let $K = V \cap C$. Then K is a convex set of dimension $m - 1$. If the base K is compact, then C is closed*

Proof. We have a map ϕ from $C \smallsetminus \mathbf{0}$ to K which associates to any point $\mathbf{c} \neq \mathbf{0}$ in C the intersection of the ray $r_{\mathbf{c}}$ with V. This map is onto by definition. The map is not defined at the origin, but otherwise each point of a ray of C maps to the same point of K.

By choosing appropriate coordinates, we may assume that V is the affine hyperplane $x_m = 1$, in which case, assigning the coordinates $(x_1, \dots, x_{m-1})$ to the point $(x_1, \dots, x_{m-1}, 1)$ of V, the map ϕ is written

$$\phi(x_1, \dots, x_m) = (\frac{x_1}{x_m}, \dots, \frac{x_{m-1}}{x_m}),$$

so ϕ is defined everywhere on $\mathbb{R}^m$ except the hyperplane $x_m = 0$.

Note that the points of $C \smallsetminus \mathbf{0}$ have $x_m > 0$, since all the rays of C meet V. If $\mathbf{k} = (k_1, \dots, k_{m-1})$ is in K, then the corresponding ray in C is $(tk_1, \dots, tk_{m-1}, t)$. This explicit formula shows that ϕ is continuous (indeed differentiable) on $C \smallsetminus \mathbf{0}$.

It is now easy to prove that K is convex. Pick two points $\mathbf{k}^0$ and $\mathbf{k}^1$ in K. We must show that $\mathbf{k}^\lambda = \lambda \mathbf{k}^1 + (1 - \lambda)\mathbf{k}^0$ is in K, for $\lambda \in (0, 1)$. The point $\mathbf{c}^1 = (k^1_1, \dots, k^1_{m-1}, 1)$ in C is above $\mathbf{k}^1$, and the point $\mathbf{c}^0 = (k^0_1, \dots, k^0_{m-1}, 1)$ in C is above $\mathbf{k}^0$. By convexity $\lambda \mathbf{c}^1 + (1 - \lambda)\mathbf{c}^0$ is in C, and its image under ϕ is $\mathbf{k}^\lambda$, which is therefore in K.

Now assume K is compact. The map ϕ sending a non-zero $\mathbf{c}$ in C to K is continuous, as we noted above . Now suppose $\mathbf{c}$ is a point in the closure of C, so that there is a sequence of points $\mathbf{c}^i$ in C converging to $\mathbf{c}$. If we can show $\mathbf{c}$ is in C, we are done. We can assume $\mathbf{c}$ is not the origin (which is in C), so we cam assume that the same holds for every term in the sequence $\{\mathbf{c}^i\}$. Thus we get a convergent

[1] This has nothing to do with basic subcones.

sequence $\phi(\mathbf{c}^i)$ in K, and since K is closed, it converges to a point in K. But that means that the ray from this point is in C, showing that C is closed. □

20.5.4 Theorem. *Let C be a closed convex cone of dimension m in $\mathbb{R}^m$. Assume V is a base for C, and let $K = V \cap C$. Then r is an extreme ray*[2] *of C, if and only if $\mathbf{p} = r \cap V$ is an extreme point*[3] *of K.*

Proof. The equivalence between the extreme rays of C and the extreme points of K follows by comparing Theorem 18.7.3 and Theorem 19.2.9, noting that the dimension of K is one less than the dimension of C. □

20.5.5 Theorem (Minkowski's Theorem for Cones)**.** *Assume C is a convex cone with a compact base K. Then every non-zero point of C can be written as a conical combination of points in the extreme rays of C. Namely, if $\mathbf{c} \in C$,*

$$\mathbf{c} = \sum_{j=1}^{n} \lambda_j \mathbf{a}_j, 0 < \lambda_j$$

where the rays $r_{\mathbf{a}_i}$ are extreme for C.

Proof. For a point $\mathbf{c} \in C$, use Minkowski's Theorem 18.7.1 to write $\phi(\mathbf{c})$ as a convex combination of the extreme points of K. By the previous theorem the rays above the extreme points of K are extreme rays of C. Choosing the same coordinate system, pick the point of the extreme ray with last coordinate equal to one, and take the same convex combination to get $\mathbf{c}$ as a convex combination of points on extreme rays. □

Finally we ask: which cones have a base? The most important result for us is

20.5.6 Theorem. *If C is a closed cone that does not contain any lines, then C has a compact base.*

[2] see Definition 19.2.5

[3] see Definition 18.1.10

Lecture 21

Convex Functions

The basic properties of convex functions are established. First we study convex functions in one variable because many of the properties of convex functions are established by reducing to the single variable case. The connection to convex sets is made through the domain of the convex function, which is always assumed to be convex, and the epigraph defined in §21.3. The main results of this lecture concern the continuity of convex functions: first and foremost, convex functions are continuous on the interior of their domain: Theorem 21.4.3.

The final topic of the lecture is not needed immediately, and can be skipped until needed later in the lectures. Precisely in order to understand how convex functions behave on the boundary of their domain, in Lecture 16 we generalized the notion of continuity to that of lower semicontinuity: §16.3. In §16.4 we proved an extension of the Weierstrass theorem to such functions, which shows that for the purposes of minimization lower semicontinuous functions are as useful as continuous ones. Using these concepts, in §21.5 we show how convex functions that are not continuous on the boundary of their domain can be made lower semicontinuous without modifying them on the interior.

21.1 The Definitions

21.1.1 Definition. A function $f : S \subset \mathbb{R}^n \to \mathbb{R}$ is *convex* if

- The domain S of f is convex;
- For any two points $\mathbf{x}_1$ and $\mathbf{x}_2$ in S, and for all λ, $0 \le \lambda \le 1$ we have:

$$f(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \le \lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2) \tag{21.1.2}$$

In other words, the graph of the function f on the segment $[\mathbf{x}_1, \mathbf{x}_2]$ is below the secant line from $(\mathbf{x}_1, f(\mathbf{x}_1))$ to $(\mathbf{x}_2, f(\mathbf{x}_2))$.

See Example 21.2.6 for the graph of a convex function in one variable, together with three of its secant lines illustrating the convexity of the function.

As usual we can rewrite (21.1.2) as

$$f(\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2) \le \lambda_1 f(\mathbf{x}_1) + \lambda_2 f(\mathbf{x}_2), \tag{21.1.3}$$

where

$$\lambda_1 \ge 0\,,\, \lambda_2 \ge 0\,, \text{ and } \lambda_1 + \lambda_2 = 1. \tag{21.1.4}$$

Just set $\lambda_1 = \lambda$ and $\lambda_2 = 1 - \lambda$.

21.1.5 Remark. If ℓ is a linear function, then for all a_1 and a_2 we have

$$\ell(a_1\mathbf{x}_1 + a_2\mathbf{x}_2) = a_1\ell(\mathbf{x}_1) + a_2\ell(\mathbf{x}_2), \tag{21.1.6}$$

so convex functions generalize linear functions in two ways:

- they are *sublinear*, meaning that the equality in (21.1.6) is replaced by an inequality;
- only certain coefficients are allowed: they must satisfy (21.1.4).

21.1.7 Definition. f is *strictly convex* if f is convex and the inequality in (21.1.2) is strict for all $\mathbf{x}_1 \ne \mathbf{x}_2$, and all λ, $0 < \lambda < 1$.

Example 21.2.6 is strictly convex. A line, such as $y = mx + b$, is convex but not strictly convex.

21.1.8 Exercise. Let f be a convex function from an open interval $\mathcal{I}$ in $\mathbb{R}$. Assume that for the two distinct points x_1 and x_2 in $\mathcal{I}$, there exists a λ, $0 < \lambda < 1$, such that we get an equality in (21.1.2), so that f is not strictly convex.

1. Show that this implies that we get equality for all λ, $0 < \lambda < 1$.
2. Consider the function $g(\lambda) = f(\lambda x_1 + (1-\lambda)x_2)$, for $0 \le \lambda \le 1$. Show that $g(\lambda)$ is equal to $f(x_2) + \lambda(f(x_1) - f(x_2))$, so that it is an affine function of λ. Thus the derivative $g'(\lambda)$ is equal to $f(x_1) - f(x_2)$, a constant. So convex functions that are not strictly convex are locally affine.
3. Here is a geometric way of saying this: suppose you have three distinct points x_1, x_2 and x_3 in $\mathcal{I}$, with x_3 between x_1 and x_2, so that $x_3 = \lambda x_1 + (1-\lambda)x_1$, for some λ between 0 and 1. Assume that the three points $(x_1, f(x_1))$, $(x_2, f(x_2))$, $(x_3, f(x_3))$ in $\mathbb{R}^2$ are aligned. Then the graph of the convex function f is a line segment over the segment $[x_1, x_2]$ in $\mathbb{R}$. Convince yourself this is the case.

21.1.9 Definition. f is *concave* if $-f$ is convex.

These lectures focus on convex functions, and leave the statements for concave functions to the reader.

21.1.10 Example. A function $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$, where $\mathbf{a}$ and $\mathbf{x}$ are in $\mathbb{R}^n$, and b is a real number, is both convex and concave, but neither strictly convex or strictly concave. Compare to Exercise 21.1.8.

We now state and prove the analog of the Convex Combination Theorem 18.1.24 for convex sets.

21.1.11 Theorem (Jensen's Inequality). *A function f defined on a convex set S is convex if and only if for any set of points $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$ in S, and any convex combination*

$$\mathbf{x} = \lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \cdots + \lambda_n\mathbf{x}_n$$

we have

$$f(\mathbf{x}) \leq \lambda_1 f(\mathbf{x}_1) + \lambda_2 f(\mathbf{x}_2) + \cdots + \lambda_n f(\mathbf{x}_n) \tag{21.1.12}$$

Proof. First note that by the Convex Combination Theorem, $\mathbf{x}$ is in S, so $f(\mathbf{x})$ makes sense.

That (21.1.12) implies convexity is trivial, and the proof that convexity implies Jensen's inequality follows that of the Convex Combinations Theorem: induction on the number r of points. We start at $r = 2$: the result is then the definition of a convex function.

Next we assume that the result is known for r, so that (21.1.12) is satisfied when $n = r$, and we prove it for $r + 1$.

We may assume that $\lambda_i > 0$ for all i, since otherwise there is nothing to prove. Let $\Gamma = \sum_{i=1}^r \lambda_i$, so $0 < \Gamma < 1$. Let $\gamma_i = \lambda_i/\Gamma$, $1 \leq i \leq r$, so that the point $\mathbf{y} = \sum_{i=1}^r \gamma_i\mathbf{x}_i$ is a convex combination of r points of S, so by the induction hypothesis,

$$f(\mathbf{y}) \leq \gamma_1 f(\mathbf{x}_1) + \gamma_2 f(\mathbf{x}_2) + \cdots + \gamma_r f(\mathbf{x}_r). \tag{21.1.13}$$

Then $\mathbf{x} = \Gamma\mathbf{y} + \lambda_{r+1}\mathbf{x}_{r+1}$, and $\Gamma + \lambda_{r+1} = 1$ so $\mathbf{x}$ is a convex combination of two points of S, and is therefore in S since S is convex. By the definition of a convex function,

$$f(\mathbf{x}) \leq \Gamma f(\mathbf{y}) + \lambda_{r+1} f(\mathbf{x}_{r+1}).$$

Now just substitute in $f(\mathbf{y})$ from (21.1.13), and the definition of Γ. □

21.1.14 Definition. The *sublevel set* S_c of a real-valued function $f(\mathbf{x})$ on S is defined by

$$S_c = \{\mathbf{x} \in S \mid f(\mathbf{x}) \leq c\}.$$

Thus, if $c < d$, $S_c \subset S_d$, and if $f(\mathbf{x}) = c$, then $\mathbf{x} \notin S_b$ for $b < c$. Note that we can take $c \in \overline{\mathbb{R}}$. Then $S_\infty = S$, and $S_{-\infty}$ is empty.

We have the following easy but important theorem:

21.1.15 Theorem. *Assume f is a convex function. For each $c \in \mathbb{R}$, S_c is a convex set. Furthermore, if we let S^0_c be the set $\{\mathbf{x} \in S \mid f(\mathbf{x}) < c\}$, it too is convex.*

Proof. Suppose $\mathbf{x}_1$ and $\mathbf{x}_2$ are in S_c, meaning that $f(\mathbf{x}_1) \leq c$ and $f(\mathbf{x}_2) \leq c$. We must show that for all λ, $0 < \lambda < 1$, $f(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq c$. By (21.1.2), we have

$$f(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq \lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2) \leq \lambda c + (1-\lambda)c = c$$

so we are done. The second assertion is proved in the same way. □

21.1.16 Example. A simple function of two variables that is convex on all of $\mathbb{R}^2$ is the paraboloid of revolution $f(x, y) = x^2 + y^2$. Then the sublevel set $_c$ is the disk of radius $\sqrt{c}$ when c is not negative, and empty when c is negative. Thus it is a convex set.

Quasiconvex functions, that we study in Lecture 24, are characterized by the property that their sublevel sets are convex. See Theorem 24.1.12.

21.2 Convex Functions in One Variable

In this section we study convex functions in one variable. Much of what is done in this section is redone later in the case of several variables, leading to some repetition in the exposition. However the functions are simpler and more transparent in this case, so beginners, especially, are urged to read this section. Furthermore it can happen, as in the result below, that to establish a property of a convex function with domain S, it is enough to establish it for the restriction of the function to $L \cap S$, where L is any line meeting S. For example:

21.2.1 Theorem. *A function f from the convex set $S \in \mathbb{R}^n$ is convex if and only if for every line L meeting S, the restriction f_L of f to $L \cap S$ is convex.*

Proof. This is obvious, because

1. The convexity of the set S is checked on lines, since Definition 18.1.3 only involves lines.
2. The inequality (21.1.2) is also checked on lines.

□

Note that the domain of a convex function restricted to a line is an interval, perhaps unbounded. We will usually assume the interval is open, because the condition imposed by convexity at an end point is weaker than at an interior point. Thus end points usually have to be handled separately. Our first result works for any interval.

21.2.2 Theorem (The Three Secant Theorem)**.** *Let f be a convex function from an interval S in $\mathbb{R}$ to $\mathbb{R}$. Pick any three points x, y and z in S, $x < y < z$. Then*

$$\frac{f(y)-f(x)}{y-x} \le \frac{f(z)-f(x)}{z-x} \le \frac{f(z)-f(y)}{z-y}$$

Each term is the slope of the secant between the corresponding points on the graph of f, so writing $\operatorname{slope}(a,b) = \frac{f(b)-f(a)}{b-a}$, we have:

$$\operatorname{slope}(x,y) \le \operatorname{slope}(x,z) \le \operatorname{slope}(y,z), \quad \text{when} \;\; x < y < z.$$

Proof. Express the point in the middle, y, as a convex combination of x and z.

$$y = (1-\lambda)x + \lambda z \tag{21.2.3}$$

Since the three points are distinct, $0 < \lambda < 1$.

Since f is convex, $f(y) \le (1-\lambda)f(x) + \lambda f(z)$, so

$$f(y) - f(x) \le \lambda(f(z) - f(x)) \tag{21.2.4}$$

Divide (21.2.4) by the positive quantity $y - x$, which is equal to $\lambda(z - x)$ by (21.2.3). This gives the left-hand inequality in the theorem:

$$\frac{f(y)-f(x)}{y-x} \le \frac{f(z)-f(x)}{z-x}.$$

We proceed similarly for the right-hand inequality, interchanging the roles of x and y, and writing $\mu = 1 - \lambda$ so

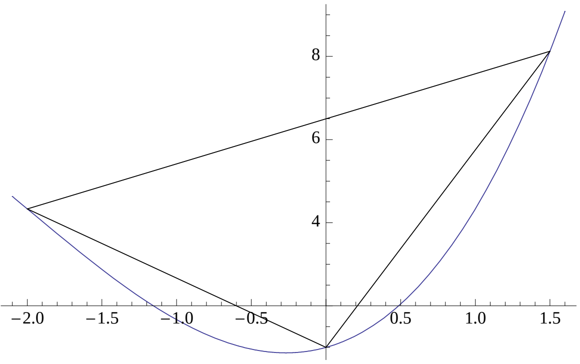
$$y = \mu x + (1-\mu)z. \tag{21.2.5}$$

The convexity of f yields $f(y) \le \mu f(x) + (1-\mu)f(z)$, so (note the change of sign)

$$f(z) - f(y) \ge \mu(f(z) - f(x)).$$

Divide by $z - y$, which is equal to $\mu(z - x)$ by (21.2.5), to get the right-hand inequality. □

21.2.6 Example. Let $f(x) = x^3/3 + 2x^2 + x + 1$. The second derivative of f is $2x + 4$, and this is non-negative for $x \geq -2$, meaning that f is convex on that interval: see Theorem 21.2.20. Take $x = -2$, $y = 0$ and $z = 1.5$. We draw the graph of the function with its three secants.



The slopes are clearly increasing as in the theorem.

If we knew that f is continuous on the interval $[a, b]$, the following theorem would follow from the Weierstrass theorem. But we do not yet know this. We will use this result to prove that f is continuous.

21.2.7 Theorem. *Assume the function $f(x)$ is convex on the interval $[a, b]$. Then*

- *f is bounded above on $[a, b]$ by M, the bigger of the two values $f(a)$ and $f(b)$,*
- *f is bounded below by*

$$2f\big(\frac{a+b}{2}\big) - M.$$

Proof. First we establish the upper bound. Let x be any point in the open interval (a, b). Then x can be written as a convex combination of a and b: $x = \lambda a + (1-\lambda)b$, where $\lambda = (b - x)/(b - a)$ and so is obviously between 0 and 1. Then since f is convex,

$$f(x) \leq \lambda f(a) + (1 - \lambda)f(b) \leq\leq \lambda M + (1 - \lambda)M \leq M,$$

where M is the greater of the two numbers $f(a)$ and $f(b)$.

For the lower bound, let c be the midpoint $(a + b)/2$ of the segment. Then any point on the segment can be written $c + x$, for $|x| < (b - a)/2$. Then by convexity,

$$f(c) \leq \frac{f(c-x)}{2} + \frac{f(c+x)}{2}.$$

We rewrite this as

$$\frac{f(c+x)}{2} \geq f(c) - \frac{f(c-x)}{2} \geq f(c) - \frac{M}{2}$$

using the upper bound M. Then a lower bound is

$$m = 2f(c) - M.$$

□

21.2.8 Theorem. *Let $f(x)$ be a convex function in one variable defined on an open interval S. Then f is continuous at all points of S.*

Proof. Pick any point x_0 in S and let u be a small positive number such that $x_0 - u$ and $x_0 + u$ are in S. Let M be an upper bound for f on the interval $[x_0 - u, x_0 + u]$. To establish continuity of f at x_0, for every $\epsilon > 0$ we must find a $\delta > 0$ so that if $|x - x_0| < \delta$, then $|f(x) - f(x_0)| < \epsilon$. Pick a number h, $0 < h < 1$.

Apply the Three Secant Theorem 21.2.2 to the triple $x_0 < x_0 + hu < x_0 + u$, to get the inequality

$$\frac{f(x_0 + hu) - f(x_0)}{hu} \le \frac{f(x_0 + u) - f(x_0)}{u},$$

so, multiplying by the positive number u,

$$\frac{f(x_0 + hu) - f(x_0)}{h} \le f(x_0 + u) - f(x_0) \le M - f(x_0). \tag{21.2.9}$$

Next use the Three Secant Theorem on the triple $x_0 - u < x_0 < x_0 + hu$ to get the inequality

$$\frac{f(x_0) - f(x_0 - u)}{u} \le \frac{f(x_0 + hu) - f(x_0)}{hu},$$

so, multiplying by the negative number $-u$,

$$\frac{f(x_0) - f(x_0 + hu)}{h} \le f(x_0 - u) - f(x_0) \le M - f(x_0). \tag{21.2.10}$$

Equations 21.2.9 and 21.2.10, show that for all $\epsilon > 0$, it suffices to take $h < \frac{\epsilon}{|M - f(x_0)|}$ to get sufficiently close. Thus we can find δ, and we are done.

□

21.2.11 Exercise. Let $[a, b]$ be a closed interval inside the open domain of the convex function f. Show that there is a constant L such that for all distinct x and y in $[a, b]$,

$$\frac{|f(y) - f(x)|}{|y - x|} \le L.$$

We say $f(x)$ is *Lipschitz continuous* with Lipschitz constant L. Hint: This follows easily from Theorem 21.2.2: use for z a point in the domain of f greater than b.

21.2.12 Exercise. Prove that the function

$$f(x) = \begin{cases} 2, & \text{if } x = -1; \\ x^2, & \text{if } -1 < x < 1; \\ 2, & \text{if } x = 1. \end{cases}$$

on the closed interval $[-1, 1]$ is convex. Note that it is not continuous at the end points. Also see Example 21.3.10.

Next we study the differentiability of a convex function. For this we apply the Three Secant Theorem again.

21.2.13 Definition. For any function f, let the *right derivative* $f'_+(x)$ be

$$f'_+(x) = \lim_{y \searrow x} \frac{f(y) - f(x)}{y - x},$$

where y approaches x from above, assuming this limit exists. Let the *left derivative* $f'_-(x)$ be

$$f'_-(x) = \lim_{y \nearrow x} \frac{f(y) - f(x)}{y - x},$$

where y approaches x from below.

21.2.14 Theorem. *If $f(x)$ is convex on the open set S, then $f'_-(x)$ and $f'_+(x)$ exist for all $x \in S$, and are increasing functions. Furthermore $f'_-(x) \le f'_+(x)$, with equality if and only if the function is differentiable at x.*

Proof. To show that $f'_+(x)$ exists, we consider a decreasing sequence $\{y_k\}$ approaching x from above. It follows from Theorem 21.2.2 that the sequence $\frac{f(y_k)-f(x)}{y_k-x}$ decreases. Since this sequence is clearly bounded from below, by, say, the slope of a secant at smaller values, by Theorem 21.2.2, Theorem 10.2.4 on bounded monotone sequences tells us that the limit exists.

The analogous result for $f'_-(x)$ is proved in the same way.

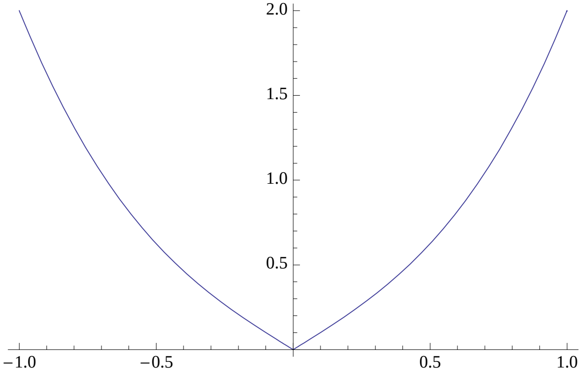
21.2.15 Exercise. Use Theorem 21.2.2 to finish the proof: show that $f'_+(x)$ is an increasing function and that $f'_-(x) \le f'_+(x)$.

□

21.2.16 Example. Let $f(x)$ be the function $|x^3 + x|$. Then f is convex on its entire domain (you should check this), and is differentiable everywhere except at $x = 0$. Indeed

$$f'(x) = \begin{cases} -3x^2 - 1, & \text{if } x < 0; \\ 3x^2 + 1, & \text{if } x > 0. \end{cases}$$

At 0 we have $f'_-(0) = -1$, and $f'_+(0) = 1$, and it is easy to verify all the conclusions of the theorem, as the following graph shows.



Here is another illustration of how convex functions are simpler than more general functions.

21.2.17 Theorem. *If the convex function f is differentiable on the open interval S, then its derivative f' is continuous, so f is $\mathcal{C}^1$.*

Proof. This is a simple consequence of the Intermediate Value Theorem 3.4.3. Indeed, Theorem 21.2.14 tells us that f' is an increasing function, so Corollary 3.4.4 shows that f' is continuous. □

21.2.18 Theorem. *If $f(x)$ is $\mathcal{C}^1$ on the open interval S, then f is convex if and only if $f'(x)$ is an increasing function on S. If $x < y < z$ are three points in S, we have*

$$\text{slope}(x, y) \le f'(y) \le \text{slope}(y, z).$$

Proof. First we give an elementary calculus proof that if $f(x)$ is $\mathcal{C}^1$ with increasing derivative on the open S, then $f(x)$ is convex on S. Pick points $x < z$ in S, and let $y = \lambda_1 x + \lambda_2 z$ be a point inbetween, where $\lambda_1 + \lambda_2 = 1$ and $\lambda_1 > 0$, $\lambda_2 > 0$ as usual. Rewriting (21.1.2), we must show

$$f(y) \le \lambda_1 f(x) + \lambda_2 f(z).$$

The left hand side is $(\lambda_1 + \lambda_2)f(y)$, so rewriting the inequality, we need:

$$\lambda_1\big(f(x) - f(y)\big) + \lambda_2\big(f(z) - f(y)\big) \ge 0. \tag{21.2.19}$$

By the Fundamental Theorem of Calculus, the left hand side is

$$-\lambda_1 \int_x^y f'(t)dt + \lambda_2 \int_y^z f'(t)dt.$$

Now we use the fact that $f'(t)$ is increasing. In each integrand we replace t by the appropriate end point value so that this expression decreases. Thus in the first integral we replace t by the upper end point y, and in the second one by the lower end point y, so that

$$\lambda_1\big(f(x) - f(y)\big) + \lambda_2\big(f(z) - f(y)\big) \geq -\lambda_1 \int_x^y f'(y)dt + \lambda_2 \int_y^z f'(y)dt.$$

Now we rearrange the right hand side, where we are integrating constants:

$$-\lambda_1 \int_x^y f'(y)dt + \lambda_2 \int_y^z f'(y)dt = -\lambda_1(y-x)f'(y) + \lambda_2(z-y)f'(y)$$

The right hand side becomes:

$$f'(y)\Big(- \lambda_1(\lambda_1 x + \lambda_2 z - x) + \lambda_2(z - \lambda_1 x - \lambda_2 z)\Big)$$

and then, grouping the terms in x and in z:

$$f'(y)\Big(\lambda_1 x(-\lambda_1 + 1 - \lambda_2) + \lambda_2 z(1 - \lambda_1 - \lambda_2)\Big) = 0.$$

So we have established (21.2.19). The other statements are corollaries of Theorem 21.2.14, since the left and right derivatives then agree. □

Finally, we assume that f is $\mathcal{C}^2$ (twice continuously differentiable). Then

21.2.20 Theorem. *If $f(x)$ is C^2 on the open interval S, then f is convex if and only if $f''(x) \geq 0$ for all $x \in S$. Furthermore if $f''(x) > 0$ for all $x \in S$, f is strictly convex on S.*

Proof. By Theorem 21.2.18 we know that f is convex if and only if $f'(x)$ is an increasing function. From single variable calculus, you know that a differentiable function is increasing if and only if its derivative is non-negative. We reviewed this in Theorems 3.1.14 and 3.2.4. □

21.2.21 Example. The function $f(x) = x^4$ is strictly convex on every interval, even though $f''(0) = 0$. Thus there is no converse to the last statement of Theorem 21.2.20.

Theorems 21.2.18 and 21.2.20 are the two standard results concerning convexity you learned in single variable calculus.[1]

[1] Calculus books talk about concave up and concave down, rather than convex and concave: see [63], §4.3.

21.2.22 Proposition. *Assume that the real-valued function f is invertible and convex (resp. concave) on the interval $S \subset \mathbb{R}$ to the interval T. Then the inverse function $f^{-1} : T \to S$ is concave (resp. convex).*

First Proof. Assume f is convex. Since f is invertible, it is either strictly increasing on the entire interval S, or strictly decreasing. We only deal with the case f is increasing. Pick two points $x_1 < x_2$ in S. Then $f(x_1) < f(x_2)$. Let $y_1 = f(x_1)$ and $y_2 = f(x_2)$. Thus $f^{-1}(y_1) < f^{-1}(y_2)$, so the inverse function is also increasing. By convexity

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda)f(x_2). \tag{21.2.23}$$

Applying the increasing function f^{-1} to (21.2.23) we get

$$\lambda x_1 + (1-\lambda)x_2 \le f^{-1}\big(\lambda f(x_1) + (1-\lambda)f(x_2)\big),$$

which can be rewritten

$$\lambda f^{-1}(y_1) + (1-\lambda)f^{-1}(y_2) \le f^{-1}\big(\lambda y_1 + (1-\lambda)y_2\big).$$

This says precisely that f^{-1} is concave. □

It is worth graphing f and its inverse function - the reflection in the 45 degree line - to see why the result is correct.

Second Proof. If f and its inverse are twice differentiable, the result also follows from a chain rule computation starting from the definition of the inverse function:

$$g(f(x)) = x.$$

Differentiate with respect to x using the chain rule to get

$$g'(f(x))f'(x) = 1.$$

so that $g'(f(x))$ is positive since $f'(x)$ must be, since f is convex. Differentiate again using the chain rule and the product rule to get:

$$g''(f(x))f'(x)^2 + g'(f(x))f''(x) = 0.$$

Since $g'(f(x))$ is positive, the sign of $g''(f(x))$ is the opposite of that of $f''(x)$, so we are done by Theorem 21.2.20. □

21.2.24 Exercise. In the first proof, write out the case where f is decreasing.

21.2.25 Exercise. Let $f(x)$ be convex on the interval $[a, b]$ and assume that $f'_+(a) > 0$. Show that $f(x)$ is strictly increasing, and therefore invertible on $[a, b]$. This is a sufficient condition for invertibility on a convex function. Write down a necessary condition and find a function that satisfies the necessary condition and yet is not invertible.

21.2.26 Exercise. Suppose that f is a convex function defined on all of $\mathbb{R}$. Assume that f is bounded from above, so that there is a constant M such that for all $x \in \mathbb{R}$, $f(x) \leq M$. Show that f is constant.

21.3 The Epigraph

We now begin our study of convex functions in several variables.

As you know, the *graph* of a real-valued function f with domain S in $\mathbb{R}^n$ is the set $\Gamma \subset \mathbb{R}^n \times \mathbb{R}$ consisting of the pairs $(\mathbf{x}, f(\mathbf{x}))$ for $\mathbf{x} \in S$. We now generalize this. We use y to denote the coordinate of the values of the function, which we call the vertical direction, so that the coordinates on $\mathbb{R}^{n+1}$ are $x_1, \dots, x_n, y$.

21.3.1 Definition. The set $M_f = \{(\mathbf{x}, y) \in \mathbb{R}^{n+1} \mid \mathbf{x} \in S, y \geq f(\mathbf{x})\}$ is called the *epigraph* of f, meaning the points above the graph.

The following theorem gives an alternate definition of convexity, used for example by Rockafellar [53].

21.3.2 Theorem. *The function $f(\mathbf{x})$ is a convex function if and only if its epigraph M_f is a convex set.*

Proof. We first prove the ⇒ implication, so we assume $f(x)$ is convex. We need to show that if the pairs $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$ are in M_f, then so is the pair

$$\lambda(\mathbf{x}_1, y_1) + (1-\lambda)(\mathbf{x}_2, y_2) \text{ for } 0 < \lambda < 1.$$

This pair can be written

$$(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \lambda y_1 + (1-\lambda)y_2),$$

so we need to prove that

$$\lambda y_1 + (1-\lambda)y_2 \geq f(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2).$$

Because $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$ are in the epigraph, $y_i \geq f(\mathbf{x}_i)$. Because f is convex,

$$f(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq \lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2) \leq \lambda y_1 + (1-\lambda)y_2.$$

This shows that $\lambda y_1 + (1-\lambda)y_2$ is in M_f, showing it is convex..

Now we prove the $\Leftarrow$ implication. We assume the epigraph is convex, so, since $(\mathbf{x}_1, f(\mathbf{x}_1))$ and $(\mathbf{x}_2, f(\mathbf{x}_2))$ are in the epigraph,

$$\lambda\big(\mathbf{x}_1, f(\mathbf{x}_1)\big) + (1-\lambda)\big(\mathbf{x}_2, f(\mathbf{x}_2)\big)$$

is too. Thus, taking the last coordinate,

$$\lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2) \geq f\big(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2\big),$$

which says that f is convex. □

21.3.3 Remark (The Boundary of M_f)**.** Since M_f is a convex set in $\mathbb{R}^{n+1}$, it has a supporting hyperplane at every point of its boundary, as we learned in Theorem 18.6.11 using the separation theorems.

What are the boundary points of M_f? Clearly any point $(\mathbf{x}^*, f(\mathbf{x}^*))$, where $\mathbf{x}^*)$ in a point in the domain of f. Furthermore, if $\mathbf{x}^*$ is a boundary point of S, then any point $(\mathbf{x}^*, y)$ with $y > f(\mathbf{x}^*)$ is an additional boundary point. This gives half lines on the boundary of M_f.

21.3.4 Remark. Now let $\mathbf{x}^*$ be a point in the interior of the domain S of f. Denote by $H_{\mathbf{x}^*}$ any supporting hyperplane at $(\mathbf{x}^*, f(\mathbf{x}^*))$. Then we can write the equation of the affine hyperplane $H_{\mathbf{x}^*}$ as

$$\langle \mathbf{a}, \mathbf{x} - \mathbf{x}^* \rangle + a_0(y - f(\mathbf{x}^*)) = 0, \tag{21.3.5}$$

for suitable constants a_0 and $\mathbf{a}$.

A hyperplane written as in (21.3.5) is *vertical* if $a_0 = 0$, where a_0 is the coefficient of the coordinate corresponding to the value of the function. The hyperplane is non-vertical otherwise.

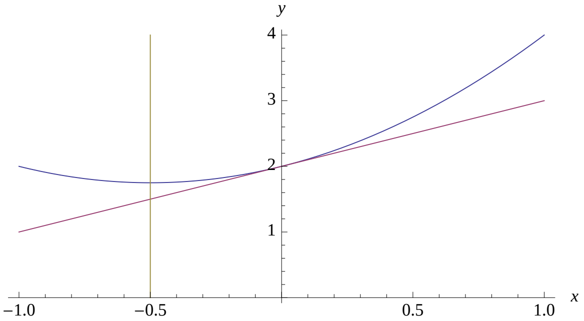
Because $\mathbf{x}^*$ is in the interior of S, we cannot have $a_0 = 0$: otherwise there would be points in the epigraph close to $(\mathbf{x}^*, f(\mathbf{x}^*))$ on either side of $H_{\mathbf{x}^*}$, contradicting the fact that it is a supporting hyperplane. Dividing (21.3.5) by $-a_0$, we get the standard equation for the non-vertical $H_{\mathbf{x}^*}$:

$$y = f(\mathbf{x}^*) + \langle \mathbf{a}, \mathbf{x} - \mathbf{x}^* \rangle = H_{\mathbf{x}^*}(\mathbf{x}). \tag{21.3.6}$$

For any $\mathbf{x}^1 \in S$, $H_{\mathbf{x}^*}$, which is not vertical, intersects the vertical line $\mathbf{x} = \mathbf{x}^1$ in a unique point with coordinates $(\mathbf{x}^1, H_{\mathbf{x}^*}(\mathbf{x}^1))$. Note that the convexity of M_f implies $f(\mathbf{x}^1) \geq H_{\mathbf{x}^*}(\mathbf{x}^1)$ for any $\mathbf{x}^1 \in S$.

A line in $\mathbb{R}^{n+1}$ is vertical if it is the intersection of n vertical and affinely independent hyperplanes. The line $\mathbf{x} = \mathbf{x}^1$ mentioned above is the intersection of the vertical hyperplanes $x_i = x_i^1$, $1 \leq i \leq n$.

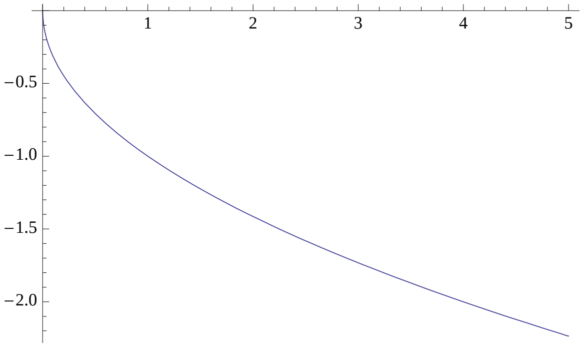
21.3.7 Example. We illustrate this with the convex function $f(x) = x^2 + x + 2$ and the point $x^* = 0$. There is a unique supporting hyperplane to the epigraph of f at 0: the tangent line to the function at $x = 0$, which has equation $y = x + 2$. Let $x_1 = -1/2$, so the intersection of the vertical line through $-1/2$ and the supporting hyperplane is the point $(-1/2, 3/2)$, which indeed is below the point on the graph at $-1/2$, which is $(-1/2, 7/4)$.



21.3.8 Lemma. *Convex functions are bounded from below on any bounded set.*

Proof. This follows because a convex function is bounded from below by the supporting hyperplane $H_{\mathbf{x}^*}$, for any $\mathbf{x}^*$ in the interior of its domain S. □

21.3.9 Example. Consider the function $f(x) = -\sqrt{x}$ on the interval $[0, \infty)$. A second derivative computation shows that it is convex. Notice that as $x \to \infty$, $f(x) \to -\infty$, showing that a convex function can fail to be bounded below on an unbounded set.



21.3.10 Example. Continuing with the same function $f(x)$, let

$$g(x) = \begin{cases} f(x), & \text{if } x > 0; \\ 1, & \text{if } x = 0. \end{cases}$$

It is easy to see that $g(x)$ is convex (check it). Obviously g is not continuous at 0. This can only happen because 0 is not an interior point of the domain of $g(x)$, as Theorem 21.2.8 shows.

21.3.11 Definition. A *subgradient hyperplane* at the point $(\mathbf{x}^*, f(\mathbf{x}^*))$ of the graph of f is a non-vertical supporting hyperplane at $(\mathbf{x}^*, f(\mathbf{x}^*))$ to the epigraph of f. We write its equation uniquely according to (21.3.6).

21.3.12 Definition. The n-vector $\mathbf{a}$ in the subgradient hyperplane equation (21.3.6) is called a *subgradient* of the function f restricted to S at $\mathbf{x}_0$.

Note that the subgradient determines the subgradient hyperplane. A subgradient, since it comes from a supporting hyperplane equation for the epigraph, satisfies the equation

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) + \mathbf{a}^T(\mathbf{x} - \mathbf{x}_0), \quad \forall \mathbf{x} \in S. \tag{21.3.13}$$

This can be written more symmetrically as

$$\langle \mathbf{x}_0, \mathbf{a} \rangle - f(\mathbf{x}_0) \geq \langle \mathbf{x}, \mathbf{a} \rangle - f(\mathbf{x}), \quad \forall \mathbf{x} \in S. \tag{21.3.14}$$

This formulation will be useful when we study the conjugate function f^* of f in §20.3:

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in S}(\langle \mathbf{x}, \mathbf{y} \rangle - f(\mathbf{x})).$$

What is the value of the conjugate at $\mathbf{a}$, where $\mathbf{a}$ is a subgradient of f at $\mathbf{x}_0$? Formulation (21.3.14) makes it clear that $f^*(\mathbf{a}) = \langle \mathbf{x}_0, \mathbf{a} \rangle - f(\mathbf{x}_0)$.

There can be many subgradients at a point. In Example 21.2.16, where $n = 1$, any a with $-1 \leq a \leq 1$ is a subgradient for f at $x = 0$.

21.3.15 Corollary. *If $f(\mathbf{x})$ is convex on the open convex set S, then the set of subgradients at any point $\mathbf{x} \in S$ is non-empty.*

Proof. The epigraph M_f, which is convex by Theorem 21.3.2, can be separated (in the sense of §18.6) from any point $(\mathbf{x}, f(\mathbf{x}))$ on its boundary. Since $\mathbf{x}$ is in the interior of S, we saw in Remark 21.3.4 that the supporting hyperplane cannot be vertical, which just means that we can find a supporting hyperplane through $(\mathbf{x}, f(\mathbf{x}))$ satisfying (21.3.6). □

As we will see in Theorem 22.1.2, the subgradient is unique when f is $\mathcal{C}^1$: it is then the gradient of f.

21.3.16 Example. Let $f(x) = -\sqrt{1 - x^2}$ on the closed interval $[-1, 1]$. Then f is convex on the entire interval, but it does not have a subgradient at either end point. This is because the tangent line to the semicircle becomes vertical as x approaches either end point.

21.4 Convex Function are Continuous on the Interior of their Domain

We now prove that convex functions in several variables are continuous at interior points of their domain, generalizing what we did in one variable (Theorem 21.2.8). The proof proceeds in the same way. We need a preliminary result:

21.4.1 Theorem. *Let $\mathbf{x}_0$ be a point in the interior of the domain S of the convex function f. Then there is a neighborhood U of $\mathbf{x}_0$ contained in the domain of f, on which $f(\mathbf{x})$ is bounded. In other words there exist numbers m and M such for all $\mathbf{y} \in U$, $m \leq f(\mathbf{y}) \leq M$.*

Proof. The lower bound follows from Lemma 21.3.8. A more elementary proof in the spirit of the one variable case is given below.

Now for the upper bound. Pick a point $\mathbf{x}_0$ in the interior of S, pick a neighborhood $N_r(\mathbf{x}_0)$ of $\mathbf{x}_0$ in S, and pick a simplex V with vertices $\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_n$ in $N_r(\mathbf{x}_0)$ such that $\mathbf{x}_0$ is the barycenter, or centroid of the simplex. See Definition 18.3.13. Thus $\mathbf{x}_0$ is in the interior of V.

Then any point $\mathbf{x}$ in the simplex can be written as a convex combination of the vertices $\mathbf{v}_i$ by Minkowski's Theorem 18.7.1:

$$\mathbf{x} = \sum_{i=0}^{n} \lambda_i \mathbf{v}_i, \text{ with } \lambda_i \geq 0 \text{ for all } i \text{ and } \sum_{i=0}^{n} \lambda_i = 1.$$

By Jensen's Inequality 21.1.11, we have

$$f(\mathbf{x}) = f\big(\sum_{i=0}^{n} \lambda_i \mathbf{v}_i\big) \leq \sum_{i=0}^{n} \lambda_i f(\mathbf{v}_i).$$

Let M be the largest of the $n+1$ numbers $f(\mathbf{v}_i)$, $0 \leq i \leq n$. Then

$$\sum_{i=0}^{n} \lambda_i f(\mathbf{v}_i) \leq \sum_{i=0}^{n} \lambda_i M = M,$$

so we are done.

Here is a simple method for getting a lower bound by using the upper bound. Again we choose the simplex so $\mathbf{x}_0$ is its barycenter, and pick a smaller neighborhood $N_s(\mathbf{x})$ inside the simplex. Then any point in $N_s(\mathbf{x})$ can be written $\mathbf{x}_0 + t\mathbf{u}$, where $\mathbf{u}$ has length s, and $-1 \leq t \leq 1$. The three points $\mathbf{x}_0 - t\mathbf{u}$, $\mathbf{x}_0$, and $\mathbf{x}_0 + t\mathbf{u}$ are aligned. Then by the argument given in the proof of Theorem 21.2.7, we see that

$$f(\mathbf{x}_0 + t\mathbf{u}) \geq 2f(\mathbf{x}_0) - M = m.$$

□

21.4.2 Remark. The proof shows that if $M = f(\mathbf{x}_0)$, then $m = f(\mathbf{x}_0)$, so the function is constant in a neighborhood of $\mathbf{x}_0$. This implies that f is constant on the interior of its domain. Thus the only convex functions with a local maximum on the interior of their domain are constant there.

21.4.3 Theorem. *A convex function f defined on an open set S is continuous.*

Proof. We use the one-variable proof. Pick any point $\mathbf{x}_0$ in S. We prove that f is continuous at $\mathbf{x}_0$. By Theorem 21.4.1 there is a closed ball B of radius $r > 0$ centered at $\mathbf{x}_0$ on which $f(\mathbf{x})$ is bounded above by M. We may as well take $M > f(\mathbf{x}_0)$. By definition, any point on the boundary of B can be written $\mathbf{x}_0 + \mathbf{u}$, where $\mathbf{u}$ is an arbitrary vector of length r. We restrict f to the line parametrized by $\mathbf{x}_0 + t\mathbf{u}$, $t \in \mathbb{R}$. Clearly any point at distance hr from $\mathbf{x}_0$ can be written $\mathbf{x}_0 + h\mathbf{u}$. We restrict h so that $0 < h < 1$. By Definition 11.1.2, we must show that for every $\epsilon > 0$ there is a δ so that for all $\mathbf{x}_0 + h\mathbf{u}$ with $|hr| < \delta$, then

$$|f(\mathbf{x}_0 + h\mathbf{u}) - f(\mathbf{x}_0)| < \epsilon.$$

By the proof of continuity in the one-variable case, it suffices to take

$$|h| < \frac{\epsilon}{M - f(\mathbf{x}_0}$$

since the bound M we are using here is a bound on each line through $\mathbf{x}_0$. We are done.

□

21.4.4 Corollary. *If $f : \mathbb{R}^n \to \mathbb{R}$ is convex on an open convex set S, and continuous on the closure $\overline{S}$, then it is convex on $\overline{S}$.*

Proof. We will prove directly that the Definition 21.1.1 of a convex function is satisfied on $\overline{S}$. By Theorem 18.4.1, $\overline{S}$ is convex, so the first condition is satisfied. Now take any two points $\mathbf{a}$ and $\mathbf{b}$ in $\overline{S}$, and take the segment $L \subset \overline{S}$ joining them. We can find sequences of points $\mathbf{a}^i$ and $\mathbf{b}^i$ in $L \cap S$ converging to $\mathbf{a}$ and $\mathbf{b}$. By the convexity of f on S, we have

$$f(\lambda\mathbf{a}^i + (1-\lambda)\mathbf{b}^i) \leq \lambda f(\mathbf{a}^i) + (1-\lambda)f(\mathbf{b}^i)$$

for all λ between 0 and 1. Because f is continuous, we can take the limit as $i \to \infty$, and get, since inequalities are preserved in the limit:

$$f(\lambda\mathbf{a} + (1-\lambda)\mathbf{b}) \leq \lambda f(\mathbf{a}) + (1-\lambda)f(\mathbf{b})$$

for all λ, so f is convex on $\overline{L}$. □

21.5 The Closure of Convex Functions

Convex functions may fail to be continuous on the boundary of their domain. So an operation called *closure* was invented by Fenchel to modify the convex function on its boundary without changing it on its interior, in such a way that it becomes as close to continuous as possible. First we make a general definition.

21.5.1 Definition. A function f from $S \subset \mathbb{R}^n$ to $\mathbb{R}$ is *closed* if its sublevels sets S_c are closed relative to S for all $c \in \overline{\mathbb{R}}$.

As an immediate consequence of Theorem 16.3.6, we get:

21.5.2 Theorem. *If f is lower semicontinuous on S, then it is closed.*

On the other hand

21.5.3 Proposition. *f is closed if and only if its epigraph is a closed set relative to $S \times \mathbb{R}$.*

Proof. The sublevel set S_c is the set of points $\mathbf{x}$ in S such that $f(\mathbf{x}) \leq c$, while the epigraph M_f is the set of points $(\mathbf{x}, y)$ in $S \times \mathbb{R}$ such that $f(\mathbf{x}) \leq y$. Then S_c is the intersection of M_f with the "horizontal" hyperplane H_c with equation $y = c$. Then if M_f is relatively closed, $M_f \cap H_c$ is too. On the other hand the relative closure of $S_{-\infty}$ implies that of M_f. □

Now assume that S is a convex set, that f is convex on S, and that U is the relative interior of S. Since S usually has dimension n, U is usually just the interior of S.

21.5.4 Definition. We redefine f to a function $\overline{f}$ on $\overline{S}$ as follows. We call $\overline{f}$ the *lower semicontinuous extension* of f. It agrees with f on U and extends to the boundary of U as follows. Let x^* belong to the boundary of U. Then x^* may be in S (for example, if S is closed), but it may not. Then

$$\overline{f}(\mathbf{x}^*) = \liminf_{\mathbf{x} \in U \to \mathbf{x}^*} f(\mathbf{x}),$$

allowing the value ∞ for $\overline{f}(\mathbf{x}^*)$.

Because f is convex, it is continuous on U. Then by construction $\overline{f}$ is lower semicontinuous on the closure $\overline{S}$ of S.

21.5.5 Theorem. *The subset S' of $\overline{S}$ where $\overline{f}$ takes finite values is convex, and $\overline{f}$ is convex on S'.*

The set $\overline{S}$ is a convex set , and, allowing infinite values in the definition of convexity, $\overline{f}$ is convex and closed on $\overline{S}$.

Proof. Consider the closure $\overline{M}_f$ of the epigraph M_f of f on S. By Theorem 18.4.1, $\overline{M}_f$ is convex, so its intersection with any vertical line (meaning a line with equations $x_i = x_i^*$, $1 \leq i \leq n$, for a fixed point $\mathbf{x}^* \in \overline{S}$) is either a closed segment that is unbounded above, or is empty. In the case we get a segment, we define $\overline{f}(\mathbf{x}^*)$ to be the lower end point of the segment. If the intersection is empty, we define $\overline{f}(\mathbf{x}^*) = \infty$. By construction, this function $\overline{f}$ has $\overline{M}_f$ as epigraph, and since $\overline{M}_f$ is a convex set, by Theorem 21.3.2 $\overline{f}$ is convex. This actually requires a slight extension of Theorem 21.3.2 allowing for infinite values of f. By construction $\overline{f}$ agrees with f on U, and since its epigraph, and therefore its sublevel sets by Theorem 16.3.6 are closed, $\overline{f}$ is lower semicontinuous, so it is the same function as in Definition 21.5.4.

Finally we need to show that S' is a convex set. Just take two points $\mathbf{x}^0$ and $\mathbf{x}^1$ where $\overline{f}$ takes finite values. Since $\overline{f}$ is convex on the segment $\left[\mathbf{x}^0, \mathbf{x}^1\right]$, its values there are also finite. □

This method of closing convex functions allows us to define quasiconvex functions in §24.1.

Lecture 22

Convex Functions and Optimization

Continuing our study of convex functions, we prove two results that will be very useful to us later:

- Corollary 22.1.4. It says that the graph of a differentiable convex function on an open set S lies above its tangent plane at any point; the following Corollary 22.1.6 shows that this characterizes differentiable convex functions on open sets.
- Theorems 22.2.1 and 22.2.2 for $\mathcal{C}^2$ convex functions. There results show that the correct generalization of the condition $f''(x) \geq 0$ (resp. $f''(x) > 0$) for a function of one variable is that the Hessian be positive semidefinite (resp. definite) for a function of several variables.

These results allow us to prove the key theorems for unconstrained minimization of convex functions on open convex sets: Theorem 22.4.1 and Corollary 22.4.4. These theorems greatly simplify the search for the minimum when the objective function is convex. You should compare Theorem 22.4.8 to the results on unconstrained optimization in §13.1 and §13.5, to appreciate how convexity facilitates the search for a minimum. The results just quoted are the most important results of this lecture.

A list of convex functions is given in §22.3. Since knowing that the objective function is convex makes finding its minimum and minimizers much easier, these examples should be studied carefully.

In §22.4 we also discuss minimization of convex functions on arbitrary convex sets, using two tools:

- Reduction to dimension 1 by considering all affine lines. This works well because the domain of the convex function is always assumed to be convex, so the intersection of the line with the domain is a segment. Another way of saying this is that from any point in the domain of a convex function one can look straight at any other point.

- The tool kit of convex sets: most importantly the separation theorems in §18.6.

Finally, in a different vein, we show in §22.6 how convexity arguments can be used to prove some of the key inequalities of analysis, starting with the arithmetic-geometric mean inequality. This section will not be used later.

22.1 Differentiable Convex Functions

The most important result of this section is Corollary 22.1.4 which compares the graph of a differentiable convex function with that of its tangent plane at any point. First, here is the main theorem on differentiability of convex functions.

22.1.1 Theorem. *If f is convex on an open convex set S, and all the partial derivatives of f exist at every point of S, then f is continuously differentiable on S.*

As we saw in Example 12.1.15, for an arbitrary function, the existence of partial derivatives does not guarantee that the function is differentiable, much less continuously differentiable ($\mathcal{C}^1$). Thus the situation for convex functions is much better. Theorem 22.1.1 is proved at the end of this section. Before dealing with it, we discuss the subgradients of a differentiable convex function.

22.1.2 Theorem. *If the convex function f, on the convex set S, is differentiable at the interior point $\mathbf{x}_0$ of S, then there is only one supporting hyperplane to its epigraph M_f at $(\mathbf{x}_0, f(\mathbf{x}_0))$. It is the tangent hyperplane to the graph of f at $\mathbf{x}_0$, so it has equation*

$$y = f(\mathbf{x}_0) + \nabla f|_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0) \tag{22.1.3}$$

Thus the only subgradient to f at $\mathbf{x}_0$ is the gradient[1].

Proof. This is an elementary fact, because we can reduce to the one-dimensional case. We argue by contradiction. Indeed if there were a separating hyperplane different from the tangent hyperplace, there would be a direction $\mathbf{v}$ through $\mathbf{x}_0$ along which the tangent hyperplane and this hyperplane do not agree. So slice the whole picture by the ordinary (meaning two-dimensional) vertical plane containing $\mathbf{v}$. This amounts to restricting f to the line through $f(\mathbf{x}_0))$ in direction $\mathbf{v}$. The restriction, which we call g is clearly convex, and is a function of one variable.

[1] or rather, the transpose of the gradient, since the gradient is the unique vector that we always write as a row.

According to our claim there is a line L through $\mathbf{x}_0$, other than the tangent line T to the graph of g through $\mathbf{x}_0$, that separates the epigraph of g from $\mathbf{x}_0$. It follows from the definition of the derivative, or Taylor's theorem (see Lecture 12) that this is impossible: if $r(x) = f(x) - f(x_0) - f'(x_0)(x - x_0)$, then $lim_{x \to x_0} \frac{r(x)}{x - x_0} = 0$. This implies that on the side of x_0 where the line L is above the tangent line, the points on the graph of f close enough to x_0 on that side will be below L, so L does not separate. □

The following corollary is one of the most useful results on convex functions: it says that at every point $\mathbf{x}$ where f is differentiable, the tangent plane approximation of the function underestimates the value of the function, not only near $\mathbf{x}$, but on the whole region where f is convex.

22.1.4 Corollary. *If $f(\mathbf{x})$ is convex on the open convex set S, and differentiable at $\mathbf{x}_0 \in S$, then*

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) + \nabla f|_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0), \quad \forall \mathbf{x} \in S. \tag{22.1.5}$$

Proof. As noted above, the equation of the tangent hyperplane to the graph of f is

$$y = f(\mathbf{x}_0) + \nabla f|_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0).$$

So the right-hand side of (22.1.5) is just the y value at the point $\mathbf{x}$ of the separating hyperplane of M_f at $\mathbf{x}_0$, so that it is less than or equal to the value of the function f at the same point $\mathbf{x}$. □

The hypotheses are weak: the function f only needs to be differentiable at the point $\mathbf{x}_0$.

22.1.6 Corollary. *If $f(\mathbf{x})$ is defined and differentiable on the open convex set S, then $f(\mathbf{x})$ is convex if and only if for each $\mathbf{x}_0 \in S$, f satisfies* (22.1.5) *for all $\mathbf{x} \in S$. Furthermore, f is strictly convex if and only if for all $\mathbf{x}_0 \in S$ and all $\mathbf{x} \neq \mathbf{x}_0 \in S$ the inequality in* (22.1.5) *is strict.*

Proof. If f is differentiable, by Corollary 22.1.4 the tangent plane to the graph of f at $\mathbf{x}_0$ is the unique candidate for a separating hyperplane for the epigraph of f and the point $\mathbf{x}_0, f(\mathbf{x}_0)$. If the epigraph is a convex set, it must have a supporting hyperplane at every point of the boundary by Theorem 18.6.11, so (22.1.5) must be satisfied at every point. The last assertion follows immediately from Definition 21.1.7 □

We can often reduce differentiability questions for convex functions to the one variable case by the following device.

22.1.7 Exercise. Given a convex function $f(\mathbf{x})$, fix a point $\mathbf{x}$ in the interior of the domain of f, and a vector $\mathbf{y}$ representing a direction at $\mathbf{x}$. Show that the function of one variable

$$g(t) = f(\mathbf{x} + t\mathbf{y})$$

is convex for those values of t that correspond to points in the domain of f. Show that this includes a small interval $-a \leq t \leq a$ around the origin. We studied a special case of this function in Exercise 21.1.8.

Hint: Pick two points t_1 and t_2 in the domain of g. You need to verify (21.1.2) for these points. Write out what this says for f. Note that

$$f(\mathbf{x} + (\lambda t_1 + (1-\lambda)t_2)\mathbf{y}) = f(\lambda(\mathbf{x} + t_1\mathbf{y}) + (1-\lambda)(\mathbf{x} + t_2\mathbf{y})).$$

Conclude by using the convexity of f.

22.1.8 Corollary. *Let f be a differentiable function on an open convex set S. Then f is convex if and only if for any two points $\mathbf{x}_1$ and $\mathbf{x}_2$ in S,*

$$\langle \nabla f|_{\mathbf{x}_1} - \nabla f|_{\mathbf{x}_2}, \mathbf{x}_1 - \mathbf{x}_2 \rangle \geq 0; \tag{22.1.9}$$

and f is strictly convex if and only if the inequality is strict for distinct points.

Proof. First assume that f is convex. Add the following two copies of (22.1.5) evaluated at $\mathbf{x}_2$ and then $\mathbf{x}_1$:

$$\begin{aligned} f(\mathbf{x}_1) &\geq f(\mathbf{x}_2) + \nabla f|_{\mathbf{x}_2}(\mathbf{x}_1 - \mathbf{x}_2), \\ f(\mathbf{x}_2) &\geq f(\mathbf{x}_1) + \nabla f|_{\mathbf{x}_1}(\mathbf{x}_2 - \mathbf{x}_1), \end{aligned}$$

and write the gradients as dot products to get (22.1.9).

Here is another way to derive the result by interpreting it in terms of the single variable differentiable convex function introduced in Exercise 22.1.7: $g(t) = f(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1))$. Compute the derivative $g'(t)$ by the chain rule, evaluate at $t = 0$ to get

$$g'(0) = \langle \nabla f|_{\mathbf{x}_1}, \mathbf{x}_2 - \mathbf{x}_1 \rangle,$$

and then at $t = 1$ to get

$$g'(1) = \langle \nabla f|_{\mathbf{x}_2}, \mathbf{x}_2 - \mathbf{x}_1 \rangle.$$

Since g is convex, $g'(1) \geq g'(0)$, which gives a second proof.

Next we show the opposite implication. Assuming (22.1.9) for all $\mathbf{x}_1$ and $\mathbf{x}_2$, we must show that f is convex. We show it satisfies (21.1.2) for all $\mathbf{x}_1$ and $\mathbf{x}_2$. This is equivalent to saying that the function $g(t)$ introduced earlier in the proof is convex, for all choices of $\mathbf{x}_1$ and $\mathbf{x}_2$, so we are done.

Finally to get strict convexity, just replace the inequalities by strict inequalities in the proof above. □

22.1.10 Corollary. *If f is a strictly convex differentiable function on an open convex set S, and $\mathbf{x}_1$ and $\mathbf{x}_2$ are distinct points in S, then $\nabla f|_{\mathbf{x}_1} \neq \nabla f|_{\mathbf{x}_2}$.*

Proof. Assume the two gradients are equal, and dot their difference with the vector $\mathbf{x}_1 - \mathbf{x}_2$, showing that the left-hand side of (22.1.9) is 0. This contradicts Corollary 22.1.8. □

Next we determine the gradient mapping when the function fails to be strictly convex: it is constant.

22.1.11 Corollary. *Assume that the convex differentiable function f fails to be strictly convex on the segment $\mathcal{I} = [\mathbf{x}_1, \mathbf{x}_2]$ in the interior of its domain. Then the gradient $\nabla f|_{\mathbf{x}}$ is constant along $\mathcal{I}$.*

Proof. By Exercise 21.1.8, the graph of f over $\mathcal{I}$ is a line. Once that is noted, the result is an easy corollary of Theorem 22.1.2. Pick a point $\mathbf{x}^*$ in the interior of $\mathcal{I}$. Because f is differentiable, its epigraph has a unique supporting hyperplane, namely the tangent hyperplane to the graph of f at $\mathbf{x}^*$ with equation

$$y = f(\mathbf{x}^*) + \nabla f|_{\mathbf{x}^*}(\mathbf{x} - \mathbf{x}^*)$$

Because $\mathbf{x}^*$ was chosen in the interior of $\mathcal{I}$, its supporting hyperplane goes through all the other points of the segment, by the same argument that proves Theorem 22.1.2, so it is a supporting hyperplane at those points too. Therefore it is the tangent hyperplane at all $\mathbf{x} \in \mathcal{I}$, including the endpoints since we know that f is $\mathcal{C}^1$ by Theorem 22.1.1. Thus the tangent hyperplane does not vary, which proves the result. □

Here is a converse to Theorem 22.1.2 that we will not prove.[2]

22.1.12 Theorem. *If the convex function f on the convex set S has a unique supporting hyperplane to its epigraph M_f at the interior point $\mathbf{x}_0$ of S, then f is differentiable at $\mathbf{x}_0$ and the supporting hyperplane is the tangent hyperplane to the graph of f at $\mathbf{x}_0$, so its equation is* (22.1.3).

Proof of Theorem 22.1.1. By Theorem 12.1.17, it is enough to show that the partial derivatives of f are continuous. We use the function $g(t)$ of Exercise 22.1.7 to apply the results of §21.2 to $g(t)$, especially Theorems 21.2.2 and 21.2.14.

Let $\mathbf{x}$ be an interior point of the domain S, and let $\mathbf{x}_i$ be a sequence of points in S approaching $\mathbf{x}$. Let $\mathbf{y}$ be an arbitrary vector of length 1 in $\mathbb{R}^n$, representing a direction, and λ a small positive number so that the points $\mathbf{x}_i \pm \lambda\mathbf{y}$ are in S. To

[2] For the proof, see [7] p. 106.

prove the result, we only need the case where $\mathbf{y}$ is a coordinate vector $\mathbf{e}_j$, but it is no harder to prove the result for any unit vector, which is what we do.

First, without making any assumption on f other than convexity, just as in the single variable case (see Definition 21.2.13), the limit

$$f'(\mathbf{x};\mathbf{y}) := \lim_{\lambda\searrow 0} \frac{f(\mathbf{x}+\lambda\mathbf{y}) - f(\mathbf{x})}{\lambda}$$

exists, since in terms of the function $g(t)$ defined above, we have written $g'_+(0) = \lim_{\lambda\searrow 0} \frac{g(\lambda)-g(0)}{\lambda}$. The key point, supplied by convexity, is that the expression inside the limit decreases to $f'(\mathbf{x};\mathbf{y})$.

The limit is called the *one-sided directional derivative* of f at $\mathbf{x}$ in the direction $\mathbf{y}$, and is written $f'(\mathbf{x};\mathbf{y})$. In the same way, we can define the one-sided directional derivative of f at $\mathbf{x}$ in the direction $-\mathbf{y}$. It is just $g'_-(0)$.

Because the partial derivative of f in the direction $\mathbf{y}$ exists by hypothesis, these two one-sided directional derivatives are equal:

$$f'(\mathbf{x};\mathbf{y}) = f'(\mathbf{x};-\mathbf{y}).$$

Call this quantity d. Since f is continuous at on the interior of its domain by Theorem 21.4.3, the sequence $f(\mathbf{x}_i+\lambda\mathbf{y})$ approaches $f(\mathbf{x}+\lambda\mathbf{y})$, and the sequence $f(\mathbf{x}_i)$ approaches $f(\mathbf{x})$.

For any $\epsilon > 0$, for all small enough $\lambda > 0$, we have

$$d \leq \frac{f(\mathbf{x}+\lambda\mathbf{y}) - f(\mathbf{x})}{\lambda} < d + \epsilon,$$

since the left-hand side is the limit of the right-hand side as λ decreases to 0. Fix a $\lambda > 0$ so that this inequality is satisfied.

Then

22.1.13 Lemma. *For i_0 sufficiently large, we get the key inequality:*

$$\frac{f(\mathbf{x}_i+\lambda\mathbf{y}) - f(\mathbf{x}_i)}{\lambda} < d + 3\epsilon, \quad \forall i \geq i_0.$$

Proof. Let

$$\epsilon' = \frac{f(\mathbf{x}+\lambda\mathbf{y}) - f(\mathbf{x})}{\lambda} - d < \epsilon$$

for our chosen λ, so $\epsilon' > 0$. Then choose i_0 large enough so that for all $i \geq i_0$,

$$|f(\mathbf{x}+\lambda\mathbf{y}) - f(\mathbf{x}_i+\lambda\mathbf{y})| < \epsilon'\lambda$$

and

$$|f(\mathbf{x}) - f(\mathbf{x}_i)| < \epsilon'\lambda.$$

Then

$$\frac{|f(\mathbf{x}_i + \lambda\mathbf{y}) - f(\mathbf{x}_i)|}{\lambda} \leq \frac{|f(\mathbf{x}_i + \lambda\mathbf{y}) - f(\mathbf{x} + \lambda\mathbf{y})|}{\lambda} + \frac{|f(\mathbf{x} + \lambda\mathbf{y}) - f(\mathbf{x})|}{\lambda} + \frac{|f(\mathbf{x}) - f(\mathbf{x}_i)|}{\lambda} < \epsilon' + d + \epsilon + \epsilon' \leq d + 3\epsilon.$$

□

We now repeat the argument with $-\mathbf{y}$. We get the parallel lemma:

22.1.14 Lemma. *For i_0 sufficiently large:*

$$d - 3\epsilon < \frac{f(\mathbf{x}_i - \lambda\mathbf{y}) - f(\mathbf{x}_i)}{\lambda}, \quad \forall i \geq i_0.$$

As ϵ goes to 0, the left-hand and the right-hand directional derivatives of f at $\mathbf{x}_i$ in the direction $\mathbf{y}$ both converge to the directional derivative of f at $\mathbf{x}$ in the direction $\mathbf{y}$, showing that the directional derivatives, and therefore the partial derivatives are continuous. □

The proofs of this result in the literature (for example, [66], p. 150, [53], p.244, and [22], pp.102) derive this result as a corollary of statements concerning the convergence of families of convex functions. More elementary books usually only prove the weaker result that f is differentiable: see [7] p. 99 or [52].

22.2 Twice Differentiable Convex Functions

22.2.1 Theorem. *If $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}^2$ (twice continuously differentiable) on an open convex set S, then $f(\mathbf{x})$ is convex on S if and only if the Hessian matrix $F(\mathbf{x})$ is positive semidefinite for all $\mathbf{x} \in S$.*

Proof. We noted in Theorem 21.2.1 that a function is convex if and only if its restriction to every line segment is convex. Furthermore by Definition 8.1.7 a symmetric matrix is positive semidefinite if and only if its restriction to any line is non-negative. So pick any point $\mathbf{x} \in S$, and consider the Hessian $F(\mathbf{x})$ at $\mathbf{x}$. Because S is open, for any non-zero vector $\mathbf{h}$ the line segment $\mathbf{x} + t\mathbf{h}$ is in S for small enough t. The second derivative of the composite function $f(\mathbf{x} + t\mathbf{h})$ considered as a function of t at $t = 0$ is easily seen by the chain rule to be $\mathbf{h}^T F(\mathbf{x})\mathbf{h}$.

After these preliminary remarks, we prove the theorem. First assume f is convex. Then its restriction to any line is convex. Such a line is determined by a point $\mathbf{x} \in S$ and an arbitrary direction $\mathbf{h}$. By Theorem 21.2.20, convexity on a line is determined by the non-negativity of the second derivative, which is our context is $\mathbf{h}^T F(\mathbf{x})\mathbf{h}$. Since this is true for all $\mathbf{h}$, Definition 8.1.7 tells us that $F(\mathbf{x})$ is positive semidefinite. Since this is satisfied for all $\mathbf{x}$, we get one implication of the theorem.

Now assume that the Hessian matrix $F(\mathbf{x})$ is positive semidefinite for all $\mathbf{x} \in S$. Then the function f is convex on every line by using the other implication of Theorem 21.2.20, and therefore it is convex on the entire open set S by Theorem 21.2.1.

□

This theorem is useful because it reduces the determination of convexity to a second derivative computation, just as in the one-variable case: Theorem 21.2.20. We see that positive semi-definiteness of the Hessian is what generalizes $f'' \geq 0$ in the one dimensional case. Positive definiteness is the generalization of $f'' > 0$, and we have

22.2.2 Theorem. *If $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}^2$ on an open convex set S, then $f(\mathbf{x})$ is strictly convex on S if the Hessian matrix $F(\mathbf{x})$ is positive definite for all $\mathbf{x} \in S$.*

The proof uses the last statement of Theorem 21.2.20 and is otherwise identical to that of Theorem 22.2.1. The converse fails, as we already know from Example 21.2.21.

22.2.3 Example (The Quadratic Function)**.** The most general quadratic function in $\mathbb{R}^n$ can be written

$$f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} + c,$$

where A is a symmetric $n \times n$ matrix, $\mathbf{b}$ a vector and c a number. Note that

$$\nabla f(\mathbf{x}) = 2A\mathbf{x}$$

and the Hessian matrix of f is just the constant matrix $2A$. So f is convex if and only if A is positive semidefinite, and is strictly convex if and only if A is positive definite. Thus we can check convexity by analyzing one matrix using the techniques of §9.4.

22.2.4 Corollary. *Assume f is $\mathcal{C}^2$ on an open convex set S, and continuous on the closure $\overline{S}$ of S. Also assume that the Hessian of f is positive semidefinite on S. Then f is convex on $\overline{S}$.*

22.2.5 Exercise. Prove Corollary 22.2.4, using Corollary 21.4.4 and Theorem 22.2.1.

22.2.6 Exercise. Study the convexity of the function $f(x, y) = x(x + y^2)$ on $\mathbb{R}^2$.

1. Find the biggest (convex) open set U in $\mathbb{R}^2$ on which f is convex.
2. Show that f is convex on the closure of U.
3. Show that this example behaves just like Peano's Example 13.5.3 at the origin: there is a minimum on each line through the origin, but the origin is not a local minimum for the function on $\mathbb{R}^2$.

22.3 Examples of Convex Functions

22.3.1 Example. The exponential function e^{ax} is convex on $\mathbb{R}$ for any $a \in \mathbb{R}$.

This is established by computing the second derivative $a^2 e^{ax}$ and using Theorem 21.2.20.

22.3.2 Example. The logarithm $\ln x$ is concave on its domain $x > 0$.

Since $\ln x$ is the inverse of e^x this follows from Proposition 21.2.22 and the convexity of e^x.

22.3.3 Example. Powers of x: x^a is convex on $x > 0$ when $a \geq 1$ or $a \leq 0$, and concave for $0 \leq a \leq 1$. Furthermore $|x|^a$, $a \geq 1$ in convex on $\mathbb{R}$.

The second derivative of x^a is $a(a-1)x^{a-2}$. So when $x > 0$ this is nonnegative when $a \geq 1$ or $a < 0$.

22.3.4 Example (The distance between two points)**.**

We can think of the distance function $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$ on $\mathbb{R}^n$ as a function from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}$. In other words it is a function of $2n$ variables: the coordinates of $\mathbf{x}$ and those of $\mathbf{y}$.

Then $d(\mathbf{x}, \mathbf{y})$ is a convex function.

Proof. Its domain, $\mathbb{R}^{2n}$, is convex, so we have to verify (21.1.2). Pick two pairs of arbitrary points $(\mathbf{x}_0, \mathbf{y}_0)$ and $(\mathbf{x}_1, \mathbf{y}_1)$ in the domain of d. We must show that for any λ between 0 and 1 we have

$$d\big((1-\lambda)\mathbf{x}_0 + \lambda\mathbf{x}_1, (1-\lambda)\mathbf{y}_0 + \lambda\mathbf{y}_1\big) \leq (1-\lambda)d(\mathbf{x}_0, \mathbf{y}_0) + \lambda d(\mathbf{x}_1, \mathbf{y}_1).$$

In terms of lengths of vectors, this says:

$$\|(1-\lambda)(\mathbf{x}_0 - \mathbf{y}_0) + \lambda(\mathbf{x}_1 - \mathbf{y}_1)\| \leq (1-\lambda)\|\mathbf{x}_0 - \mathbf{y}_0\| + \lambda\|\mathbf{x}_1 - \mathbf{y}_1\|.$$

First we apply the triangle inequality (see Theorem 5.4.17) to $(1-\lambda)(\mathbf{x}_0 - \mathbf{y}_0)$ and $\lambda(\mathbf{x}_1 - \mathbf{y}_1)$ and their sum to get

$$\|(1-\lambda)(\mathbf{x}_0 - \mathbf{y}_0) + \lambda(\mathbf{x}_1 - \mathbf{y}_1)\| \leq \|(1-\lambda)(\mathbf{x}_0 - \mathbf{y}_0)\| + \|\lambda(\mathbf{x}_1 - \mathbf{y}_1)\|.$$

Then we can pull out the positive scalar factors $(1-\lambda)$ and λ to get the result. □

22.3.5 Example (The distance between a point and a convex set)**.** Now let C be a convex set in $\mathbb{R}^n$, and define the distance $D_C(\mathbf{x})$ as

$$D_C(\mathbf{x}) = \inf_{\mathbf{c}\in C} d(\mathbf{x}, \mathbf{c}).$$

Then $D_C(\mathbf{x})$ is a convex function.

Proof. In Exercise 11.1.5 we showed that this function is continuous for any set C, not just convex sets: we will not need this result here, as we will get convexity directly. Pick any $\mathbf{x}_0$ and $\mathbf{x}_1$ in $\mathbb{R}^n$. By the definition of $\inf$, for any $\epsilon > 0$, we can find elements $\mathbf{c}_0$ and $\mathbf{c}_1$ in C, such that

$$\begin{aligned} d(\mathbf{x}_0, \mathbf{c}_0) &\le D_C(\mathbf{x}_0) + \epsilon;\\ d(\mathbf{x}_1, \mathbf{c}_1) &\le D_C(\mathbf{x}_1) + \epsilon. \end{aligned}$$

To prove that D_C is convex, for any λ between 0 and 1, we must show

$$D_C\big((1-\lambda)\mathbf{x}_0 + \lambda\mathbf{x}_1)\big) \le (1-\lambda)D_C(\mathbf{x}_0) + \lambda D_C(\mathbf{x}_0). \tag{22.3.6}$$

Now

$$\begin{aligned} D_C\big((1-\lambda)\mathbf{x}_0 + \lambda\mathbf{x}_1)\big) &= \inf_{\mathbf{c}\in C} d\big((1-\lambda)\mathbf{x}_0 + \lambda\mathbf{x}_1, \mathbf{c}\big)\\ &\le d\big((1-\lambda)\mathbf{x}_0 + \lambda\mathbf{x}_1), (1-\lambda)\mathbf{c}_0 + \lambda\mathbf{c}_1\big)\\ &\le (1-\lambda)d(\mathbf{x}_0, \mathbf{c}_0) + \lambda d(\mathbf{x}_1, \mathbf{c}_1)\\ &\le (1-\lambda)D_C(\mathbf{x}_0) + \lambda D_C(\mathbf{x}_1) + \epsilon. \end{aligned}$$

To get to the second line of this chain of inequalities, we use the fact that C is convex, so $(1-\lambda)\mathbf{c}_0 + \lambda\mathbf{c}_1$ is in C. On the next line we use the fact that $d(\mathbf{x}, \mathbf{y})$ is a convex function, as we showed in the previous example. On the last line we use the definition of $\mathbf{c}_0$ and $\mathbf{c}_1$, plus the obvious fact $(1-\lambda)\epsilon + \lambda\epsilon = \epsilon$.

Since this chain of inequalities is true for all $\epsilon > 0$, by letting ϵ go to 0, it implies (22.3.6), and we are done. □

22.3.7 Example (A Thickened Convex Set)**.** Let $C \subset \mathbb{R}^n$ be a convex set. For any non-negative number r, let $C_r = \{\mathbf{x} \in \mathbb{R}^n \mid D_C(\mathbf{x}) \le r\}$, where D_C is defined in the previous example. So C_r is the set of points at distance at most r from C. Then C_r is a closed convex set containing C.

Indeed, it is the sublevel set (see Theorem 21.1.15) of the convex function D_C. Note that C_0 is the closure of C. Furthermore, if C is bounded, then C_r is compact.

22.3.8 Example. If f and g are convex functions, and a and b are non-negative, then $af + bg$ is convex.

We need to show that

$$af(\lambda\mathbf{x}^1+(1-\lambda)\mathbf{x}^2)+bg(\lambda\mathbf{x}^1+(1-\lambda)\mathbf{x}^2) \le a\lambda(f(\mathbf{x}^1)+g(\mathbf{x}^1))+b\lambda(f(\mathbf{x}^2)+g(\mathbf{x}^2))$$

This is simply the sum of the convexity inequalities for f and g multiplied by the nonnegative constants a and b, so that the inequality is preserved.

22.3.9 Example. Composition: assume the functions $h : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ can be composed, with $f(\mathbf{x}) = h(g(\mathbf{x}))$. Then

- f is convex if h is convex and nondecreasing, and g is convex.
- f is convex if h is convex and nonincreasing, and g is concave.

We just sketch the first case. Since g is convex,

$$g(\lambda\mathbf{x}^1 + (1-\lambda)\mathbf{x}^2) \le \lambda g(\mathbf{x}^1) + (1-\lambda)g(\mathbf{x}^2).$$

Apply the increasing function h to both sides, so that in direction of the inequality is preserved,

$$h(g(\lambda\mathbf{x}^1 + (1-\lambda)\mathbf{x}^2)) \le h(\lambda g(\mathbf{x}^1) + (1-\lambda)g(\mathbf{x}^2)).$$

Using convexity of h on the right-hand side of this equation, we get the desired conclusion.

22.3.10 Example. If $f : \mathbb{R}^m \to \mathbb{R}$ is a convex function, and A is a $m \times n$ matrix, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^n$, then $f(A\mathbf{x} + \mathbf{b})$ is a convex function where defined. The proof is similar to that in Example 22.3.8.

22.3.11 Example. If f and g are convex functions, then the max function h defined by $h(x) = max\{f(x), g(x)\}$ is convex.

More generally, the pointwise least upper bound (the supremum) of any (possibly infinite) family of convex functions is convex.

Let $\{f_\alpha\}$ be a collection (perhaps infinite) of convex functions, all defined on an open convex set S. The least upper bound (or supremum or $\sup$) F of this family is defined as follows:

$$F(\mathbf{x}) = \sup_\alpha f_\alpha(\mathbf{x})$$

Let M_α be the epigraph of f_α. By Theorem 21.3.2 M_α is a convex set. The supremum F of the f_α has as its epigraph the intersection $\cap_\alpha M_\alpha$. As we saw in Theorem 18.1.15, this intersection of convex sets is convex, so we use the opposite implication of Theorem 21.3.2 to conclude.

22.3.12 Exercise.

1. Show that the set $S = \{(x, y, z) \mid x > 0, y > 0, z > 0\}$ in $\mathbb{R}^3$ is convex.
2. Show that on S, the function

   $$f(x, y, z) = \frac{1}{xyz} + x + y + z$$

   is convex by computing its Hessian and quoting a theorem.
3. In an alternate approach, establish that $\ln \frac{1}{xyz}$ is convex on S, use Example 22.3.8, and Example 22.3.9, to give an alternate proof of 2.
4. Minimize the function f: show that the minimum value is 4, and determine where it is attained.

This function is related to the Cobb-Douglas function of Example 13.4.1.

22.4 Optimization of Convex Functions

We start with the theorem that makes the minimization of convex functions so pleasant.

22.4.1 Theorem. *Let f be a convex function defined on the convex set S. Then if f has a local minimum at a point $\mathbf{x}_0$, it has a global minimum there. The set of points $M \subset S$ where the global minimum of f is attained is either empty or convex. If f is strictly convex in a neighborhood of a local minimizer $\mathbf{x}_0$, then $\mathbf{x}_0$ is the unique global minimizer.*

Proof. Assume the local minimum at $\mathbf{x}_0$ is not the global minimum: that means that there is a point $\mathbf{x}_1 \in S$ with $f(\mathbf{x}_1) < f(\mathbf{x}_0)$. Consider the segment $[\mathbf{x}_0, \mathbf{x}_1]$ parametrized by λ: $(1-\lambda)\mathbf{x}_0 + \lambda\mathbf{x}_1$, with $0 \leq \lambda \leq 1$. Since S is convex, this segment is in S, and for any λ, $0 \leq \lambda \leq 1$, (21.1.2) holds, so

$$f\big((1-\lambda)\mathbf{x}_0 + \lambda\mathbf{x}_1\big) \leq (1-\lambda)f(\mathbf{x}_0) + \lambda f(\mathbf{x}_1) = f(\mathbf{x}_0) + \lambda\big(f(\mathbf{x}_1) - f(\mathbf{x}_0)\big).$$

Because $f(\mathbf{x}_1) < f(\mathbf{x}_0)$, the value of f at any point of the segment corresponding to $\lambda > 0$ is less than that at $\mathbf{x}_0$. This contradicts the assertion that $\mathbf{x}_0$ is a local minimizer, since there are points in S arbitrarily close to $\mathbf{x}_0$ with smaller value than $f(\mathbf{x}_0)$.

Next we prove that the set of minimizers M is convex or empty. If M is empty, there is nothing to prove. If M is non-empty, let c be the minimum value of f on

S. Then the sublevel set S_c is the same as the level set of f at c, namely M, since all the lower level sets are empty. By Theorem 21.1.15, S_c is convex, so we are done.

The final assertion is left to you as an exercise. □

22.4.2 Exercise. Show that if f is strictly convex in a neighborhood of a local minimizer $\mathbf{x}_0$, then $\mathbf{x}_0$ is the unique global minimizer.

Hint: We have already shown that $\mathbf{x}_0$ is a global minimizer. Assume there is a second global minimizer $\mathbf{x}_1$, and use the segment $[\mathbf{x}_0, \mathbf{x}_1]$ to derive a contradiction to strict convexity near $\mathbf{x}_0$.

Next an easy result improving on Theorem 13.1.1 in several ways. First, f is not assumed to be $\mathcal{C}^1$; secondly, the convexity of f guarantees that the point $\mathbf{x}^*$ with the horizontal subgradient (the generalization of a critical point) is a minimum. Because of Theorem 22.4.1, we do not need to distinguish between local and global minimizers.

Finally note that in general there need not be a subgradient at a point $\mathbf{x} \in S$, since Corollary 21.3.15 only establishes the existence of subgradients for points in the interior of S, and we are not assuming that S is open. Indeed, Examples 21.3.9 at $x = 0$ and 21.3.16 provide counterexamples for points on the boundary of S.

22.4.3 Theorem. *Let f be a convex function defined on a convex set S. Then $\mathbf{x}^*$ is a minimizer if and only if the epigraph of f has a horizontal subgradient hyperplane $y = f(\mathbf{x}^*)$ at $\mathbf{x}^*$, meaning that the vector $\mathbf{a}$ in Definition 21.3.12 is the zero vector.*

Proof. First we assume $y = f(\mathbf{x}^*)$ is the equation of a subgradient hyperplane. By the definition of the subgradient hyperplane, all points of the epigraph of f are above or on a subgradient hyperplane equation. Thus for all $\mathbf{x} \in S$, $f(\mathbf{x}) \geq f(\mathbf{x}^*)$, so $\mathbf{x}^*$ is a minimizer.

On the other hand if $\mathbf{x}^*$ is a minimizer, then by definition $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for all $f(\mathbf{x}) \in S$, so $y = f(\mathbf{x}^*)$ is the equation of a subgradient hyperplane. □

For the rest of the discussion of minima, we assume that f is $\mathcal{C}^1$.

22.4.4 Corollary. *Assume f is a $\mathcal{C}^1$ convex function on the convex set $S \subset \mathbb{R}^n$. If f is $\mathcal{C}^1$ in a neighborhood of $\mathbf{x}^*$, then $\mathbf{x}^*$ is a minimizer if and only if*

$$\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \quad \forall \mathbf{x} \in S. \tag{22.4.5}$$

Proof. This follows from Corollary 22.1.4, which tells us that

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \nabla f|_{\mathbf{x}^*}(\mathbf{x} - \mathbf{x}^*), \quad \forall \mathbf{x} \in S.$$

The corollary applies because we require that f be differentiable in a neighborhood of $\mathbf{x}^*$. This hypothesis is only needed when the point $\mathbf{x}^* \in S$ is on the boundary of S. □

Here is a geometric interpretation of this result. The gradient $\nabla f(\mathbf{x}^*)$ points in the direction of steepest ascent of f at $\mathbf{x}^*$, assuming it is non-zero. (22.4.5) says that the angle that $\nabla f(\mathbf{x}^*)$ makes with $\mathbf{x} - \mathbf{x}^*$ is acute, so that when we are moving from $\mathbf{x}^*$ to $\mathbf{x}$ we are going uphill, at least initially. Since by convexity, the segment $[\mathbf{x}, \mathbf{x}^*]$ is in S, this is necessary at a minimum. Since $\mathbf{x}$ is arbitrary, we move uphill in all possible directions, so we are at a minimum. We will have more to say about this in Lecture 23.

So far we have made no assumptions (other than convexity) on the domain S. Recall that if S is open, Corollary 21.3.15 says that subgradients exist at every point of S. Furthermore, when f is differentiable at a point $\mathbf{x}$ in the interior of S, Theorem 22.1.2 tells us that the only subgradient at $\mathbf{x}$ is the gradient $\nabla f(\mathbf{x}^*)$. On the other hand Theorem 13.1.1 tells us that for an arbitrary $\mathcal{C}^1$ function f, at a minimizer $\mathbf{x}^*$ in the interior of the domain, the gradient vanishes. Thus we see the relation of Theorem 22.4.3 to our prior results.

22.4.6 Corollary. *The hypotheses of Corollary 22.4.4 are still in force. Assume that $\mathbf{x}^*$ is a minimizer for f on S. Then another point $\mathbf{x} \in S$ is a minimizer if and only if*

$$\nabla f|_{\mathbf{x}} = \nabla f|_{\mathbf{x}^*} \quad \textit{and} \quad \nabla f|_{\mathbf{x}^*}(\mathbf{x} - \mathbf{x}^*) = 0. \tag{22.4.7}$$

Proof. First we show that all points $\mathbf{x}$ satisfying the two equations are minimizers. Then

$$\begin{aligned} f(\mathbf{x}^*) &\geq f(\mathbf{x}) + \nabla f|_{\mathbf{x}}(\mathbf{x}^* - \mathbf{x}), && \text{by Corollary 22.1.4 ;} \\ &= f(\mathbf{x}) + \nabla f|_{\mathbf{x}^*}(\mathbf{x}^* - \mathbf{x}), && \text{by the first equation;} \\ &= f(\mathbf{x}), && \text{by the second equation.} \end{aligned}$$

Since $\mathbf{x}^*$ is a minimizer, so is $\mathbf{x}$.

Finally, we show that all minimizers satisfy the two equations. The key point is that on the segment $[\mathbf{x}^*, \mathbf{x}]$, f fails to be strictly convex, so by Corollary 22.1.11, the gradient of f is constant. Thus the gradient is constant on the entire locus of minimizers. That shows that the first equation is satisfied.

For the second equation, note that Corollary 22.4.4 applied at $\mathbf{x}^*$ shows that

$$\langle \nabla f|_{\mathbf{x}^*}, \mathbf{x} - \mathbf{x}^* \rangle \geq 0.$$

Applied at $\mathbf{x}$, it gives

$$\langle \nabla f|_{\mathbf{x}}, \mathbf{x}^* - \mathbf{x} \rangle \geq 0.$$

Since we just showed that $\nabla f|_{\mathbf{x}^*} = \nabla f|_{\mathbf{x}}$, these two real numbers are the negative one of the other, so they must both be 0 and we are done. □

For a $\mathcal{C}^1$ convex function on an open set, we have the converse to Theorem 13.1.1:

22.4.8 Theorem. *Let f be a $\mathcal{C}^1$ convex function defined on the open convex set S. Then $\mathbf{x}^*$ is a minimizer if and only if $\nabla f(\mathbf{x}^*) = \mathbf{0}$.*

Proof. If $\nabla f(\mathbf{x}^*) = \mathbf{0}$, then Corollary 22.1.4 says that $\forall \mathbf{x} \in S$, $f(\mathbf{x}) \geq f(\mathbf{x}^*)$, which proves one implication.

On the other hand, if $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for all $\mathbf{x} \in S$, then, as already noted, Theorem 13.1.1 gives the result, without even requiring that f be convex. □

Next we consider maxima. We have a beautiful result on open sets.

22.4.9 Theorem. *Let f be a convex function defined on the open convex set S. Then f has a maximizer $\mathbf{x}^*$ on S if and only if f is the constant function.*

Proof. It is enough to prove the theorem for f restricted to any line L through $\mathbf{x}*$. Because S is open, there is a closed segment $[x_0, x_1]$ containing the maximizer x^* in its interior: so $x_0 < x^* < x_1$. Then the Three Secants Theorem 21.2.2 shows that if x^* is a maximizer, $f(x_0) = f(x^*) = f(x_1)$. Doing this for all lines, we see that the function is constant. □

On more general S, there is no general result. In particular f might have several local maximizers, some of which are not global maximizers, it could have several global maximizers that are isolated from each other. Generally speaking, convexity does not help much with locating maxima. Here is a useful special result.

22.4.10 Theorem. *Let f be a continuous convex function on a bounded polyhedron P. Then f attains its maximum at an extreme point of P.*

Proof. Because f is continuous and P is compact, f attains its maximum somewhere on P by the Weierstrass Theorem 16.2.2. Call such a maximizer $\mathbf{x}^*$. By Theorem 18.7.8, P has a finite number of extreme points, its vertices, and by Minkowski's Theorem 18.7.1 $\mathbf{x}^*$ can be written as a convex combination of them:

$$\mathbf{x}^* = \sum_{i=1}^{m} \lambda_i \mathbf{x}^i, \quad \sum \lambda_i = 1; \quad \forall i, \lambda_i \geq 0.$$

Then by convexity of f, Jensen's Inequality 21.1.11 gives

$$f(\mathbf{x}^*) \leq \sum_{i=1}^{m} \lambda_i f(\mathbf{x}^i).$$

Since $\mathbf{x}^*$ is a maximizer, $f(\mathbf{x}^*) \geq f(\mathbf{x}^i)$ for all i. This, together with Jensen's inequality, implies that at least one of the $\mathbf{x}^i$ is a maximizer. □

22.4.11 Corollary. *A linear function f restricted to a convex polyhedron attains its maximum and minimum at extreme points.*

Proof. Since linear functions are convex, this is immediate for the maximum, and follows for the minimum since the function $-f$ is also convex. □

22.5 A Convex Minimization Example

This example was first mentioned by Fermat, although the exceptional cases were only settled in the 19th century, when it became a textbook favorite: see the textbooks of Bertrand, Serret ([58], §154), and Goursat ([27], §62). A beautiful modern reference with a more complete historical background is Kuhn [36].

22.5.1 Example. Given a triangle with vertices A, B and C, the problem is to find a point M in the plane that minimizes the sum of the distances to the three vertices. To simplify the notation, we place A at the origin, B at the point $(c, 0)$ and C at the point (a, b). We use (x, y) for the coordinates of M. The problem is to minimize

$$f(x,y) = \sqrt{x^2+y^2} + \sqrt{(x-c)^2+y^2} + \sqrt{(x-a)^2+(y-b)^2}$$

The function $f(x, y)$ is convex: Indeed, the distance of a fixed point to a variable point is a convex function (see Example 22.3.4), and the sum of convex functions is convex (see Example 22.3.8).

We set both partials to 0:

$$\begin{aligned} \frac{x}{\sqrt{x^2+y^2}} + \frac{x-c}{\sqrt{(x-c)^2+y^2}} + \frac{x-a}{\sqrt{(x-a)^2+(y-b)^2}} &= 0; \qquad (22.5.2)\\ \frac{y}{\sqrt{x^2+y^2}} + \frac{y}{\sqrt{(x-c)^2+y^2}} + \frac{y-b}{\sqrt{(x-a)^2+(y-b)^2}} &= 0. \end{aligned}$$

We see a potential problem: while f is defined for all points in $\mathbb{R}^2$, it is not differentiable at the vertices of the triangle. So we first attempt to find the minimum away from the vertices of the triangle by setting the gradient to zero. By convexity there can be at most one such solution which is the global minimum. If it is impossible to solve the gradient equations, the minimum must occur at a vertex. Each one of the six terms in (22.5.2) is the $\cos$ or the $\sin$ of three angles θ, ϕ, ψ, so we have

$$\begin{aligned} \cos\theta + \cos\phi + \cos\psi &= 0;\\ \sin\theta + \sin\phi + \sin\psi &= 0. \end{aligned}$$

Move the ψ-terms to the right-hand side, square each equation and add, to get:

$$2+\cos\theta\cos\phi+\sin\theta\sin\phi=1$$

Using a trigonometric addition formula, this gives

$$1+2\cos(\theta-\phi)=0, \text{so } \cos(\theta-\phi)=-\frac{1}{2},$$

so $\theta-\phi$ measures $2\pi/3$ radians. Now this is the angle $\widehat{AMB}$. Repeating this for the other two possibilities, we see that the angles $\widehat{BMC}$ and $\widehat{CMA}$ also measure $2\pi/3$ radians. Then geometrically, M can be constructed by elementary plane geometry as follows. Let A_1 (resp. B_1, C_1) be the third vertex of an equilateral triangle whose other vertices are B (resp. C, A) and C (resp. A, B) on the outside of the triangle ABC. Let C_A (resp.C_B, C_C) be the circles of center A_1 (resp. B_1, resp. C_1) passing through the points B (resp. C, A) and C (resp. A, B). Then M must be the common intersection of these circles. According to Kuhn [36], this point, when it exists, is called the Torricelli point, in honor of the 17th century mathematician who claimed that the intersection is the solution to Fermat's problem. However if an angle of the original triangle is greater than $2\pi/3$ radians the circles do not intersect and there is no solution. In that case the solution occurs at a vertex, indeed the vertex where the angle is at least $2\pi/3$ radians.

If all angles in the triangle are at most $2\pi/3$ radians, then the three circles intersect and that it the unique minimum we are looking for. This example is also used by Levi in [41], §1.3, where a quick physical solution using potential energy solves the problem when the angles of the original triangle are less than $2\pi/3$ radians. The other case is not discussed.

22.6 The Arithmetic-Geometric Mean Inequality

In this section we derive some of the most important inequalities of mathematics from convexity results. We start with the arithmetic-geometric mean inequality and a historical perspective: we turn to the French mathematician Augustin Louis Cauchy (1789-1857), and his influential 1821 textbook *Cours d'Analyse* [15]. For an annotated translation in English, see [11].

In the Note II at the end of [15] (page 291 of the annotated translation), Cauchy first gives an description of inequalities of real numbers, which has became the standard description given for ordered fields: see §14.1. Later he proves what we know as the Cauchy-Schwarz inequality 5.4.6.

Cauchy concludes Note II with the Arithmetic-Geometric Mean Inequality ('un théorème digne de remarque'), a theorem due to Gauss, which we now prove:

22.6.2. Then we extend the result using convexity to the generalized arithmetic-geometric mean inequality. This allows us to prove some of the most famous inequalities in analysis: Hölder's inequality which then yields the Cauchy-Schwarz inequality 5.4.6.

First two definitions:

22.6.1 Definition. The *geometric mean* of n positive real numbers $a_1, \dots, a_n$ is

$$g = \sqrt[n]{a_1 a_2 \dots a_n}$$

and the *arithmetic mean* is

$$a = \frac{a_1 + \dots + a_n}{n}$$

22.6.2 Theorem (The Arithmetic-Geometric Mean Inequality)**.** *The geometric mean of n positive real numbers $a_1, \dots a_n$ is no greater that their arithmetic mean. In other words $g \leq a$, with equality if and only if all the numbers are the same.*

Proof. This is Cauchy's 1821 proof. We give a different proof using the optimization techniques developed in this course in §32.3. We raise both sides of the equation to the n-th power and prove the equivalent statement

$$a_1 a_2 \dots a_n \leq \Bigl(\frac{a_1 + \dots + a_n}{n}\Bigr)^n. \tag{22.6.3}$$

For $n = 2$, writing the two numbers as a and b, obviously

$$ab = \Bigl(\frac{a+b}{2}\Bigr)^2 - \Bigl(\frac{a-b}{2}\Bigr)^2 < \Bigl(\frac{a+b}{2}\Bigr)^2,$$

as long as $a \neq b$. We continue by induction on m in order to establish the result for any n that is of the form 2^m. For example, when $n = 4$,

$$abcd < \Bigl(\frac{a+b}{2}\Bigr)^2 \Bigl(\frac{c+d}{2}\Bigr)^2 < \Bigl(\frac{a+b+c+d}{2}\Bigr)^4,$$

where the right most inequality comes from substituting $a + b$ for a and $c + d$ for b in the previous case. So the result is proved for all powers of 2.

To handle the case where n is not a power of 2, we let 2^m be the smallest power of 2 greater than n, and we extend the number of terms to 2^m by letting the last $2^m - n$ terms be equal to the same number k, where

$$k = \frac{a_1 + \dots + a_n}{n}, \tag{22.6.4}$$

in other words, k is the arithmetic mean of the first n terms. Using the result for 2^m we get

$$a_1 a_2 \dots a_n k^{2^m - n} < \Big(\frac{a_1 + \dots + a_n + (2^m - n)k}{2^m}\Big)^{2^m}. \tag{22.6.5}$$

Then substituting in the value for k from (22.6.4), we see that the right-hand side of (22.6.5) simplifies to k^{2m}, so that we get

$$a_1 a_2 \dots a_n k^{2^m - n} < k^{2m}.$$

Divide by $k^{2^m - n}$ to get:

$$a_1 a_2 \dots a_n < k^n,$$

which by (22.6.4) gives the result. □

We now generalize this result by using the convexity of $-\ln[x]$.

22.6.6 Theorem. *Let a and b be positive numbers with $a \neq b$. Then for any λ, $0 < \lambda < 1$,*

$$a^\lambda b^{1-\lambda} < \lambda a + (1 - \lambda) b. \tag{22.6.7}$$

Proof. Assume $a < b$. Write down the definition of the convexity of $-\ln[x]$ on the segment $[a, b]$ and take the exponential on both sides. □

Note that the case $\lambda = 1/2$ gives us a second proof of Cauchy's Theorem 22.6.2 in the case $n = 2$. More generally,

22.6.8 Theorem. *For any n-tuple of positive numbers a_1, ..., a_n, and any set of λ_i, $1 \leq i \leq n$, such that $\lambda_i \geq 0$ and $\sum \lambda_i = 1$,*

$$\prod_{i=1}^n x_i^{\lambda_i} \leq \sum_{i=1}^n \lambda_i x_i. \tag{22.6.9}$$

Proof. Just use Jensen's Inequality 21.1.11. □

22.6.10 Corollary. *Now let $\lambda_i = 1/n$ for all i. Then*

$$\Big(\prod_{i=1}^n x_i\Big)^{\frac{1}{n}} \leq \frac{\sum_{i=1}^n x_i}{n},$$

so we have recovered the general arithmetic-geometric mean inequality for all n.

We now use (22.6.7) to prove:

22.6.11 Theorem (Hölder inequality)**.** *For* $p > 1$*,* $q > 1$*,* $1/p + 1/q = 1$*, and* $\mathbf{x}$ *and* $\mathbf{y}$ *in* $\mathbb{R}^n$*,*

$$\sum_{i=1}^{n} |x_i||y_i| \le \Bigl(\sum_{i=1}^{n} |x_i|^p\Bigr)^{\frac{1}{p}} \Bigl(\sum_{i=1}^{n} |y_i|^q\Bigr)^{\frac{1}{q}}. \tag{22.6.12}$$

Proof. Let

$$a = \frac{|x_i|^p}{S_{\mathbf{x}}}, b = \frac{|y_i|^q}{S_{\mathbf{y}}}$$

where

$$S_{\mathbf{x}} = \Bigl(\sum_{j=1}^{n} |x_j|^p\Bigr), S_{\mathbf{y}} = \Bigl(\sum_{j=1}^{n} |y_j|^q\Bigr)$$

so that the general arithmetic-geometric mean inequality (22.6.7) applied to a and b gives

$$\Bigl(\frac{|x_i|^p}{S_{\mathbf{x}}}\Bigr)^{\frac{1}{p}} \Bigl(\frac{|y_i|^q}{S_{\mathbf{y}}}\Bigr)^{\frac{1}{q}} \le \frac{|x_i|^p}{pS_{\mathbf{x}}} + \frac{|y_i|^q}{qS_{\mathbf{y}}}.$$

Then sum all the inequalities over i:

$$\sum_{i=1}^{n} \Bigl(\frac{|x_i|^p}{S_{\mathbf{x}}}\Bigr)^{\frac{1}{p}} \Bigl(\frac{|y_i|^q}{S_{\mathbf{y}}}\Bigr)^{\frac{1}{q}} \le \sum_{i=1}^{n} \frac{|x_i|^p}{pS_{\mathbf{x}}} + \sum_{i=1}^{n} \frac{|y_i|^q}{qS_{\mathbf{y}}} = \frac{1}{p} + \frac{1}{q} = 1,$$

and clear the denominators to get the result. □

22.6.13 Corollary. *When* $p = q = 2$*, the Hölder inequality reduces to the Cauchy-Schwarz Inequality 5.4.6.*

Finally, we could use the Hölder inequality to prove another famous inequality, the Minkowski inequality. See, for example, Carothers [14], p.44, for how these inequalities are used to provide norms on infinite dimension vector spaces of functions.

22.6.14 Exercise. Let a and b be positive numbers with $a < b$. Create two infinite sequences $\{a_n\}$ and $\{g_n\}$ by letting

$$a_1 = \frac{a+b}{2} \text{ and } g_1 = \sqrt{ab}, \tag{22.6.15}$$

and

$$a_n = \frac{a_{n-1} + g_{n-1}}{2} \text{ and } g_n = \sqrt{a_{n-1}g_{n-1}}, \text{ for } n \ge 2.$$

Prove by induction that

$$a_n > a_{n+1} > \cdots > g_{n+1} > g_n, \text{ for all } n.$$

Prove that the two sequences $\{a_n\}$ and $\{g_n\}$ converge. State the theorem you use.

Show that both sequences converge to the same value, called the *arithmetic-geometric mean* of a and b.

22.6.16 Exercise. Show that the geometric mean function of $\mathbf{x} = (x_1, x_2, \dots, x_n)$:

$$f(\mathbf{x}) = \big(\prod_{i=1}^{n} x_i\big)^{1/n} \tag{22.6.17}$$

is concave on the open positive quadrant $\mathbb{R}^n_{++}$ by computing the Hessian matrix F of $-f$ (notice the minus sign) and showing it is positive semidefinite.

22.6.18 Exercise. Show that the set of points in the quadrant $x_1 > 0$, $x_2 > 0$ in $\mathbb{R}^2$ given by $x_1 x_2 \geq 1$ is convex. Hint: this is sometimes called the “hyperbolic” set. Why?

Generalize to $\mathbb{R}^n$: show that the set of points $\mathbf{x}$ in the open first quadrant $\mathbb{R}^n_{++}$ such that

$$\prod_{i=1}^{n} x_i \geq 1$$

is convex.

Hint: Apply the definition of convexity directly, using (22.6.7).

Lecture 23

Convex Optimization with Differentiability

In Lecture 22, we proved some of the basic theorems for the optimization of convex function. Here we continue this study using the techniques we introduced in the last five chapters, namely Lagrange multipliers and Kuhn-Tucker conditions. Indeed we show how the results obtained for general nonlinear optimization in Lectures 28, 29, 31 and 32 can be improved and simplified using convexity hypotheses.

We first define the standard problem and review the key results from Lecture 22. Then, in §23.2, we quickly cover the case where there are only equality constraints, and compare with the general Lagrange multiplier set-up. Then we cover a simple but important example: convex optimization over the positive quadrant: §23.3.

The first important new theorem of the lecture is Theorem 23.4.1, which gives a sufficient condition for a feasible solution to be a minimizer: the KKT conditions without any contraint qualification.

The classic example in §23.5 shows that a constraint qualification is needed in some situations. A mild constraint Qualification, called the Slater condition (§23.7), allows the formulation of a necessary condition in terms of Lagrange multipliers for a point $\mathbf{x}^*$ to be a minimizer. We need to establish the convexity of the value function in §23.6 to derive the necessary condition in §23.7. The lecture ends with a geometric example: §23.8.

23.1 The Standard Problem in Convex Optimization

23.1.1 Definition. The standard convex minimization problem is

Minimize $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, subject to $A\mathbf{x} = \mathbf{b}$, and $g_k(\mathbf{x}) \leq 0$, $1 \leq k \leq p$.

Here $f(\mathbf{x})$ and the $g_k(\mathbf{x})$ are convex functions, and A is a $m \times n$ matrix of maximal rank m. As usual we write $\mathbf{g}(\mathbf{x})$ for the vector of $g_k(\mathbf{x})$.

The feasible set F is an intersection of convex sets, and is therefore convex, unless empty. We write f^* for the minimal value of $f(\mathbf{x})$ on F. Note that we allow the value $-\infty$, which means that f is unbounded below on F. We write $\mathbf{x}^*$ for any feasible vector that minimizes, namely, such that $f(\mathbf{x}^*) = f^*$. If $f^* = -\infty$, then there is no minimizer $\mathbf{x}^*$.

Recall that from Theorem 22.4.1, any local minimizer $\mathbf{x}^*$ for the standard problem is a global minimizer. Now assume that f is $\mathcal{C}^1$. Then Corollary 22.4.4 says that a point $\mathbf{x}^*$ is a minimizer for Problem 23.1.1 if and only if

$$\nabla f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) \geq 0, \quad \text{for all feasible } \mathbf{x}. \tag{23.1.2}$$

On the left-hand side of this expression, we are multiplying the row vector ∇f with the column vector $\mathbf{x} - \mathbf{x}^*$.

Furthermore, if there is an open ball in $\mathbb{R}^n$ centered at $\mathbf{x}^*$ contained in the feasible set F, then $\mathbf{x}^*$ is a minimizer if and only if $\nabla f(\mathbf{x}^*) = \mathbf{0}$, since then one can move in all possible directions in $\mathbb{R}^n$ at $\mathbf{x}^*$. This always happens if there are no equality constraints, and if none of the inequality constraints is active at $\mathbf{x}^*$. It can happen in other situations, and when it does, we have a minimizer.

23.2 Convex Optimization with Equality Constraints Only

Assume that f is $\mathcal{C}^1$, and that there are no inequality constraints. So the only constraints are the matrix constraints $A\mathbf{x} = \mathbf{b}$. We make the harmless Rank Assumption 19.6.2 that the $m \times n$ matrix A has rank m, where $m \leq n$.

For $\mathbf{x}^*$ to be a minimizer, we need (23.1.2) to be satisfied.

Since both $\mathbf{x}$ and $\mathbf{x}^*$ satisfy the constraint $A\mathbf{x} = \mathbf{b}$, their difference $\mathbf{v} = \mathbf{x} - \mathbf{x}^*$ is in the nullspace $\mathcal{N}(A)$ of A: $A\mathbf{v} = A\mathbf{x} - A\mathbf{x}^* = \mathbf{b} - \mathbf{b} = \mathbf{0}$. Now (23.1.2) says that $\nabla f(\mathbf{x}^*) \geq 0$ on the entire linear space $\mathcal{N}(A)$. But this can only happen if $\nabla f(\mathbf{x}^*) = 0$ on $\mathcal{N}(A)$, in other words $\nabla f(\mathbf{x}^*) \perp \mathcal{N}(A)$. By the Four Subspaces Theorem 7.2.3 , this implies that $\nabla f(\mathbf{x}^*)$ is in $\mathcal{R}(A^T)$, which just says that $\nabla f(\mathbf{x}^*)$ can be written as a linear combination of the rows of A, in other words:

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{n} \lambda_i \mathbf{a}^i = \mathbf{0},$$

where $\mathbf{a}^i$ denotes the i-th row of A.

Thus we have recovered a 'convex' version of the Lagrange Multiplier Theorem 28.3.9:

23.2.1 Theorem. *Let $\mathbf{x}^*$ be a minimizer for Problem 23.1.1 with only equality constraints. There are unique numbers $\lambda_1^*, \dots, \lambda_m^*$, such that*

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{n} \lambda_i^* \mathbf{a}^i = \mathbf{0}, \tag{23.2.2}$$

where $\mathbf{a}^i$ denotes the i-th row of A.

This theorem improves the ordinary Lagrange Theorem, because we do not need to check a second order condition: indeeed, the objective function is convex. We do not need to assume regularity for the constraints, because the constraints are linear.

Furthermore, if $f(\mathbf{x})$ is linear, so we write it $\mathbf{c}^T\mathbf{x}$, then (23.2.2) becomes $\mathbf{c} = A^T\lambda^*$, so we recover an analog of the Duality Theorem 25.5.1 of linear optimization, with the λ playing the role of the dual variables $\mathbf{y}$. If we add the inequality constraints studied in the next section, we recover duality for the Asymmetric Problem 25.1.5 of linear optimization.

23.3 Example: Convex Optimization over the Positive Quadrant

In this section we assume the optimization problem is

$$\text{Minimize } f(\mathbf{x}) \text{ subject to } \mathbf{x} \succeq \mathbf{0}. \tag{23.3.1}$$

23.3.2 Theorem. *A necessary condition for a feasible $\mathbf{x}^*$ to be a solution to Problem 23.3.1 is that*

$$\nabla f(\mathbf{x}^*) \succeq \mathbf{0} \quad \textit{and} \quad \nabla f(\mathbf{x}^*)\mathbf{x}^* = 0.$$

Proof. By (23.1.2), for $\mathbf{x}^*$ to be a minimizer, we need

$$\nabla f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) \geq 0, \quad \text{for all } \mathbf{x} \succeq \mathbf{0}. \tag{23.3.3}$$

The expression $\nabla f(\mathbf{x}^*)\mathbf{x}$ is a linear function of $\mathbf{x}$ with coefficients $\nabla f(\mathbf{x}^*)$. It goes to $-\infty$ for suitable $\mathbf{x} \succeq \mathbf{0}$ unless the gradient is non-negative: $\nabla f(\mathbf{x}^*) \succeq \mathbf{0}$. Indeed, assume the i-th coordinate of the gradient is negative: then take $\mathbf{x}$ to be the vector with zeroes in all other positions So we assume this is the case, since we want a finite minimum.

When $\mathbf{x} = \mathbf{0}$, (23.3.3) becomes

$$\nabla f(\mathbf{x}^*)\mathbf{x}^* \leq 0.$$

Since both $\mathbf{x}^*$ and $\nabla f(\mathbf{x}^*)$ are non-negative, this can only occur if

$$\nabla f(\mathbf{x}^*)\mathbf{x}^* = 0,$$

which is complementary slackness: for each index j: either the j-th coordinate of $\nabla f(\mathbf{x}^*)$ or that of $\mathbf{x}^*$ is 0. □

Theorem 23.4.1 shows this is also a sufficient condition. This is worked out in Example 23.4.3.

23.4 Sufficient Conditions for Minimality

We now treat Problem 23.1.1 in full generality.

23.4.1 Theorem. *Sufficient conditions that a feasible $\mathbf{x}^*$ be a solution to Problem 23.1.1 are that there exist a m-vector λ and a non-negative p-vector μ such that*

$$\nabla f(\mathbf{x}^*) + \lambda^T A + \mu^T \nabla \mathbf{g}(\mathbf{x}^*) = \mathbf{0} \quad \textit{and} \quad \langle \mu, \mathbf{g}(\mathbf{x}^*)\rangle = 0. \tag{23.4.2}$$

These are the usual KKT conditions. There are no constraint qualifications.

Proof. For any feasible $\mathbf{x}$ we compute:

$$\begin{aligned}
f(\mathbf{x}) - f(\mathbf{x}^*) &\geq \langle \nabla f(x^*), \mathbf{x} - \mathbf{x}^*\rangle, && f \text{ convex: Corollary 22.1.4;}\\
&= \langle -\lambda^T A - \mu^T \nabla \mathbf{g}(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle, && \text{using (23.4.2);}\\
&= -\langle \lambda^T A, \mathbf{x} - \mathbf{x}^*\rangle - \langle \mu^T \nabla \mathbf{g}(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle, && \text{by linearity;}\\
&= -\langle \lambda, A(\mathbf{x} - \mathbf{x}^*)\rangle - \langle \mu, \nabla \mathbf{g}(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*)\rangle, && \text{by selfadjointness;}\\
&= -\langle \mu, \nabla \mathbf{g}(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*)\rangle, && \mathbf{x} - \mathbf{x}^* \in \mathcal{N}(A);\\
&\geq -\langle \mu, \mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{x}^*)\rangle, && g \text{ convex: Corollary 22.1.4;}\\
&= -\langle \mu, \mathbf{g}(\mathbf{x})\rangle, && \text{by complementary slackness;}\\
&\geq 0, && \text{because } \mu \succeq \mathbf{0} \text{ and } \mathbf{g}(\mathbf{x}) \preceq \mathbf{0}.
\end{aligned}$$

Thus $\mathbf{x}^*$ is a minimizer. □

We will get a necessary condition in §23.7, after an example showing that a constraint qualification is needed (§23.5), and a discussion of the value function (§23.6), used in our proof of the necessary condition.

23.4.3 Example. We now finish the example of §23.3 using Theorem 23.4.1, by showing that the conditions of Theorem 23.3.2 are the KKT conditions of Theorem 23.4.1. Write the n positivity constraints as $g_i(\mathbf{x}) = -x_i \leq 0$, to follow our convention. Then $\nabla g_i = -\mathbf{e}_i$, where $\mathbf{e}_i$ is the i-th unit vector. So if we choose $\mu = \nabla f(\mathbf{x}^*)$, from our previous work we see that $\mu \succeq \mathbf{0}$ and complementary slackness holds. The Lagrange equations follow, because

$$\nabla f(\mathbf{x}^*) + \mu^T \nabla g(\mathbf{x}^*) = \nabla f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*) \nabla g(\mathbf{x}^*) = \nabla f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*) = \mathbf{0}.$$

Thus we have a necessary and sufficent condition for minimality.

23.5 An Example

We consider the problem in $\mathbb{R}^2$ with objective function $f(x_1, x_2) = x_1$ and two inequality constraints:

$$g_1(x_1, x_2) = x_2 - t \leq 0 \text{ , } g_2(x_1, x_2) = x_1^2 - x_2 \leq 0$$

Note the parameter t. If t is negative, then the feasible set is empty, so we assume $t \geq 0$. When $t = 0$, the feasible set is just one point: the point $(0, 0)$. For each value of $t \geq 0$, we set this up as a KKT problem. A quick graph shows that the minimum occurs at the left intersection of the two bounding curves:

$$x_2 = t, x_1^2 = x_2, \text{ so } x_1 = -\sqrt{t} \text{ and } x_2 = t$$

We solve this using KKT. The Lagrangian is $\mathcal{L}(x_1, x_2, \mu_1, \mu_2) = x_1 + \mu_1(x_2 - t) + \mu_2(x_1^2 - x_2)$, so the Lagrange equations are

$$1 + 2\mu_2 x_1 = 0$$
$$\mu_1 - \mu_2 = 0$$

so $\mu_1 = \mu_2 = \frac{1}{2\sqrt{t}}$. Thus, if $\sqrt{t} > 0$, the μ_i are non-negative as required, and the problem can be solved using multipliers.

Note however that when $t = 0$, the feasible set is reduced to one point: $(0, 0)$, and the method breaks down: the minimum cannot be found through the Lagrange method.

This shows that even when one deals with a convex optimization problem, one cannot always find the solution using Lagrange multipliers. As we will see in §23.7, there is a constraint qualification that tells us when one can: the Slater condition.

As a final remark, note that we can compute the minimum value function $w(t)$ (see Definition 23.6.2) in terms of t: it is $w(t) = -\sqrt{t}$. It is convex as required by Theorem 23.6.3. The difficulty of the value $t = 0$ is that it is not an interior point of the domain of the value function. Thus some of the standard methods of dealing with convex functions do not apply. If you refer back to Lecture 21, you will see that many of the results there are only valid for points in an open set of the domain of the convex function. This excludes the value $t = 0$.

23.6 The Value Function for Convex Minimization

Now we look at the 'perturbed' problem, where f, g and A are as in (23.1.1), without requiring that f and the g_k be differentiable: they just need to be convex.

23.6.1 Definition (The Standard Perturbed Problem)**.**

$$\text{Minimize } f(\mathbf{x}), \text{ subject to } A\mathbf{x} = \mathbf{b} \text{ and } \mathbf{g}(\mathbf{x}) \preceq \mathbf{e}.$$

where

- $f(\mathbf{x})$ is a convex function of n variables;
- A is a constant $m \times n$ matrix of rank m, and $\mathbf{b}$ a variable m-vector;
- $\mathbf{g}(\mathbf{x})$ is a collection of p convex functions, and $\mathbf{e}$ a variable p-vector.

The following theorem generalizes what we did in the linear case in §25.8. It shows the power of the convexity hypotheses. First a definition:

23.6.2 Definition. Let $\mathbf{d}$ be the compound $(m + p)$-vector $(\mathbf{b}, \mathbf{e})$. For any fixed value of $\mathbf{d}$ such that the feasible set of Problem 23.6.1 is nonempty, and f has a finite minimum, let $w(\mathbf{d})$ denote that minimum value. As $\mathbf{d}$ varies we get a function $w(\mathbf{d})$ of $\mathbf{d}$, called the *minimum value function*, or just the *value function* for the problem.

23.6.3 Theorem. *The set of vectors* $\mathbf{d}$ *such that (23.6.1) has a finite solution is either empty or is a convex subset W of $\mathbb{R}^{m+p}$, and in the latter case the value function $w(\mathbf{d})$ is a convex function on W.*

Proof. If W is empty, there is nothing to prove. So assume it is non-empty. If it is just one point, again there is nothing to prove. So pick any two points $\mathbf{d}^0$ and $\mathbf{d}^1$ in W. To establish the convexity of W, we need to show there is a finite solution at every point of the segment $\lambda\mathbf{d}^0 + (1 - \lambda)\mathbf{d}^1$, for $0 < \lambda < 1$. Since there is a solution at $\mathbf{d}^0 = \left[\mathbf{b}^0, \mathbf{e}^0\right]$, there is a minimizer $\mathbf{x}^0$ such that

$$A\mathbf{x}^0 = \mathbf{b}^0, \text{ and } \mathbf{g}(\mathbf{x}^0) \leq \mathbf{e}^0.$$

Similarly at $\mathbf{d}^1 = \left[\mathbf{b}^1, \mathbf{e}^1\right]$, there is a minimizer $\mathbf{x}^1$ such that

$$A\mathbf{x}^1 = \mathbf{b}^1, \text{ and } \mathbf{g}(\mathbf{x}^1) \leq \mathbf{e}^1.$$

We claim $\lambda\mathbf{x}^0 + (1 - \lambda)\mathbf{x}^1$ is feasible for the value $\lambda\mathbf{d}^0 + (1 - \lambda)\mathbf{d}^1$. First

$$A(\lambda\mathbf{x}^0 + (1 - \lambda)\mathbf{x}^1) = \lambda A\mathbf{x}^0 + (1 - \lambda)A\mathbf{x}^1 = \lambda\mathbf{b}^0 + (1 - \lambda)\mathbf{b}^1$$

by linearity. Next, by convexity of each of the g_i, we get

$$\mathbf{g}(\lambda\mathbf{x}^0 + (1-\lambda)\mathbf{x}^1) \le \lambda\mathbf{g}(\mathbf{x}^0) + (1-\lambda)\mathbf{g}(\mathbf{x}^1) \le \lambda\mathbf{e}^0 + (1-\lambda)\mathbf{e}^1.$$

Putting these two results together, we see that $\lambda\mathbf{x}^0 + (1-\lambda)\mathbf{x}^1$ is feasible, for each value of λ. Finally we have to rule out the possibilty that f goes to $-\infty$ along the segment. This is ruled out by Lemma 21.3.8 applied to f restricted to the segment $[\mathbf{x}^0, \mathbf{x}^1]$. This concludes the proof that W is convex.

Next we prove that the value function is convex. As required, it is defined on a convex set, namely W. So we need to show, for any $\mathbf{d}^0$ and $\mathbf{d}^1$ in W, that

$$w(\lambda\mathbf{d}^0 + (1-\lambda)\mathbf{d}^1) \le \lambda w(\mathbf{d}^0) + (1-\lambda)w(\mathbf{d}^1), \qquad \text{for } 0 < \lambda < 1.$$

23.6.4 Definition. For each $\lambda \in [0,1]$, denote by $F(\lambda)$ the set of feasible $\mathbf{x}$ for $\mathbf{d} = \lambda\mathbf{d}^0 + (1-\lambda)\mathbf{d}^1$.

By definition of w,

$$w(\lambda\mathbf{d}^0 + (1-\lambda)\mathbf{d}^1) = \inf_{\mathbf{x}\in F(\lambda)} f(\mathbf{x}) \le \inf_{\mathbf{x}\in \lambda F(0)+(1-\lambda)F(1)} f(\mathbf{x}).$$

The last inequality follows because the set of $\mathbf{x} = \lambda\mathbf{x}^0 + (1-\lambda)\mathbf{x}^1$, for any $\mathbf{x}^0 \in F(0)$ and $\mathbf{x}^1 \in F(1)$, is feasible at λ by the convexity of W, and so is contained in $F(\lambda)$. Since f is convex:

$$f(\lambda\mathbf{x}^0 + (1-\lambda)\mathbf{x}^1) \le \lambda f(\mathbf{x}^0) + (1-\lambda)f(\mathbf{x}^1),$$

so that

$$\inf_{\mathbf{x}\in \lambda F(0)+(1-\lambda)F(1)} f(\mathbf{x}) \le \lambda \inf_{\mathbf{x}\in F(0)} f(\mathbf{x}) + (1-\lambda) \inf_{\mathbf{x}\in F(1)} f(\mathbf{x}).$$

The right-hand side of this inequality is $\lambda w(\mathbf{d}^0) + (1-\lambda)w(\mathbf{d}^1)$, so assembling the chain of inequalities, we have shown that w is convex. □

Finally we want to establish an Envelope Theorem, namely a theorem that computes the gradient of $w(\mathbf{d})$ when it exists. As is shown at the end of the next section, when the minimizer can be determined using Lagrange multipliers, the gradient of $w(\mathbf{d})$ is the collection of Lagrange multipliers.

23.7 The Slater Constraint Qualification and a Necessary Condition for Convex Optimization

23.7.1 Definition. We say that the optimization problem 23.1.1 satisfies the *Slater condition* if there is a feasible $\mathbf{x}^0$ such that $g_k(\mathbf{x}^0) < 0$ for each inequality constraint g_k, $1 \leq k \leq p$.

Thus $\mathbf{x}^0$ is in the interior of the constraint set defined by inequalities. Notice that the example in §23.5 fails this condition when $t = 0$: the feasible set contains only one point, $\mathbf{x}^0 = (0,0)$, and $g_1(0,0) = 0$ and $g_2(0,0) = 0$.

23.7.2 Theorem. *Assume that the Slater condition holds at a feasible point $\mathbf{x}^*$ for the convex optimization problem 23.1.1. Then a necessary condition for $\mathbf{x}^*$ to be a minimizer is that it satisfy the KKT conditions of Theorem 23.4.1.*

Proof. For simplicity we assume there are no equality constraints. If the Slater condition holds, we show that there are nonnegative constants $\mu_1, \dots, \mu_p$ such that the Legrangian

$$\mathcal{L}(\mathbf{x}) = f(\mathbf{x}) + \sum_{k=1}^{p} \mu_k g_k(\mathbf{x})$$

has a minimum at $\mathbf{x}^*$, and such that complementary slackness holds at $\mathbf{x}^*$: $\mu^T \mathbf{g}(\mathbf{x}^*) = 0$. Differentiating the Lagrangian at the minimum $\mathbf{x}^*$ gives the required result.

By Theorem 23.6.3 the value function $w(\mathbf{z})$ is convex, and defined on a convex set Ω in $\mathbb{R}^p$. Consider its epigraph (see §21.3) in $\mathbb{R}^{p+1}$, where the last coordinate corresponds to the value of the function. The value $\mathbf{z} = (z_1, \dots, z_p)$ corresponds to the feasible set given by the constraints $g_k(\mathbf{x}) - z_k \leq 0$. Thus the original problem corresponds to the value $\mathbf{z} = \mathbf{0}$. By hypothesis $w(\mathbf{0})$ is defined. The Slater condition tells us that $\mathbf{0}$ is an interior point of the domain of $w(\mathbf{z})$, so that the supporting hyperplane H to the epigraph at the point $(w(\mathbf{0}), \mathbf{0})$ in $\mathbb{R}^{p+1}$ is not vertical, and H not being vertical means that the coefficient of z_{p+1} in the equation of H is not 0. So we can normalize it to 1, and write the equation of H as:

$$\sum_{k=1}^{p} \mu_i z_i + z_{p+1} = w(\mathbf{0}) \tag{23.7.3}$$

For any $\mathbf{z}$ in the domain of w, since $w(\mathbf{z})$ is above the supporting hyperplane H (see 23.7.3), we get

$$w(\mathbf{z}) + \mu^T \mathbf{z} \geq w(\mathbf{0}). \tag{23.7.4}$$

If $\mathbf{b} \preceq \mathbf{c}$, then $w(\mathbf{b}) \geq w(\mathbf{c})$, since the minimization at $\mathbf{c}$ takes place over a larger set than at $\mathbf{b}$. As we 'relax the constraints', taking all the z_k positive, w decreases, so (23.7.4) tells us

$$\mu^T \mathbf{z} \geq w(\mathbf{0}) - w(\mathbf{z}) \geq 0. \tag{23.7.5}$$

This forces $\mu \succeq 0$, since we can do this one coordinate of $\mathbf{z}$ at a time.

By hypothesis, $\mathbf{x}^*$ is a minimizer of f at $\mathbf{z} = \mathbf{0}$. Since $\mathbf{g}(\mathbf{x}^*) \preceq \mathbf{0}$, $w(\mathbf{g}(\mathbf{x}^*)) \geq w(\mathbf{0})$. But then the existence of $\mathbf{x}^*$ shows that $w(\mathbf{g}(\mathbf{x}^*)) = w(\mathbf{0})$.

Now let $\mathbf{x}$ be any feasible vector for $\mathbf{z} = \mathbf{0}$. Then using (23.7.4)

$$f(\mathbf{x}) \geq w(\mathbf{0}) = w(\mathbf{g}(\mathbf{x})) \geq w(\mathbf{0}) - \mu^T \mathbf{g}(\mathbf{x}). \tag{23.7.6}$$

Apply this to $\mathbf{x} = \mathbf{x}^*$ to get:

$$w(\mathbf{0}) \geq w(\mathbf{0}) - \mu^T \mathbf{g}(\mathbf{x}^*), \quad \text{or} \quad \mu^T \mathbf{g}(\mathbf{x}^*) \geq 0.$$

Since μ is a non-negative vector, and $\mathbf{g}(\mathbf{x})$ a non-positive vector, this implies complementary slackness: $\mu^T \mathbf{g}(\mathbf{x}^*) = 0$.

Then (23.7.6) implies, together with complementary slackness, that

$$f(\mathbf{x}) + \mu^T \mathbf{g}(\mathbf{x}) \geq w(\mathbf{0}) = f(\mathbf{x}^*) = f(\mathbf{x}^*) + \mu^T \mathbf{g}(\mathbf{x}^*),$$

so that $\mathbf{x}^*$ does minimize the Lagrangian, and we are done. □

Notice how the Lagrange multipliers μ were generated from the coefficients of the supporting hyperplane of the epigraph of the value function at the point associated to the value $\mathbf{0}$. Also note that the Slater condition was only used once: to show that this hyperplane is not vertical.

This establishes:

23.7.7 Theorem (Envelope Theorem)**.** *If the value function $w(\mathbf{z})$ of our standard problem 23.1.1 without equality constraints is differentiable at a point $\mathbf{z}$ in the interior of its domain, then by the proof of Theorem 23.7.2, unique Lagrange multipliers μ_k can be found, and*

$$\nabla w(\mathbf{z}) = \mu, \textit{ namely} \quad \frac{\partial w}{\partial z_k}(\mathbf{z}) = \mu_k.$$

It would be easy to add equality constraints. You should think carefully about how this theorem generalizes Theorem 25.8.2.

23.8 A Geometric Example

In Example 18.3.24 we found the smallest ball containing a regular simplex. We now take an arbitrary n-simplex in $\mathbb{R}^n$, and ask for the ball of smallest radius containing it.

Let $\mathbf{a}^i$, $0 \leq i \leq n$ be the vertices of the simplex. They are affinely independent. We use as variables the coordinates $(x_1, \dots, x_n)$ of the center of the ball, and its radius r. The objective function is

$$f(\mathbf{x}, r) = r,$$

since we wish to minimize the radius. The constraints, stating that the distance of the center to each of the $n+1$ points is bounded by the radius, are

$$g_i(\mathbf{x}, r) = \|\mathbf{x} - \mathbf{a}^i\| - r \leq 0, \text{ for } 0 \leq i \leq n. \tag{23.8.1}$$

By Example 22.3.4, the constraints are all convex, and the objective function is clearly convex (see Example 22.3.3), so that we are dealing with a convex problem (see Definition 23.1.1).

We know that a solution exists, since we can find a big disk containing all the points. We would like to solve the problem using Kuhn-Tucker. First we check that the Slater condition is satisfied. This is the case, since each the feasible set for each constraint is a right "circular cone", which all intersect as long as the radius r is large enough. Indeed the perturbed problem 23.6.1 with constraints $g_i(\mathbf{x}, r) \leq \mathbf{e}$ satisfies the Slater condition as long as $e_j > -r$.

So we can solve the problem using Kuhn-Tucker: we now want to solve the problem explicitly, which means writing $\mathbf{x}^*$, r^* and μ^* is terms of the data, namely the vectors $\mathbf{a}^0$, …, $\mathbf{a}^n$. Note that by Theorem 22.4.1, once we have found one minimizer, the minimum value of the objective function, namely the radius, is determined: any local minimum is a global minimum.

The Lagrangian is written

$$\mathcal{L}(\mathbf{x}, r, \mu) = r + \sum_{i=0}^{n} \mu_i \big(\|\mathbf{x} - \mathbf{a}^i\| - r\big)$$

We are looking for a solution $(\mathbf{x}^*, r^*, \mu^*)$ of the system obtained by setting the partials of $\mathcal{L}$ to 0:

$$\sum_{i}^{m} \frac{\mu_i}{\|\mathbf{x} - \mathbf{a}^i\|}(x_j - a^i_j) = 0, \qquad \text{for the partial w.r.t. } x_j; \tag{23.8.2}$$

$$1 = \sum_{i}^{m} \mu_i, \qquad \text{for the partial w.r.t. } r. \tag{23.8.3}$$

Additionally the constraints (23.8.1) must be satisfied at the solution:

$$g_i(\mathbf{x}^*, r^*) = \|\mathbf{x}^* - \mathbf{a}^i\| - r^* \leq 0,$$

the multipliers μ^* must be non-negative, and complementary slackness must hold: $\mu^*\mathbf{g}(\mathbf{x}^*) = 0$.

Equations 23.8.2 say, using (23.8.3):

$$x_j = \sum_{i=0}^{n} \mu_i a^i_j \,,\, 1 \le j \le n \,, \text{so} \quad \mathbf{x} = \sum_{i=0}^{n} \mu_i \mathbf{a}^i. \tag{23.8.4}$$

Since the μ are nonnegative, $\mathbf{x}$ is a convex linear combination of the points $\mathbf{a}^i$ for which $\mu_i > 0$.

We reorder the points so that the first $k+1$ constraints are active.

By complementary slackness, μ_i is 0 unless the corresponding constraint is active, meaning that the point $\mathbf{a}^i$ is on the boundary of the ball: $\|\mathbf{x} - \mathbf{a}^i\| = r$ for $0 \le i \le k$. Then the system of equations we need to solve is

$$\begin{aligned} \sum_{i=0}^{k} \mu_i &= 1; \\ \|\mathbf{x} - \mathbf{a}^i\| &= r, \quad \text{for } 0 \le i \le k; \\ \mathbf{x} &= \sum_{i=0}^{k} \mu_i \mathbf{a}^i. \end{aligned}$$

with the remaining points feasible.

23.8.5 Lemma. *At a solution* $(\mathbf{x}^*, r^*, \mu^*)$ *we have*

$$\|\mathbf{x}^*\|^2 + r^{*2} = \sum_{i=0}^{k} \mu^*_i \|\mathbf{a}^i\|^2.$$

Proof. Indeed, for the active constraints, the only ones with positive μ, we have

$$r^2 = \|\mathbf{x} - \mathbf{a}^i\|^2 = \langle \mathbf{x}, \mathbf{x}\rangle - 2\langle \mathbf{x}, \mathbf{a}^i\rangle + \langle \mathbf{a}^i, \mathbf{a}^i\rangle \text{ for the i-th constraint.} \tag{23.8.6}$$

Now multiply the i-th equation by μ_i, and sum over i using (23.8.3), to get

$$r^2 = \langle \mathbf{x}, \mathbf{x}\rangle - 2\langle \mathbf{x}, \sum_{i}^{m} \mu_i \mathbf{a}^i\rangle + \sum_{i}^{m} \mu_i \langle \mathbf{a}^i, \mathbf{a}^i\rangle.$$

We get, using (23.8.4) to substitute out $\sum_i^m \mu_i \mathbf{a}^i$,

$$r^2 = -\|\mathbf{x}\|^2 + \sum_{i=0}^{n} \mu_i \|\mathbf{a}^i\|^2,$$

as claimed. □

We know that there must be at least one active constraint (indeed, at least two, as we shall see later), so $k \geq 1$. By changing coordinates we place the first point at the origin (call it $\mathbf{a}^0$). The remaining k points $\mathbf{a}^i$ (numbered from 1 to k) corresponding to active constraints are linearly independent by Theorem 18.2.21. So the $n \times k$ matrix whose columns are the $\mathbf{a}^i$:

$$A = \left[a^i_j\right], \qquad \text{where } i \text{ is the column index,}$$

has rank k and (23.8.4) can be written

$$\mathbf{x} = A\mu. \tag{23.8.7}$$

The fact that the first point is the origin, and the constraint is active, says that $r^2 = \|\mathbf{x}\|^2$. Subtracting that equation from the other equations in (23.8.6) expressing that the constraints are active gives

$$\mathbf{a}^{iT}\mathbf{x} = \frac{\|\mathbf{a}^i\|^2}{2},$$

something that can be seen via elementary plane geometry. Grouping these equations all in one matrix equation we get

$$A^T\mathbf{x} = \mathbf{b}, \qquad \text{where } \mathbf{b} \text{ is the vector } \Big(\frac{\|\mathbf{a}^i\|^2}{2}\Big).$$

Replace $\mathbf{x}$ by its value given in (23.8.7) to get

$$A^TA\mu = \mathbf{b}.$$

Because n is at least as large as k, and A has maximal rank, Proposition 13.3.3 tells us that the $k \times k$ symmetric matrix is positive-definite and therefore invertible. This allows us to solve for μ. Then (23.8.7) gives us the center $\mathbf{x}$, and finally $r^2 = \|\mathbf{x}\|^2$, so we have all the variables.

It remains to check that the remaining (non-active) points are feasible.

The weakness of this method is that we have to try various combinations of active points to determine which one yields the solution.

This can be set up systematically in a way that resembles the construction of regular simplices in Example 18.3.24.

For the $n + 1$ points $\mathbf{a}^i$, let d_{ij} denote the distance between $\mathbf{a}^i$ and $\mathbf{a}^j$. Pick two points for which this distance is maximum, and let $\mathbf{c}^1$ be the midpoint of the segment they form. Then the ball $\overline{N}_{r_1}(\mathbf{c}^1)$, where $r_1 = d_{ij}/2$ contains the two points $\mathbf{a}^i$ and $\mathbf{a}^j$, and the constraint is effective. If all the remaining points are in $\overline{N}_{r_1}(\mathbf{c}^1)$, we are done. Otherwise pick a third point $\mathbf{a}^k$ such that the distance to $\mathbf{c}^1$

is maximum. By the argument given in Example 18.3.24 we can find a new center $\mathbf{c}^2$ and a new radius $r_2 > r_1$ so that the three points are effective constraints for the ball. Continuing in this way, we can arrive at the unique solution on the first try.

This also shows that any number of active constraints between 2 and n is possible.

23.8.8 Exercise. Generalize what is done, by considering m distinct points $\mathbf{a}^i$, $1 \leq i \leq m$, in $\mathbb{R}^n$. As before, the goal is to find the closed ball of minimum radius containing the points.

Hint: First, you can throw out any point that is in the convex hull of the remaining points: why? It is easy to reduce to the case the points generate the ambient space, so that you can assume that the first $n+1$ points are affinely independent. Then proceed as in the previous example.

Lecture 24

Quasiconvexity and Optimization

Quasiconvex functions generalize convex function. They have a simple and natural definition (24.1.1). We study them in the same way as convex functions, by successively adding differentiability requirements. Unlike convex functions, quasiconvex functions need not be continuous. After an extended example §24.5 showing that the internal rate of return is a quasiconcave function: Proposition 24.5.4, we conclude in §24.6 with a short account about the minimization of quasiconvex functions.

24.1 Quasiconvex functions

At the end of §21.1, we noted that quasiconvex functions are exactly those functions whose sublevels sets are convex. This could be used as a definition. We will take a different starting point, so that our definition is parallel with the definition for convex functions. The characterization by sublevel sets becomes Theorem 24.1.12.

24.1.1 Definition. A function $f : S \subset \mathbb{R}^n \to \mathbb{R}$ is *quasiconvex*[1] if

- The domain S of f is convex;
- For any two points $\mathbf{x}_1$ and $\mathbf{x}_2$ in S, and for all λ, $0 \leq \lambda \leq 1$ we have:

$$f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}. \tag{24.1.2}$$

If the second property is strengthened by adding: For any two points $\mathbf{x}_1$ and $\mathbf{x}_2$ in S with $f(\mathbf{x}_1) < f(\mathbf{x}_2)$, and any λ, $0 < \lambda < 1$, we have:

$$f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) < f(\mathbf{x}_2), \tag{24.1.3}$$

[1] Other references for quasiconvex function are [10] p. 95 and [5], p. 135.

then f is *strictly quasiconvex.*

The function f is *quasiconcave* (resp. strictly quasiconcave) if $-f$ is quasiconvex (resp. strictly quasiconvex).

24.1.4 Theorem. *Convex functions are quasiconvex. Indeed, they are strictly quasiconvex.*

Proof. It is enough to show that (21.1.2) implies (24.1.2) which is clear. Another proof would be to combine Theorem 21.1.15 and Theorem 24.1.12 below. The last statement is an exercise for the reader: look at Exercise 21.1.8. □

Quasiconvexity, like convexity (see Theorem 21.2.1), can be checked on lines:

24.1.5 Theorem. *A function f from the convex set $S \subset \mathbb{R}^n$ is quasiconvex if and only if for every line L meeting S the restriction f_L of f to $L \cap S$ is quasiconvex.*

The proof is identical. In some books, the following definition is used:

24.1.6 Definition. Let f be a function defined on a convex set $S \subset \mathbb{R}^n$. For any line L in $\mathbb{R}^n$ intersecting S, let f_L be the restriction of f to the interval $S_L = S \cap L$. If S_L contains two distinct points $\mathbf{x}_0$ and $\mathbf{x}_1$, we can write $f_L = f(\mathbf{x}_0 + t(\mathbf{x}_1 - \mathbf{x}_0)$, for t varying on an interval I containing $[0, 1]$. If S_L contains just one point $\mathbf{x}$, then of course $f_L(\mathbf{x}) = f(\mathbf{x})$.

Then f is *unimodal* on S, if for any line in $\mathbb{R}^n$, f_L either

1. is weakly monotonic on S_L, meaning that it is either weakly increasing, so that $f_L(s) \leq f_L(t)$ for all $s < t$ in I, or weakly decreasing, so that $f_L(s) \geq f_L(t)$ for all $s < t$ in I;
2. or changes direction exactly once, in the following way: there is a value t_0 in the interval such that for all $t < t_0$, f_L is weakly decreasing, so that $f_L(t) \geq f_L(t_0)$, and for all $t > t_0$, f_L is weakly increasing, so that $f_L(t_0) \leq f_L(t)$.

Notice that the definition does not depend on how S_L is parametrized: examine what happens when you interchange $\mathbf{x}_0$ and $\mathbf{x}_1$, for instance. This definition is somewhat unsatisfactory because it does not deal with the situation where f_L changes direction in the opposite way: starting out by weakly increasing, and then weakly increasing. This means we do not have a name for the corresponding concept for quasiconcave functions. If needed, we could call it *reverse unimodal*. In any case, we have the

24.1.7 Theorem. *A function f defined on a convex set S is quasiconvex if and only if it is unimodal.*

24.1.8 Exercise. Prove this by comparing the definitions of quasiconvexity and unimodality.

Because of this theorem, quasiconvex functions are sometimes called unimodal: see [10], p. 95. This makes it easy to show that quasiconvex functions need not be continuous.

24.1.9 Example. The function f on the interval $[-1, 2]$ given by

$$f(x) = \begin{cases} 1, & \text{if } -1 \le x < 0; \\ 0, & \text{if } 0 \le x < 1; \\ 1, & \text{if } 1 \le x \le 2. \end{cases}$$

is obvious unimodal and therefore quasiconvex. Yet it is not continuous, so it is not convex

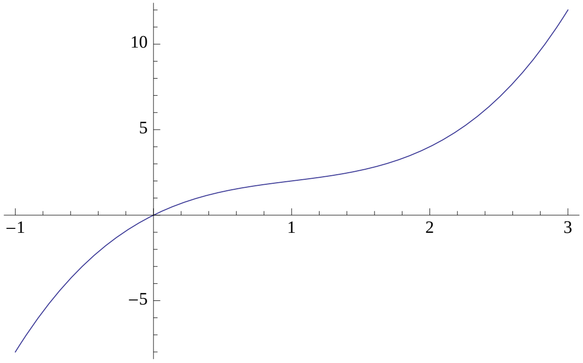
24.1.10 Exercise. Establish which step functions are unimodal.

If Item (2) in Definition 24.1.1 never occurs, then the function f is said to be quasilinear or quasimonotone. Obviously such a function is both quasiconvex and quasiconcave. Indeed, the only functions that are both quasiconvex and quasiconcave are quasilinear, just as the only functions that are both convex and concave are linear.

Quasiconvex function can have inflection points, as shown in Example 24.1.11. This makes the analysis of its critical points more difficult than those of convex functions.

24.1.11 Example (A quasiconvex functions with an inflection point)**.** .

Let $f(x) = x^3 - 3x^2 + 4x$. Its first derivative has no real roots, so $f'(x)$ increases on $\mathbb{R}$. Its second derivative $6x - 6$ is zero at $x = 1$, where it passes from negative to positive, so $f(x)$ is concave on $(\infty, 1)$ and convex on $(1, \infty)$. However since it is unimodal it is quasiconvex.



The argument shows that any cubic whose derivative has no real roots is quasiconvex (and quasiconcave, too). It also applies to cubics where the first derivative has a double root.

24.1.12 Theorem. *A function $f : S \subset \mathbb{R}^n \to \mathbb{R}$ is quasiconvex if and only if its sublevel sets S_c are convex for all $c \in \mathbb{R}$.*

Proof. First assume f is quasiconvex. Take two points $\mathbf{x}_1$ and $\mathbf{x}_2$ in S_c, so that $f(\mathbf{x}_1) \le c$ and $f(\mathbf{x}_2) \le c$. Then by (24.1.2) any point $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$, $0 \le \lambda \le 1$, on the segment $[\mathbf{x}_1, \mathbf{x}_2]$ satisfies

$$f(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \le \max\{f(\mathbf{x}_1), f(\mathbf{x}_2\} \le c,$$

so it is in S_c as required.

Now assume all the sublevel sets S_c are convex. Take any two points $\mathbf{x}_1$ and $\mathbf{x}_2$ in S, and let c be the bigger of $f(\mathbf{x}_1)$ and $f(\mathbf{x}_2)$. Then any point on the segment $[\mathbf{x}_1, \mathbf{x}_2]$ joining $\mathbf{x}_1$ and $\mathbf{x}_2$ is in S_c by convexity of S_c, so (24.1.2) is satisfied. □

Here are some additional properties of quasiconvex functions that extend some of the examples of convex functions given in §22.3.

24.1.13 Theorem. *Composition: assume the functions $h : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ can be composed, with $f(\mathbf{x}) = h(g(\mathbf{x}))$. Then f is quasiconvex if h is nondecreasing and g is quasiconvex.*

Let $\{f_\alpha\}$ be a collection (perhaps infinite) of convex functions, all defined on an open convex set S. Then $F(\mathbf{x}) = \sup_\alpha f_\alpha(\mathbf{x})$ is quasiconvex.

Proof. The first statement follows immediately from Definition 24.1.1. For the second statement, let S_c^α be the sublevel set of level c for the function f_α. Then the sublevel set S_c of level c of F is

$$S_c = \bigcap_\alpha S_c^\alpha.$$

As the intersection of convex sets, by Theorem 18.1.15 it is convex, so F is quasiconvex by Theorem 24.1.12. □

24.1.14 Example. Consider the function $f(x_1, x_2) = -x_1x_2$ restricted to the positive quadrant $x_1 \ge 0$ and $x_2 \ge 0$. Then the sublevel set S_c is given by $-x_1x_2 \le c$ in the positive quadrant, or $x_1x_2 \ge -c$. So S_c is the entire quadrant when $c \ge 0$. When $c < 0$, it is the set above a branch of the hyperbola $x_2 = -c/x_1$, which is convex as we saw in Exercise 22.6.18. Theorem 24.1.12 says $f(\mathbf{x})$ is quasiconvex.

Yet it is not convex: For example it fails the second derivative test for convexity. Its Hessian is the constant matrix

$$\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$$

so its characteristic polynomial is $\lambda^2 - 1$. Thus it has one positive and one negative eigenvalue.

In the next three sections, we make increasingly stringent regularity requirements on f: first we assume that it is continuous, then that it is $\mathcal{C}^1$ and finally that it is $\mathcal{C}^2$. In each case we replace f by its restriction to a line segment, and use single variable calculus to get the desired result. We use the following notation:

24.1.15 Notation. Here is how we reduce to the single variable case. We already used something similar in Definition 24.1.6. Given two points x_0 and x_1 in the convex domain S of f, we restrict to the line segment Σ joining the two points. Write $\mathbf{v} = \mathbf{x}_1 - \mathbf{x}_0$. We parametrize Σ by

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1 = \mathbf{x}_0 + t\mathbf{v} \quad \text{for } 0 \le t \le 1.$$

The parametrization is set up so that the notation $\mathbf{x}_t$ is consistent at the end points $t = 0$ and $t = 1$ and each point in the segment is expressed as a convex combination of $\mathbf{x}_0$ and $\mathbf{x}_1$. Then let the function g of a single variable $t \in \mathbb{R}$ be:

$$g(t) = f(\mathbf{x}_0 + t\mathbf{v}) = f(\mathbf{x}_t), \quad 0 \le t \le 1.$$

We write $g_{\mathbf{x}_0\mathbf{x}_1}$ if we need to emphasize the segment it comes from. Usually we do not.

If $f(\mathbf{x})$ is $\mathcal{C}^1$, then by the chain rule:

$$\begin{aligned} g'(t) &= \langle \nabla f(\mathbf{x}_0 + t\mathbf{v}), \mathbf{v} \rangle = \langle \nabla f(\mathbf{x}_t), \mathbf{v} \rangle; \\ g'(0) &= \langle \nabla f(\mathbf{x}_0), \mathbf{v} \rangle; \\ g'(1) &= \langle \nabla f(\mathbf{x}_1), \mathbf{v} \rangle. \end{aligned}$$

and if $f(\mathbf{x})$ is $\mathcal{C}^2$, writing $F(\mathbf{x})$ for the Hessian of f at $\mathbf{x}$:

$$\begin{aligned} g''(t) &= \mathbf{v}^T F(\mathbf{x}_0 + t\mathbf{v})\mathbf{v} = \mathbf{v}^T F(\mathbf{x}_t)\mathbf{v}; \\ g''(0) &= \mathbf{v}^T F(\mathbf{x}_0)\mathbf{v}; \\ g''(1) &= \mathbf{v}^T F(\mathbf{x}_1)\mathbf{v}. \end{aligned}$$

24.2 Continuous Quasiconvex Functions

If the function f is continuous, we can slightly weaken Condition 24.1.2 needed for quasiconvexity. This technical result can be skipped.

24.2.1 Theorem. *A continuous function $f(\mathbf{x})$ is quasiconvex on S if and only if for $\mathbf{x}_0$ and $\mathbf{x}_1$ in S, with $f(\mathbf{x}_0) < f(\mathbf{x}_1)$, we have*

$$f(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_0) \le f(\mathbf{x}_1). \tag{24.2.2}$$

Proof. The condition is necessary, since by hypothesis $\max\{f(\mathbf{x}_0), f(\mathbf{x}_1)\} = f(\mathbf{x}_1)$. To show sufficiency, we need to consider the case $f(\mathbf{x}_1) = f(\mathbf{x}_0)$ and establish (24.2.2) in that case. We reduce to a single-variable calculus lemma.

24.2.3 Lemma. *Let $g(t)$ be a continuous function on the closed interval $[0,1] \subset \mathbb{R}$, such that $g(0) = g(1) = c$. Assume that for all r and s in $[0,1]$ such that $g(r) < g(s)$, then $g(t) \le g(s)$ for all t in the segment bounded by r and s. The lemma applies when $r < s$ and when $s < r$.*

Then for any $t \in [0,1]$, $g(t) \le c$.

Proof. Suppose not. Then there is a $s \in (0,1)$ with $g(s) = d > c$. Since g is continuous, by the intermediate value theorem g takes on all values between c and d on the interval $(s,1)$, (and on the interval $(0,s)$, though we do not use that fact.). So pick a $r \in (s,1)$ with $g(r) = e$, $c < e < d$. Then since $g(0) = c$, $g(s) = d$, $g(r) = e$ and $0 < s < r$, the hypothesis is violated. □

The theorem follows from the lemma by restricting to the segment $[\mathbf{x}_0, \mathbf{x}_1]$ using the usual function $g(t) = f(\mathbf{x}_0 + t(\mathbf{x}_1 - \mathbf{x}_0))$. The lemma shows that whenever $f(\mathbf{x}_0) = f(\mathbf{x}_1) = c$, f only takes on values $\le c$ at points between $\mathbf{x}_0$ and $\mathbf{x}_1$, which is exactly what we had to prove. □

24.3 Continuously Differentiable Quasiconvex Functions

Assume f is $\mathcal{C}^1$. Then we have the following result[2]

24.3.1 Theorem. *A continuously differentiable function $f(\mathbf{x})$ is quasiconvex if and only if for all $\mathbf{x}_0$ and $\mathbf{x}_1$ in S with $f(\mathbf{x}_0) \le f(\mathbf{x}_1)$ we have*

$$\langle \nabla f(\mathbf{x}_1), \mathbf{x}_1 - \mathbf{x}_0 \rangle \ge 0. \tag{24.3.2}$$

[2] Due to Arrow and Enthoven ([2]), and occasionally called the fundamental theorem of quasiconvex functions: [1], p. 446.

In terms of the function $g = g_{\mathbf{x}_0\mathbf{x}_1}$ of 24.1.15, the expression in (24.3.2) is $g'(1)$.

Proof. First we assume f is quasiconvex. If $\nabla f(\mathbf{x}_1) = \mathbf{0}$, there is nothing to prove: we have equality in (24.3.2). At a point where $\nabla f(\mathbf{x}_1)$ is not the zero vector, $\nabla f(\mathbf{x}_1)$ is the normal vector to the level set of f of level $c = f(\mathbf{x}_1)$ at $\mathbf{x}_1$. Thus the hyperplane passing through $\mathbf{x}_1$ with normal vector $\nabla f(\mathbf{x}_1)$ is a supporting hyperplane to the convex sublevel set S_c. By Theorem 16.3.6, since f is continuous, S_c is closed. Now $\mathbf{x}_0$ is in S_c. Thus Corollary 18.6.4 applies: in fact it shows we have strict inequality in (24.3.2), so we are done.

Next we show that (24.3.2) is a sufficient condition for quasiconvexity. Assume that for any pair of points $\mathbf{x}_0$ and $\mathbf{x}_1$ in S with $f(\mathbf{x}_0) \le f(\mathbf{x}_1)$, (24.3.2) is satisfied. The $\mathcal{C}^1$ function $g(t)$ from 24.1.15 on the segment $[\mathbf{x}_0, \mathbf{x}_1]$ has a maximizer t^*, $0 < t^* < 1$, by the Weierstrass theorem, and $g'(t^*) = 0$. To show that f is quasiconvex, we must show that $g(t^*) \le g(1) = f(\mathbf{x}_1)$. Assume not. Then $g(t^*) > g(1)$. Since g is continuous, the mean value theorem gives a t_1, $t^* < t_1 < 1$, and $g'(t_1) < 0$. Then (24.3.2) fails on the segment $[\mathbf{x}_0, \mathbf{x}_0 + t_1\mathbf{v}]$, since the derivative computations at the end of 24.1.15 show that the expression on the left-hand side of (24.3.2) is $g'(t_1)$. Obviously t_1 can also be chosen so that $g(t_1) > g(1)$, so that $f(\mathbf{x}_0) \le f(\mathbf{x}_0 + t_1\mathbf{v})$, so we have our contradiction. □

The first part of the proof shows:

24.3.3 Corollary. *Let $f(\mathbf{x})$ be a differentiable function on an open set S. Assume f is quasiconvex. Then for all $\mathbf{x}_0$ and $\mathbf{x}_1$ in S such that $f(\mathbf{x}_0) \le f(\mathbf{x}_1)$ and such that $\nabla f(\mathbf{x}_1) \ne 0$, we have*

$$\langle \nabla f(\mathbf{x}_1), \mathbf{x}_1 - \mathbf{x}_0 \rangle > 0.$$

24.4 $\mathcal{C}^2$ Quasiconvex Functions

Finally we assume that f is twice continuously differentiable. We write $F(\mathbf{x})$ for the Hessian of f at $\mathbf{x}$.

24.4.1 Theorem. *Assume $f(\mathbf{x})$ is a $\mathcal{C}^2$ function on the open convex set S.*

- ***The Necessary Condition.*** *Assume $f(\mathbf{x})$ is quasiconvex. Then for all $\mathbf{x} \in S$, and all vectors $\mathbf{v} \in \mathbb{R}^n$ such that $\langle \mathbf{v}, \nabla f(\mathbf{x})\rangle = 0$, $\mathbf{v}^T F(\mathbf{x})\mathbf{v} \ge 0$.*
- ***The Sufficient Condition.*** *Assume that for all $\mathbf{x} \in S$, and all vectors $\mathbf{v} \ne \mathbf{0} \in \mathbb{R}^n$ such that $\langle \mathbf{v}, \nabla f(\mathbf{x})\rangle = 0$, $\mathbf{v}^T F(\mathbf{x})\mathbf{v} > 0$. Then f is quasiconvex.*

Proof. As usual, we reduce to the one-variable case. We fix a point $\mathbf{x} \in S$ and a non-zero direction $\mathbf{v}$ at $\mathbf{x}$. Then we get a line $\mathbf{x} + t\mathbf{v}$, as t varies in $\mathbb{R}$, and this line intersects S in an open interval I that contains $t = 0$, since $\mathbf{x} \in S$ and S is open and convex. All line segments in S can be obtained in this way. By scaling $\mathbf{v}$ we may assume that $\mathbf{x} + \mathbf{v}$ is in S. We will use this later. Consider the restriction $f(\mathbf{x} + t\mathbf{v})$ of f to the interval I, for $\mathbf{x} \in S$. To show that f is quasiconvex is to show that for each choice of $\mathbf{x} \in S$ and $\mathbf{v}$ the function $f(\mathbf{x} + t\mathbf{v})$ is quasiconvex, as noted in Theorem 24.1.5.

First we establish the necessary condition. For the composite function $g(t) = f(\mathbf{x}+t\mathbf{v})$ of one variable t, the hypotheses say that when $g'(t) = 0$, then $g''(t) \geq 0$. Indeed, if $g''(t) < 0$, then g has a strict local maximum at x, and this contradicts Theorem 24.1.7.

Now on to the sufficient condition. It implies that if $g'(t) = 0$, then $g''(t) > 0$, showing that g has a strict minimum at t. Thus all critical points are strict minima, so the restriction of f to the line is unimodal and therefore quasiconvex. Thus f itself is quasiconvex. □

As already noted, a quasiconvex function can have inflection points. Take one that has an inflection point at a critical point. Such a function will not satisfy the sufficient condition. A simple example is $f(x) = x^3$. It is unimodal so it is quasiconvex, and yet it fails the sufficient condition at $x = 0$. On the other hand the function $f(x) = -x^4$ satisfies the necessary condition, since the only time the hypothesis is met is when $x = 0$, in which case the second derivative is 0. And yet $f(x) = -x^4$ is obviously not quasiconvex on any interval containing the origin.

Here is a variant of the sufficient condition in this theorem: it eliminates the strict inequality when f has no critical points. This result is due to Otani [49]. It gives a beautiful variant for quasiconvex function of Theorem 22.2.1.

24.4.2 Theorem. *Assume that f is defined on an open convex set S, and that the gradient $\nabla f(\mathbf{x})$ of f is non-zero at every point $\mathbf{x} \in S$. Then for every $\mathbf{x} \in S$, the $(n-1) \times (n-1)$ symmetric submatrix $F_\perp(\mathbf{x})$ of the Hessian F restricted to the orthogonal complement of $\nabla f(\mathbf{x})$ exists. Then $F_\perp(\mathbf{x})$ is positive semidefinite for all $\mathbf{x} \in S$ if and only if the function f is quasiconvex.*

Proof. We introduced $F_\perp(\mathbf{x})$ in (17.7.2). The necessity of the condition for quasiconvexity is the necessary condition of Theorem 24.4.1, so there is nothing to prove.

To establish the sufficient condition, we show that any sublevel set S_c of f is convex. By Theorem 21.1.15 this establishes that f is quasiconvex. Since the gradient of f is never $\mathbf{0}$, we can apply the results of §17.7 at any point $\mathbf{x}^* \in S$. Assume $f(\mathbf{x}^*) = c$. Then in a small enough neighborhood of $\mathbf{x}^*$, by the implicit function theorem the level set can be written $x_b = g(\mathbf{x}_f)$. Here the $\mathbf{x}_f$ are the $n-1$ free variables, and x_b is the remaining variable: the bound variable.

The tangent space $T_{\mathbf{x}^*}$ to $f(\mathbf{x}) = c$ at the point $\mathbf{x}^*$ is given by

$$\sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) = 0.$$

as we learned in §17.3. This expresses it as the orthogonal complement of the gradient $\nabla f(\mathbf{x}^*)$ at $\mathbf{x}^*$. Thus in our usual notation for hyperplanes (see Example 18.1.7), $T_{\mathbf{x}^*}$ is written $H_{\nabla f(\mathbf{x}^*),c}$.

We will show that $T_{\mathbf{x}}$ is a supporting hyperplane to S_c at $\mathbf{x}$. Then Corollary 18.6.12 tells us that S_c is convex.

Pick a point $\mathbf{x}^*$ with $f(\mathbf{x}^*) = c$.

24.4.3 Lemma. *As long as the parametrization of the level set by the function $g(\mathbf{x}_f)$ remains valid in a neighborhood of $\mathbf{x}^*$, the graph of g, for values of the free variables close to $\mathbf{x}_f$, lies in the negative half-space $H^-_{\nabla f(\mathbf{x}),c}$.*

Proof. We use (17.7.2), and the interpretation of $F_\perp$ as the restriction of F to the tangent space, so that by hypothesis $F_\perp$ is positive semidefinite along the graph of g. Thus if we orient the x_1 axis so that $\nabla f(\mathbf{x}^*)$ has positive coordinates in that direction, then the graph of g is concave and therefore lies below $T_{\mathbf{x}^*}$ - meaning on the opposite side of $\nabla f(\mathbf{x}^*)$. □

To finish the proof, we show that the sublevel set S_c is the intersection over all $\mathbf{x}$ in S satisfying $f(\mathbf{x}) = c$ of the half-spaces $H^-_{\nabla f(\mathbf{x}),c}$. This will show S_c is convex.

So suppose not: pick any point $\mathbf{x}_0$ in S_c, and assume that $\mathbf{x}_0$ lies in the positive half-space $H^+_{\nabla f(\mathbf{x}_1),c}$ for some $\mathbf{x}_1$. As usual form the function $g(t)$ (nothing to do with the implicit function) equal to $f(\mathbf{x}_0 + t(\mathbf{x}_1 - \mathbf{x}_0))$. It clearly reaches a maximum strictly greater than c at a point $t_0 \in (0,1)$. At that point, the level curve must be tangent to the line $\mathbf{x}_0 + t(\mathbf{x}_1 - \mathbf{x}_0)$. A moment's thought will tell you that the sublevel set cannot be convex there: in fact it is concave locally.

□

We conclude with a theorem of Arrow-Enthoven (Theorem 4 of [2]), originally proved many years before the previous theorem from which it now follows.

24.4.4 Corollary. *Assume $f(x,y)$ is $\mathcal{C}^2$ and defined on the positive quadrant. Also assume that the partial derivatives f_x and f_y are both positive at every point (x_0, y_0) in the first quadrant. Consider the expression*

$$f_x^2 f_{yy} - 2 f_x f_y f_{xy} + f_y^2 f_{xx}. \tag{24.4.5}$$

The function f is quasiconcave if and only if this expression is everywhere non-positive.

Proof. Because the partials are not zero, f has no critical points and Theorem 24.4.2 applies. The expression in (24.4.5) divided by f_x^2 is the Hessian of f restricted to the tangent lines of the critical points, as we saw in (17.5.9).

Since $f_x^2 > 0$, the assumption that (24.4.5) is positive is equivalent to the Hessian being positive. So simply multiplying f by $-f$ we can apply the theorem, so $-f$ is quasiconvex and f is quasiconcave. This is the bordered Hessian. □

24.4.6 Remark. This corollary has an interesting economic interpretation. We think of f as the utility function of a consumer, where x and y represent the quantities of two goods between which the consumer can choose. Then the level curves of $f(x, y)$ are the indifference curves of the consumer.

24.5 Example: the Internal Rate of Return

The internal rate of return will provide us with an example of a quasiconcave function f, so a function such that $-f$ is quasiconvex. Thus $f : S \subset \mathbb{R}^n \to \mathbb{R}$ is quasiconcave if and only if its domain S is convex and if the superlevel sets $S^c = \{\mathbf{s} \in S \mid f(\mathbf{x}) \geq c\}$ are convex for all $c \in \mathbb{R}$.

Let $\mathbf{x} = (x_0, x_1, \dots, x_n)$ denote the cash flow in dollars of an investor over n equally spaced time intervals. By convention x_i positive means the investor receives money, negative means the investor dispenses money. We assume that x_0 is negative. The interest rate r over the entire period is assumed to be positive and constant. At time 0, when the investor invests x_0 dollars, the *present value* of the cash flow is, by definition:

$$PV(\mathbf{x}, r) = x_0 + \sum_{i=1}^{n} \frac{x_i}{(1+r)^i}. \tag{24.5.1}$$

We assume:

$$x_0 < 0, \text{ as already mentioned, and } \sum_{i=0}^{n} x_i > 0. \tag{24.5.2}$$

Each inequality delimits an open half-space, so that the intersection C of these two half-spaces in $\mathbb{R}^{n+1}$ is convex by Theorem 18.1.15.

In terms of the function PV, (24.5.2) says that $PV(\mathbf{x}, 0) = \sum_{i=0}^{n} x_i > 0$, and for large enough r, $PV(\mathbf{x}, r) < 0$, since the limit of $PV(\mathbf{x}, r)$ as $r \to \infty$ is x_0. Since the function $PV(\mathbf{x}, r)$ is continuous as a function of r alone (just check (24.5.1)), this means, by the intermediate value theorem, that for each fixed $\mathbf{x} \in \mathbb{R}^{n+1}$, there is a positive interest rate r that yields $PV(\mathbf{x}, r) = 0$.

There might be several such r, since our hypotheses given by (24.5.2) do not guarantee that for all $\mathbf{x}$ the function $PV(\mathbf{x}, r)$ is a decreasing function of r.

24.5.3 Example. Assume that n is odd, and let $y = 1/(1+r)$, so that (24.5.1) becomes a polynomial of degree n in y:

$$P(y) = \sum_{i=0}^{n} x_i y^i.$$

We can let y vary from 0 (corresponding to an infinite interest rate r) to 1, corresponding to $r = 0$. We can easily arrange for $P(y)$ to have several roots between 0 and 1. Take $n-1$ numbers t_i, $0 < t_1 < t_2 < \dots < t_{n-1} < 1$ and set

$$P(y) = (y - t_1)(y - t_2) \dots (y - t_{n-2})(y - t_{n-1})^2,$$

so this polynomial has all its roots real, between 0 and 1, and exactly one double root t_{n-1}. We let x_i be the coefficient of degree i of $P(y)$. Then $P(0) = x_0$ is negative because n is odd, and $P(1)$, the sum of the coefficients, is positive, because all the roots are smaller than 1. So this polynomial, for any choice of the t_i as above, meets the requirements (24.5.2), and yet $PV(\mathbf{x}, r)$ is not a decreasing function of r.

For each $\mathbf{x}$ denote by $IRR(\mathbf{x})$ the smallest positive interest rate that makes $PV(\mathbf{x}, r) = 0$. Thus IRR is a function of the cash flow $\mathbf{x}$, given implicitly, called the *internal rate of return* of the cash flow $\mathbf{x}$. Is $r(x_1, \dots, x_n)$ locally a differentiable function of $\mathbf{x}$? We could hope to prove this using the Implicit Function Theorem 17.6.6. Example 24.5.3 shows that this will not work in general. Indeed, because of the double root corresponding to $IRR(\mathbf{x})$, the partial derivative with respect to r at this $\mathbf{x}$ is 0.

Thus the IRR is a value function: For each cash flow $\mathbf{x}$, it is an optimum value for the remaining variable r. Thus we are not too surprised to find:

24.5.4 Proposition. *The internal rate of return $IRR(\mathbf{x})$ is a quasiconcave function on C.*

Proof. Fix a real number c. Then notice:

$$IRR(\mathbf{x}) \geq c \iff PV(\mathbf{x}, r) > 0, \text{ for all } r \text{ such that } 0 \leq r < c. \tag{24.5.5}$$

This gives a description of the superlevel set S^c of $IRR(\mathbf{x})$ that we now exploit.

For a fixed number r, the set $V_r = \{\mathbf{x} \in \mathbb{R}^{n+1} \mid PV(\mathbf{x}, r) < 0\}$ is an open half-space: indeed (24.5.1) is linear in the x_i. Recall that open half-spaces are convex. Now take the intersection of all the V_r for $0 \leq r < c$. This is a (uncountably infinite) intersection of half-spaces. Such an intersection is convex by Theorem 18.1.15. Then intersect with the convex set C: this is convex. This intersection is the set of $\mathbf{x} \in C$ satisfying the right-hand side of (24.5.5), so it is the superlevel set S^c of the function IRR. The convexity of all the S^c says that IRR is quasiconcave. □

24.6 The Optimization of Quasiconvex Functions

Quasiconvex functions form a broader class of functions than convex functions so they are harder to minimize. We record just two results. First, much of Theorem 22.4.1 remains, since it relies only on the convexity of the sublevel sets, which is still true for quasiconvex functions. The proof is the same as in the convex case. Note that a quasiconvex function could have a local minimum that is not a global minimum: for instance a step function. However if you require that the local minimum be strict, this cannot happen.

24.6.1 Theorem. *Let f be a quasiconvex function defined on the open convex set S. Then if f has a strict local minimum at $\mathbf{x}_1$, it has a global minimum there. The set of points $M \subset S$ where the global minimum of f is attained is either empty or convex.*

As already noted, the analog of Theorem 22.4.8 is false for quasiconvex functions.

Now a second result for $\mathcal{C}^1$ quasiconvex functions. Note that this result only applies when the minimizer is not in the interior of S, because in the interior, a necessary condition to be a minimizer in that $\nabla f(\mathbf{x}_0) = 0$, which cannot happen in the theorem below.

24.6.2 Theorem. *A sufficient condition for $\mathbf{x}_0$ to be a global minimizer of the quasiconvex $\mathcal{C}^1$ function f on the feasible set S is that*

$$\langle \nabla f(\mathbf{x}_0), \mathbf{x}_1 - \mathbf{x}_0 \rangle > 0, \quad \forall \mathbf{x}_1 \in S, \mathbf{x}_1 \neq \mathbf{x}_0. \tag{24.6.3}$$

Proof. Once again we reduce to a segment by the methods of Notation 24.1.15. So we write $\mathbf{v} = \mathbf{x}_1 - \mathbf{x}_0$, $g(t) = f(\mathbf{x}_0 + t\mathbf{v})$. Then the expression in (24.6.3) is $g'(0)$, as we have already noted. □

Compare this to Corollary 22.4.4, which treats the analogous question when f is convex. Our hypothesis here is stronger, since we require strict inequality, and we only have a sufficient condition, not a necessary condition. The issue once again is that f might have inflection points, which need to be excluded.

24.6.4 Exercise. Let P be a polytope in $\mathbb{R}^n$, and let $f(\mathbf{x})$ be a continuous quasiconvex function on P. By the Weierstrass Theorem 16.2.2 $f(\mathbf{x})$ has a maximum on P. Show that it has a maximum at an extreme point of P, namely a vertex of the polytope in a minimal representation.

Hint: Let c be the maximum value of f at the extreme points of P. Determine S_c using the fact that it is convex.

Part VII

Linear Optimization

Lecture 25

Linear Optimization and Duality

In linear optimization both the objective function and the constraints are linear. The previous lecture allowed us to describe the feasible set of any linear optimization problem. Using those results, we show that a minimizer, if one exists, to the most important linear optimization problems occurs at one of the finite number of extreme points of the convex feasible set. This reduces the problem to the examination of the values of objective function at a finite number of points—finite, but potentially very large, so a systematic way of testing the extreme points is needed. That is the simplex method, to be studied in Lecture 27.

Here we focus on the Duality Theorem 25.5.1, which associates to any linear optimization problem a second apparently unrelated problem. In reality the two problems are closely connected, and are generally solved in tandem. After the duality theorem we deal with the Equilibrium Theorem 25.6.1, which introduces the notion of complementary slackness. §25.7 shows how the duality theorem and the equilibrium theorem show the path to the simplex method, to be studied in Lecture 27. Next we show how these two results allow us to find a simple expression for the "shadow price".

25.1 The Problem

We start by defining a *linear optimization problem.* As usual we only deal with minimization.

First, the objective function f is linear. We work in $\mathbb{R}^n$, and write

$$f(\mathbf{x}) = c_1x_1 + \cdots + c_nx_n = \mathbf{c}^T\mathbf{x}, \tag{25.1.1}$$

for a n-vector $\mathbf{c}$ of constants and a n-vector of variables $\mathbf{x} = (x_j)$, $1 \leq j \leq n$, always pairing the running index j with n.

Next, all the constraints are affine, so each one describes either an affine hyperplane

$$a_1x_1 + \dots + a_jx_j + \dots + a_nx_n = b, \tag{25.1.2}$$

or an affine half-space

$$a_1x_1 + \dots + a_jx_j + \dots + a_nx_n \geq b \tag{25.1.3}$$

The number of constraints is denoted m. The running index i will be associated with m. We denote the subset of $(1, \dots, i, \dots, m)$ of indices i where we have an inequality in the constraint by $\mathcal{I}$.

Finally, some (usually all) of the variables x_j are constrained to be non-negative. We denote the subset of $(1, \dots, j, \dots, n)$ of indices j where x_j is constrained to be non-negative by $\mathcal{J}$. So

$$x_j \geq 0, \text{for } j \in \mathcal{J}. \tag{25.1.4}$$

We could absorb these constraints into the inequality constraints (25.1.3), but it is usually better to handle them separately, as they indicate that the corresponding variables only make sense when they are nonnegative.

We allow a mixture of equality and inequality constraints. By adding slack variables, as we saw in §19.7 , one can always reduce an inequality constraint to an equality constraint. For simplicity we will usually only consider the situation where there are either only equality constraints (other than the positivity constraints), or only inequality constraints, and leave it to you to see that any linear optimization problem can be reduced to one of these.

Thus we have several different kinds of linear optimization problems, depending on the structure of the constraint set. We studied all possible constraint sets in §19.6, 19.7 , and 19.8.

We give names to the two most important linear minimization problems, the ones with constraint set studied in §19.6 and 19.7.

First we list the problem one always reduces to in order to do computations.

25.1.5 Definition. The *asymmetric*, or *canonical*[1] minimization problem is:
Minimize $\mathbf{c}^T\mathbf{x}$ subject to the constraints $A\mathbf{x} = \mathbf{b}$ and $\mathbf{x} \succeq \mathbf{0}$.

Secondly, the problem that is the most pleasant to handle theoretically, because its associated dual has the same form as itself, as we will see in Definition 25.3.9.

25.1.6 Definition. The *symmetric* or *standard*[2] minimization problem is:
Minimize $\mathbf{c}^T\mathbf{x}$ subject to the constraints $A\mathbf{x} \succeq \mathbf{b}$ and $\mathbf{x} \succeq \mathbf{0}$.

[1] Called standard in [42] and [10], p. 146; but canonical in [24], p. 75, [23] p. 14,[22], p. 61, and [7], p.144. The name canonical seems more traditional, so we will stick with it.

[2] Called canonical in [42]; inequality form in [10], p. 146; and standard in in [24], p.74, [23] p. 12, [22], p. 61, and [7], p.140.

It is enough to understand the symmetric (standard) and asymmetric (canonical) problems to gain full understanding of the material.

In both cases A is an $m \times n$ matrix $[a_{ij}]$ of constants, and $\mathbf{b}$ an m-vector of constants, called the *constraint vector*. The vector $\mathbf{c}$ is usually called the *cost* vector and the goal is to minimize cost. Strict inequalities are never allowed in the constraint set, because we want them to be closed sets. In these notes we require that all the inequalities go in the '$\geq$' direction. This can always be achieved by multiplying by -1 the inequalities that go in the other direction.

Since $\mathbf{a}^i$ denotes the i-th row of A, we will also write $\mathbf{a}^i \cdot \mathbf{x} = b_i$ or $\mathbf{a}^i \cdot \mathbf{x} \geq b_i$ for the i-th constraint.

Here is the general case, building on the notation of Definition 19.8.1.

25.1.7 Definition. The linear minimization problem associated to the objective function $\mathbf{c}^T\mathbf{x}$, to the $m \times n$ matrix $A = [a_{ij}]$ and to the m-vector of constants $\mathbf{b}$, and to the indexing sets $(\mathcal{I}, \mathcal{J})$ is

Minimize $\mathbf{c}^T\mathbf{x}$ subject to the constraints

- $x_j \geq 0$ when $j \in \mathcal{J}$,
- $\mathbf{a}^i \cdot \mathbf{x} \geq b_i$ when $i \in \mathcal{I}$,
- $\mathbf{a}^i \cdot \mathbf{x} = b_i$ when $i \notin \mathcal{I}$.

Thus in both the symmetric and asymmetric minimization problems $\mathcal{J} = (1, \dots, n)$, while in the symmetric problem $\mathcal{I} = (1, \dots, m)$, and in the asymmetric problem $\mathcal{I}$ is empty.

The constraints define the feasible set F of the problem.

25.1.8 Proposition. *The feasible set for any linear optimization problem (in particular for 25.1.5 and 25.1.6), if non-empty, is closed and convex.*

Proof. Each one of the inequalities describes a closed half-space, and each equality constraint describes a linear subspace . A closed half-space is both closed and convex, as is a linear subspace. The set where all the inequalities are satisfied the intersection of these closed and convex spaces, and therefore is either empty or both closed and convex: see Theorem 18.1.15. □

Does a linear optimization problem always have a solution? No. First of all, the feasible set F might be empty. Even if the F is non-empty, the set of values $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$, for $\mathbf{x} \in F$, could be unbounded negatively, so there is no minimum value.

Here is a case where the answer is yes.

25.1.9 Theorem. *If the set F defined by the constraints in 25.1.7 is non-empty and bounded, then any linear optimization problem with feasible set F has a solution.*

Proof. The objective function is continuous, and it is defined on a closed set F. Since F is bounded, it is compact, so by the Weierstrass Theorem 16.2.2 there is a point $\mathbf{x}^* \in F$ that is a minimizer for the objective function $\mathbf{c} \cdot \mathbf{x}$. □

Note that a linear optimization problem may have a solution even if the feasible set is unbounded: e.g. Example 1.3.1.

In conclusion, we have established:

25.1.10 Theorem. *A linear optimization problem falls into one of the following three categories:*

(e) its feasible set F is empty.

(u) F is non-empty but the objective function is unbounded on F, which itself must be unbounded.

(f) F is non-empty and the objective function has a finite optimum value. This is always the case if F is a non-empty and bounded set.

Next a general result building on convexity. The objective function, since linear, is a convex function, and we are studying it on a convex set F. Thus Theorem 22.4.1 holds. We use P to denote the feasible set of the problem. P is a polyhedron in the affine space given by the equality constraints.

25.1.11 Theorem. *Consider the objective function $f(\mathbf{x}) = \langle \mathbf{c}, \mathbf{x} \rangle$ on P. Assume $f(\mathbf{x})$ attains its minimum value e on P, so that we are in case (f) of Theorem 25.1.10. Let P_e be the intersection of P with the affine hyperplane $H_{\mathbf{c},e}$, namely, the locus of points $\mathbf{x}$ such that $f(\mathbf{x}) = e$. Then*

1. *$H_{\mathbf{c},e}$ is a supporting hyperplane[3] for the convex set P, with P in the positive half-space $H^+_{\mathbf{c},e}$ associated to the hyperplane.*

2. *P_e is convex and closed. It is the locus of minimizers for the optimization problem.*

3. *Any extreme point of P_e is an extreme point of P;*

4. *If P_e is compact, then it has extreme points by Minkowski's Theorem 18.7.1. Pick such an extreme point, call it x^*. Then x^* is an extreme point of P, so that $f(\mathbf{x})$ has a minimizer $\mathbf{x}^*$ that is an extreme point for the polyhedron P.*

5. *If P_e is a single point x^*, then x^* is an extreme point of P.*

[3] See Definition 18.6.10.

We do not claim that P_e has extreme points. Indeed, by Exercise 18.7.10, this will not be true without extra hypotheses. However for both the standard and the canonical problems, we saw in Theorem 19.6.13 and Proposition 19.7.5 that there are always extreme points.

Proof. For Item (1), note that for any 'cost' e, the intersection of the hyperplane $H_{\mathbf{c},e}$ with the feasible set P is the set of feasible points where the objective function takes the value e. $H_{\mathbf{c},e}$ divides $\mathbb{R}^n$ into two half-spaces. The normal vector $\mathbf{c}$ of $H_{\mathbf{c},e}$ points in the direction of increasing cost, and determines the positive half-space $H^+_{\mathbf{c},e}$.

The key point is that minimality of the objective function at $\mathbf{x}^0$ means that the hyperplane $H_{\mathbf{c},e}$ is a supporting hyperplane for the convex set F at $\mathbf{x}^0$, and F is in the half-space H^+_e. Indeed, to be a supporting hyperplane at $\mathbf{x}^0$ means that the hyperplane passes through $\mathbf{x}^0$, so that the cost at $\mathbf{x}^0$ is e, and the fact that P lies in H^+_e means that the cost at any point of P is at least e.

Item (2): Because P_e is the intersection of the two closed and convex subspaces P and $H_{\mathbf{c},e}$, it is closed and convex.

Item (3): To show that any extreme point of P_e is an extreme point of P use Theorem 18.7.3. If $\mathbf{p}$ is extreme for P_e, there are $n-1$ linearly independent constraint hyperplanes h_k in $H_{\mathbf{c},e}$, all active at $\mathbf{p}$. These hyperplanes are the intersection with $H_{\mathbf{c},e}$ of constraint hyperplanes H_k in $\mathbb{R}^n$, which are active at $\mathbf{p}$. $H_{\mathbf{c},e}$ is another active constraint hyperplane, clearly linearly independent from the previous ones, so we have n independent linearly independent active constraints to P at $\mathbf{p}$, showing that $\mathbf{p}$ is extreme for P.

Item (4) is an immediate consequence of Minkowski's Theorem 18.7.1, and item (5) is a special case of (4). □

This theorem does not help us find the solution of a linear optimization problem. It just tells us what the set of all solutions looks like, assuming there is one solution: it is a closed convex set, a fact that already follows from Theorem 22.4.1.

25.2 Basic Solutions

25.2.1 Remark. In this section we treat the asymmetric (canonical) problem 25.1.5. We imagine that the problem comes from the symmetric problem 25.1.6 by adding a slack variable 19.7.2 to each of the m constraint equation, so we denote the total number of variables by $n+m$ (called n previously). So A is an $m\times(n+m)$ matrix of rank m. The slack variables that were called z_i are now noted x_{n+i}, $1\le i\le m$.

First we analyze the relationship between the symmetric and the associated asymmetric problems.

We start with the symmetric linear optimization problem 25.1.6: Minimize $\mathbf{c}^T\mathbf{x}$ subject to the constraints $A\mathbf{x} \geq \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$, where $\mathbf{x}$ is the n-vector of unknowns, A is a given $m \times n$ matrix with $m < n$ of rank m, and $\mathbf{c}$ and $\mathbf{b}$ are vectors of length n and m respectively, as forced by the other dimensions.

We pass to the asymmetric problem given in Proposition 19.7.5, so we write, in block notation (see §6.10), $A' = [A, -I]$, where I is the $m \times m$ identity matrix, $\mathbf{x}' = [\mathbf{x}, \mathbf{z}]$, so that $\mathbf{x}'$ is an $n + m$ vector with $x'_j = x_j$ for $1 \leq j \leq n$, and $x'_{n+i} = z_i$, for $1 \leq i \leq m$. We also write $\mathbf{c}'$ for the $n + m$ vector with $c'_i = c_i$ for $1 \leq i \leq n$ and all the other entries 0. So we have a second minimization problem:

Minimize $\mathbf{c}'^T\mathbf{x}'$ subject to the constraints $A'\mathbf{x}' = \mathbf{b}$ and $\mathbf{x}' \geq 0$.

The original problem and this problem are equivalent, in the following sense:

25.2.2 Theorem.

- *The projection from $\mathbb{R}^{n+m}$ to $\mathbb{R}^n$ obtained by omitting the last m coordinates is a one-to-one map of the feasible sets, from the feasible set of the equality problem to the feasible set of the inequality problem.*

- *The two problems are of the same type in the sense of Theorem 25.1.10.*

- *If they both have a finite minimum, then the minimum value is the same. Furthermore a minimizer for the inequality problem is the projection of a minimizer of the equality problem.*

Proof. The first point is established in Proposition 19.7.5. The rest follows easily from the fact that the cost c'_{n+i} associated to each slack variable x'_{n+i} is zero. □

25.2.3 Example. We go back to Example 19.6.14, this time starting from the symmetric problem:

Minimize $c_1x_1 + c_2x_2$ subject to $-x_1 - 2x_2 \geq -6$ and $\mathbf{x} \succeq \mathbf{0}$.

We multiplied the constraint by -1 to get the inequality in the right direction. Our recipe for getting an asymmetric problem is to add a slack variable $x_3 \geq 0$, make the constraint $-x_1 - 2x_2 - x_3 = -6$, and keep the same objective function. Thus in $\mathbb{R}^3$, if we think of the x_3-axis as the vertical direction, the level sets of the objective function are vertical planes. Thus we see geometrically why Theorem 25.2.2 is true in this case: the vertical lines of projection are contained in the level sets of the objective function. Finally note that there are many asymmetric problems associated to a given symmetric problem: in this example we could make the constraint $-x_1 - 2x_2 - px_3 = -6$, for any positive number p.

So for the rest of this section we forget the symmetric problem, and focus on the asymmetric problem, which we write as in Remark 25.2.1. Note that the rank of A is m.

By Theorem 19.6.13, if the feasible set F is non-empty, then a basic solution[4] exists. By reordering the variables, we can assume that the m columns where the basic solution is non-zero are the first ones, which allows us to write a vector of variables $\mathbf{x}$ as a block vector $(\mathbf{x}_B, \mathbf{x}_N)$, where $\mathbf{x}_B$ is an m-vector and $\mathbf{x}_N$ an n-vector. If $\mathbf{x}$ is basic, then it is written $(\mathbf{x}_B, \mathbf{0})$, so in block multiplication notation

$$A\mathbf{x} = [A_B, A_N] \begin{bmatrix} \mathbf{x}_B \\ \mathbf{0} \end{bmatrix} = A_B\mathbf{x}_B = \mathbf{b} \tag{25.2.4}$$

Recall from Definition 18.3.17 that a positivity constraint $x_j \geq 0$, or an inequality constraint $\mathbf{a}^i \cdot \mathbf{x} \geq b_i$, is *active* at a point $\mathbf{x}^*$ if the inequality becomes an equality at $\mathbf{x}^*$, so that $x^*_j = 0$ or $\mathbf{a}^i \cdot \mathbf{x}^* = b_i$. Otherwise it is *slack*.

Here is a summary of what we know about the asymmetric problem:

25.2.5 Theorem. *Consider the asymmetric problem 25.1.5.*

1. *Its feasible set F is non-empty if and only if $\mathbf{b}$ is in the cone C_A generated by the columns of A. If F is non-empty, F has extreme points, which are finite in number. The basic solutions of $A\mathbf{x} = \mathbf{b}$ and $\mathbf{x} \succeq \mathbf{0}$ are the extreme points of F.*

2. *If Problem 25.1.5 has a solution, namely a vector $\mathbf{x}^0$ in F such that $\mathbf{c}^T\mathbf{x}^0$ is minimal for all $\mathbf{x} \in F$, then there is a solution $\mathbf{x}^1$ that is* basic.

Proof. The statement about the non-emptiness of F is immediate from Definition 19.3.1 of the finite cone C_A. The equivalence of basic solutions and extreme points is Theorem 19.6.13, and the existence of basic solutions comes from Theorem 19.4.1, which says that a finite cone is the union of its basic subcones. The finiteness of the extreme points follows from Corollary 18.7.5.

Item (2) follows from Theorem 25.1.11 and Theorem 19.6.13. From Theorem 25.1.11, item (1), we see that the objective function $\langle \mathbf{c}, \mathbf{x} \rangle$ assumes its minimum value e at a point $\mathbf{x}^0$ at which $H_{\mathbf{c},e}$ is a supporting hyperplane for the convex set F.

We need to find a basic solution on the hyperplane $H_{\mathbf{c},e}$, meaning a solution with at most m non-zero coordinates. We get a new minimization problem by adding the (equality) constraint corresponding to $H_{\mathbf{c},e}$. There are now $m + 1$ constraints (unless the new constraint is a linear combination of the previous ones, in which case no further work is necessary) and the cost function is constant on the entire feasible set $F_1 = F \cap H_{\mathbf{c},e}$ of the new problem. So we do not need to consider the cost function, and simply apply Theorem 19.4.1 to find a basic element. □

[4] See Definition 19.6.11.

This theorem reduces the computation of the minimum of the cost function on the feasible set to a search through the finite number of extreme points of the feasible set F. However, since a system corresponding to a matrix of size $m \times n$ could have as many as $\binom{m+n}{m}$ extreme points, this is not an efficient search technique even for reasonably small numbers. The great discovery that made the search for the optimum solution efficient is the Simplex Method, which is covered in Lecture 27.

25.2.6 Example. We continue with Example 25.2.3 considered as a asymmetric problem. There are three extreme points for the feasible set, since in the constraint $x_1 + 2x_2 + x_3 = 6$ the coefficients of all the x_j are non-zero. Indeed, as we noted in the closely related Example 19.6.10, the extreme points are the three points $(6, 0, 0)$, $(0, 3, 0)$ and $(0, 0, 6)$. If the objective function is $c_1x_1 + c_2x_2 + c_3x_3$, then the minimum is the smallest of the three values $6c_1$, $3c_2$, $6c_3$.

25.2.7 Example. We transform Example 1.3.1, a symmetric minimization problem into the associated asymmetric problem, for which the matrix of constraints is

$$A = \begin{bmatrix} 1 & 2 & -1 & 0 \\ 2 & 1 & 0 & -1 \end{bmatrix} \quad \text{and } \mathbf{b} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}.$$

The cost is $\mathbf{c}^T = (1, 1, 0, 0)$.

We want to find the extreme points of the feasible set. There are potentially as many as 6 such, since we choose 2 out of 4 columns. However since there are exactly as many extreme points as for the associated symmetric problem, where there are only 3 extreme points, most of the potential candidates are not extreme points. Why not? Because they are not feasible, meaning that they are not in the first octant. For example, the last two columns of A form a basic submatrix. The basic solution associated to this submatrix is $(0, 0, -4, -5)$ which is not feasible. In the same way, columns 1 and 3 form a basic submatrix. The associated basic solution is $(5/2, 0, -3/2, 0)$, which again is not feasible.

The minimum occurs at the vertex $(2, 1)$ of the feasible region, and is equal to 3.

We will consider more complicated examples later. See Example 25.6.7.

25.3 A Necessary and Sufficient Condition for a Minimum

In this section we consider linear optimization problems where all the constraints are inequalities. This includes the symmetric problem 25.1.6. If you want, just imagine that we are dealing with that case. The techniques used apply to the general problem, as Theorem 25.3.17 shows. We use Corollary 19.5.3 of the Farkas Theorem to give a necessary and sufficient condition for a feasible point $\mathbf{x}^*$ to be a minimum. The key ideas of linear optimization appear in this section, which should be studied carefully.

Thus we consider the optimization problem:

$$\text{Minimize } \mathbf{c}^T\mathbf{x} \text{ subject to the constraints } A\mathbf{x} \succeq \mathbf{b}, \tag{25.3.1}$$

for a $m \times n$ matrix A, with no additional positivity constraints. These constraints defines the feasible set F. If the problem is the Symmetric Problem 25.1.6, then the matrix A contains the $n \times n$ identity matrix as a submatrix of rows.

25.3.2 Definition. Given a point $\mathbf{x}^*$ in F, let $I(\mathbf{x}^*)$ be the collection of indices of constraints i, $1 \leq i \leq m$, that are active at $\mathbf{x}^*$, so that $\langle \mathbf{a}^i, \mathbf{x}^* \rangle = b_i$ for $i \in I(\mathbf{x}^*)$.

25.3.3 Theorem. *Let $\mathbf{x}^*$ be a feasible point for* (25.3.2). *Then $\mathbf{x}^*$ is a minimizer for the objective function $\mathbf{c}^T\mathbf{x}$ on F if and only if there is a $\mathbf{y}^* \succeq 0$ in $\mathbb{R}^m$, with $y_i^* = 0$ for all $i \notin I(\mathbf{x}^*)$, and ${\mathbf{y}^*}^T A = \mathbf{c}^T$.*

25.3.4 Example. This example illustrates Lemma 25.3.5 below. We are in $\mathbb{R}^2$, with three constraints $x_1 \geq 0$, $x_2 \geq 0$ and $x_1 + 2x_2 \leq 1$. So the three constraint vectors, the rows of A, are $\mathbf{a}^1 = (1, 0)$, $\mathbf{a}^2 = (0, 1)$, and $\mathbf{a}^3 = (-1, -2)$, and $\mathbf{b} = (0, 0, -1)$. The minus signs in the last constraint occur because we are required to write the constraints with a '$\geq$' sign, so we multiply the constraint by -1. The first two constraints are active at the origin, our $\mathbf{x}^*$. The objective function is $\mathbf{c} \cdot \mathbf{x}$, where we leave $\mathbf{c}$ unspecified. For an arbitrary vector $\mathbf{z} \in \mathbb{R}^2$, write $\mathbf{x}^* + \epsilon\mathbf{z}$. We ask for which $\mathbf{z}$ is $\mathbf{x}^* + \epsilon\mathbf{z}$ feasible for small enough $\epsilon > 0$. Evaluating on the first constraint gives $\langle \mathbf{a}^1, \epsilon\mathbf{z} \rangle = \epsilon z_1 \geq 0$, so $z_1 \geq 0$, and on the second constraint gives $z_2 \geq 0$. Evaluating on the third constraint gives $\epsilon z_1 + \epsilon z_2 \leq 1$, and by taking ϵ small enough, no condition is imposed on $\mathbf{z}$. Thus $\mathbf{x}^* + \epsilon\mathbf{z}$ is feasible precisely when $\mathbf{z} \succeq \mathbf{0}$.

Now if $\mathbf{x}^*$ is a minimizer, then evaluating the objective function on any feasible point should give a value greater than or equal to the value of the objective function at $\mathbf{x}^*$, which in this example is 0. Evaluate the objective function at $\mathbf{x}^* + \epsilon\mathbf{z}$, where $\mathbf{z} \succeq \mathbf{0}$, so that the point is feasible. We get $\epsilon\mathbf{c} \cdot \mathbf{z}$. The question is: when is this expression positive for all possible choices of non-negative z_1 and z_2? Clearly it is necessary and sufficient that both c_1 and c_2 be non-negative. Thus the origin is a minimizer on the first quadrant for $\mathbf{c} \cdot \mathbf{x}$ if and only if the level lines of the objective function have negative slope.

Proof. Let q be the number of elements in $I(\mathbf{x}^*)$. Note that q is positive, since otherwise $\mathbf{x}^*$ is an interior point of the feasible set. An interior point cannot be a minimizer of a linear function ℓ, since ℓ has no critical points (unless it is constant).

We prove the theorem in two steps. First the $\Longrightarrow$ implication. We assume $\mathbf{x}^*$ is a minimizer. We write an expression for all the feasible points near $\mathbf{x}^*$. As usual $\mathbf{a}^i$ denotes the i-th row of A.

25.3.5 Lemma. *The vector $\mathbf{x}^* + \epsilon\mathbf{z}$ is feasible, for a small enough positive ϵ, if and only if $\mathbf{z} \cdot \mathbf{a}^i \geq 0$ for all $i \in I(\mathbf{x}^*)$.*

Proof. Take the dot product of $\mathbf{x}^* + \epsilon\mathbf{z}$ with $\mathbf{a}^i$. If the i-th constraint is not active at $\mathbf{x}^*$, by choosing ϵ small enough we can satisfy the i-th inequality for any $\mathbf{z}$. If the i-th constraint is active, the non-negativity of $\langle \mathbf{a}^i, \mathbf{z}\rangle$ is necessary and sufficient to satisfy the i-th constraint . □

By hypothesis $\mathbf{x}^*$ is a minimizer, so the value of the objective function at any nearby feasible point $\mathbf{x}^* + \epsilon\mathbf{z}$ must be at least as large as the value at $\mathbf{x}^*$, so

$$\langle \mathbf{c}, \mathbf{x}^* + \epsilon\mathbf{z}\rangle \geq \langle \mathbf{c}, \mathbf{x}^*\rangle, \text{ which implies } \langle \mathbf{c}, \mathbf{z}\rangle \geq 0.$$

Therefore by Lemma 25.3.5:

$$\langle \mathbf{a}^i, \mathbf{z}\rangle \geq 0, \text{ for all } i \in I(\mathbf{x}^*) \text{ implies } \langle \mathbf{c}, \mathbf{z}\rangle \geq 0.$$

Now we get to the key idea of the proof. Let A_I be the submatrix of A formed by the rows of A with index $i \in I(\mathbf{x}^*)$, in other words, the active constraints.

Then we see:

$$\text{For all } \mathbf{z} \text{ such that } \langle \mathbf{z}, \mathbf{a}^i\rangle \geq 0 \text{ for } i \in I(\mathbf{x}^*), \text{ then } \langle \mathbf{z}, \mathbf{c}\rangle \geq 0. \tag{25.3.6}$$

By Corollary 19.5.3 of the Farkas alternative applied to the transpose of A_I, where $\mathbf{z}$ plays the role of $\mathbf{y}$ and $\mathbf{c}$ the role of $\mathbf{b}$, the implication of (25.3.6) is equivalent to the statement that there exist non-negative numbers y_i, $i \in I(\mathbf{x}^*)$ such that

$$\mathbf{c} = \sum_{i \in I(\mathbf{x}^*)} y_i \mathbf{a}^i.$$

Now let $y_i = 0$, if $i \notin I(\mathbf{x}^*)$. So we have an m-vector $\mathbf{y} \succeq \mathbf{0}$ with $\mathbf{y}^{*T} A = \mathbf{c}^T$, and we are done.[5]

$\Longleftarrow$ Now we prove the other (easier) implication. We are given a feasible $\mathbf{x}^*$, its collection $I(\mathbf{x}^*)$ of active constraints, and a $\mathbf{y}^* \succeq 0$ in $\mathbb{R}^m$ with $y_i^* = 0$ for all

[5] We will meet the condition $y_i^* = 0$ if $i \notin I(\mathbf{x}^*)$ again later in this chapter: it goes by the name *complementary slackness*.

$i \notin I(\mathbf{x}^*)$ and $\mathbf{y}^{*T}A = \mathbf{c}^T$. Our goal is to show $\mathbf{x}^*$ is a minimizer. Let $\mathbf{x}$ be any feasible vector, and let $\mathbf{z} = \mathbf{x} - \mathbf{x}^*$.

For any $i \in I(\mathbf{x}^*)$, so that $\langle \mathbf{a}^i, \mathbf{x}^* \rangle = b_i$, we have

$$\langle \mathbf{a}^i, \mathbf{z} \rangle = \langle \mathbf{a}^i, \mathbf{x} \rangle - \langle \mathbf{a}^i, \mathbf{x}^* \rangle \geq b_i - b_i = 0.$$

so in particular, since $y_i^* \geq 0$, $y_i^* \langle \mathbf{a}^i, \mathbf{z} \rangle \geq 0$.

For $i \notin I(\mathbf{x}^*)$, we cannot control the sign of $\langle \mathbf{a}^i, \mathbf{z} \rangle$. But since by hypothesis the i-th coordinate of $\mathbf{y}^*$ is 0, we have $y_i^* \langle \mathbf{a}^i, \mathbf{z} \rangle = 0$.

Thus altogether,

$$\mathbf{c}^T \mathbf{z} = \mathbf{y}^{*T} A\mathbf{z} \geq 0$$

which in turn implies that $\mathbf{c}^T\mathbf{x} = \mathbf{c}^T\mathbf{z} + \mathbf{c}^T\mathbf{x}^* \geq \mathbf{c}^T\mathbf{x}^*$, so that $\mathbf{x}^*$ is a minimizer as required. □

Thus the existence of a finite solution to any linear minimization problem without equalities is equivalent to the non-emptiness of a set that looks like that the feasible set of another optimization problem. Indeed, since $\mathbf{y}^*$ is non-negative, the equation $A^T\mathbf{y}^* = \mathbf{c}$ in the theorem says that $\mathbf{c}$ is in the cone C generated by the rows of A, so our conclusion is that C is non-empty.

Which optimization problem? The following corollary of Theorem 25.3.3 will help us decide.

25.3.7 Corollary. *With the notation of the theorem, we have*

$$\mathbf{y}^{*T}\mathbf{b} = \mathbf{y}^{*T}A\mathbf{x}^* = \mathbf{c}^T\mathbf{x}^*. \tag{25.3.8}$$

Proof. By putting the constraints that are active at $\mathbf{x}^*$ first, we can decompose A and $\mathbf{y}^*$ into two block submatrices of compatible size for multiplication (see §6.10),

$$A = \begin{bmatrix} A_I \\ A_N \end{bmatrix} \text{ and } \mathbf{y}^* = \begin{bmatrix} \mathbf{y}_I^* \\ \mathbf{0} \end{bmatrix}$$

corresponding to the active constraints and the non-active constraints at $\mathbf{x}^*$. We do block multiplication $\mathbf{y}^{*T}A$. Because of the zeroes in $\mathbf{y}^*$, we get $\mathbf{y}^{*T}A = \mathbf{y}_I^{*T}A_I$, and since A_I corresponds to the active constraints at $\mathbf{x}^*$, multiplying by $\mathbf{x}^*$, we get $A_I\mathbf{x}^* = \mathbf{b}$. This gets us the left-hand equality in 25.3.8. Theorem 25.3.3 tells us that $\mathbf{y}^{*T}A = \mathbf{c}^T$, so multiplying by $\mathbf{x}$ on the right and using the first conclusion, we get $\mathbf{y}^{*T}\mathbf{b} = \mathbf{y}^{*T}A\mathbf{x}^* = \mathbf{c}^T\mathbf{x}^*$ as required, □

Notice the symmetry. It suggests that the dual objective function is $\mathbf{y} \cdot \mathbf{b}$. We have two collections of non-negative variables $\mathbf{x}$ and $\mathbf{y}$ linked by the matrix A. If we take the transpose of 25.3.8, the relationship is preserved. In the course of taking the transpose the roles of $\mathbf{x}$ and $\mathbf{y}$ are exchanged, as are those of $\mathbf{b}$ and $\mathbf{c}$. This suggests that there is a linear optimization problem in $\mathbb{R}^m$ to which Theorem 25.3.3 applies, and that yields our original problem back. This is indeed the case, and this is the theory of duality of linear optimization, which we will study further in §25.4 and §25.5 . Duality associates to any linear optimization problem (called the primal) a second linear optimization problem called the dual. Then the dual of the dual gives back the primal. The importance of the dual is that the solutions of the dual give information about the original problem. The dual problem often has a meaningful interpretation that sheds light on the original problem. We will see this in the applications in Lecture 26.

For time being, let us just write down the dual problem in the special case of the symmetric problem, so that A contains the $n \times n$ identity matrix with corresponding entries of $\mathbf{b}$ equal to 0. We revert to our old notation, so the identity matrix has been removed from A.

25.3.9 Definition. The dual of the symmetric minimization problem 25.1.6 is the symmetric minimization problem:

$$\text{Minimize } -\mathbf{y}^T\mathbf{b} \text{ subject to } -A^T\mathbf{y} \succeq -\mathbf{c}, \text{ and } \mathbf{y} \succeq \mathbf{0}, \tag{25.3.10}$$

or, which is equivalent (this is the form we will use most often):

$$\text{Maximize } \mathbf{y}^T\mathbf{b} \text{ subject to } \mathbf{y}^T A \preceq \mathbf{c}^T, \text{ and } \mathbf{y} \succeq \mathbf{0}. \tag{25.3.11}$$

25.3.12 Remark. Note that the point $\mathbf{y}^*$ in the statement of Theorem 25.3.3 satisfies the constraints of the feasible set of the dual. This shows that if the primal problem has a finite minimum, then the feasible set of the dual problem is non-empty.

25.3.13 Exercise. Apply Theorem 25.3.3 to the symmetric dual problem 25.3.9, and show that you get Problem 25.1.6 back.

This exercise explains why this is called duality: if you apply it twice, you get the original problem back. Finally, this explains why we have called problem 25.1.6 the symmetric problem: Its dual is of the same form as itself.

25.3.14 Algorithm. Here is the algorithm for passing from the primal to the dual, or from the dual to the primal in the case of the symmetric minimization problem.

- Replace a minimization problem by a maximization problem, and vice-versa;
- Act on the constraint matrix A by left multiplication ($\mathbf{y}^T A$) instead of right multiplication ($A\mathbf{x}$), and vice-versa. This interchanges n and m. Another way of saying this is: replace A by A^T.

- Make the optimization vector $\mathbf{c}$ the constraint vector $\mathbf{b}$, and vice-versa.
- Reversal the direction of the inequality in the constraint equation: the '$\geq$' in $A\mathbf{x} \geq \mathbf{b}$ replaced by '$\leq$' in $\mathbf{y}^T A \leq \mathbf{c}^T$.

25.3.15 Exercise. In $\mathbb{R}^2$, solve the optimization problem: maximize the function $f(y_1, y_2) = y_2$, subject to $\mathbf{y} \succeq \mathbf{0}$ and $y_1 + y_2 \leq 3$. Formulate and solve the associated primal problem (which is the dual of the dual).

25.3.16 Exercise. Write down the algorithm for passing to the dual when it is written as (25.3.10).

Consider the following generalization of Theorem 25.3.3, keeping the same notation as before, but adding a $r \times n$ matrix B and a r-vector $\mathbf{d}$ of equality constraints.

25.3.17 Theorem. *Let $\mathbf{x}^*$ be a feasible point for the problem:*
Minimize $\mathbf{c}^T\mathbf{x}$ subject to $A\mathbf{x} \succeq \mathbf{b}$ and $B\mathbf{x} = \mathbf{d}$.
Then $\mathbf{x}^*$ is a minimizer for the problem if and only if there is a
$\mathbf{y}^* \succeq 0$ in $\mathbb{R}^m$ with $y_i^* = 0$ for all $i \notin I(\mathbf{x}^*)$,
$\mathbf{w}^*$ in $\mathbb{R}^r$,
such that $\mathbf{y}^{*T} A + \mathbf{w}^{*T} B = \mathbf{c}^T$.

Note that there is no positivity constraint on $\mathbf{w}$.

25.3.18 Exercise. Prove this theorem.

We also have the analog of Corollary 25.3.7.

25.3.19 Corollary. *With the notation of the theorem, we have, using block matrix notation:*

$$\begin{bmatrix} \mathbf{y}^{*T} & \mathbf{w}^{*T} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{d} \end{bmatrix} = \begin{bmatrix} \mathbf{y}^{*T} & \mathbf{w}^{*T} \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} \mathbf{x}^* = \mathbf{c}^T\mathbf{x}^*. \tag{25.3.20}$$

25.3.21 Exercise. Prove this corollary.

25.4 The Duality Principle

We have already examined duality for the symmetric problem. Here we look at it for the asymmetric problem and then we write down the general duality, and finally we explain why these problems are paired. We also extend Corollary 25.3.19 to the weak duality theorem below.

25.4.1 Example (Dual of the Asymmetric Minimization Problem)**.** We call problem 25.1.5 the *primal*. Its dual is

$$\text{Maximize } \mathbf{y}^T\mathbf{b} \text{ subject to } \mathbf{y}^T A \leq \mathbf{c}^T \tag{25.4.2}$$

There is no positivity constraint on the m-vector $\mathbf{y}$.

25.4.3 Remark. As in the case of the symmetric optimization problem (see Remark 25.3.12), in the case of the asymmetric problem the m-vector $\mathbf{w}^*$ in the statement of Theorem 25.3.17 satisfies the constraints of the feasible set of its dual. The matrix B and the vector $\mathbf{d}$ in the statement of Theorem 25.3.17 correspond to A and $\mathbf{b}$ in (25.4.2). As in the symmetric case, this shows that if the primal problem has a finite minimum, then the feasible set of the dual problem is non-empty.

Next we write down the dual problem for the general minimization problem 25.1.7, after writing down the algorithm used to get it.

25.4.4 Algorithm. For the general algorithm we modify Algorithm 25.3.14 by adding two items and by modifying the last item.

- Interchange the indexing sets $\mathcal{I}$ and $\mathcal{J}$.
- Require that $y_i \geq 0$ for $i \in \mathcal{I}$.
- Reversal the direction of the inequality in the constraint equations: the '$\geq$' in $\sum_{j=1}^n a_{ij}x_j \geq b_i$, for $i \in \mathcal{I}$ is replaced by '$\leq$' in $\sum_{i=1}^m y_i a_{ij} \leq \mathbf{c}_j$, for $j \in \mathcal{J}$.

25.4.5 Definition. The dual of Problem 25.1.7 is
Maximize $\mathbf{y}^T\mathbf{b}$ subject to the constraints

- $y_i \geq 0$ when $i \in \mathcal{I}$,
- $\sum y_i a_{ij} \leq c_j$ when $j \in \mathcal{J}$,
- $\sum y_i a_{ij} = c_j$ when $j \notin \mathcal{J}$.

25.4.6 Exercise. For the general linear optimization problem, show that Theorem 25.3.17 proves that there is a vector satisfying the constraints of the dual problem if the primal problem has a finite minimum.

25.4.7 Remark. Why are the primal and dual paired? The duality is suggested by the theorems and corollaries of §25.3. Then there is the statement, now checked in all cases, that if the primal has a finite minimum (so that we are in case (f) of Theorem 25.1.10), then the feasible set of the dual in non-empty (so that the dual is not in case (e)). The next step is given by the following theorem.

25.4.8 Theorem (Weak Duality). *Assume that the feasible set F_p of the primal minimization problem and the feasible set F_d of the dual maximization problem are non-empty. Pick any $\mathbf{x} \in F_p$ and $\mathbf{y} \in F_d$. Then*

$$\mathbf{y}^T\mathbf{b} \leq \mathbf{c}^T\mathbf{x}. \tag{25.4.9}$$

Thus the minimum of the primal is bounded by the maximum of the dual, namely

$$\max_{\mathbf{y}\in F_d} \mathbf{y}^T\mathbf{b} \leq \min_{\mathbf{x}\in F_p} \mathbf{c}^T\mathbf{x} \tag{25.4.10}$$

In particular if one can find feasible $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{y}^T\mathbf{b} = \mathbf{c}^T\mathbf{x}$, then $\mathbf{x}$ is a minimizer and $\mathbf{y}$ a maximizer for their respective problems, and the minimum of the primal is the maximum of the dual.

Proof. We write $\mathbf{y}^T A\mathbf{x}$ in two different ways. First as

$$\mathbf{y}^T A\mathbf{x} = \mathbf{y}^T(A\mathbf{x}) = \sum_{i=1}^{m} y_i\big(\sum_{j=1}^{n} a_{ij}x_j\big)$$

When $i \in \mathcal{I}$, $y_i \geq 0$ and $\sum_{j=1}^n a_{ij}x_j \geq b_i$, so $y_i\big(\sum_{j=1}^n a_{ij}x_j\big) \geq y_ib_i$.

When $i \notin \mathcal{I}$, $\sum_{j=1}^n a_{ij}x_j = b_i$, so $y_i\big(\sum_{j=1}^n a_{ij}x_j\big) = y_ib_i$.

Putting both cases together, we get

$$\mathbf{y}^T A\mathbf{x} \geq \mathbf{y}^T\mathbf{b}. \tag{25.4.11}$$

Now perform the matrix multiplications is the opposite order:

$$\mathbf{y}^T A\mathbf{x} = (\mathbf{y}^T A)\mathbf{x} = \sum_{j=1}^{n} \big(\sum_{i=1}^{m} y_ia_{ij}\big)x_j$$

When $j \in \mathcal{J}$, $x_j \geq 0$ and $\sum_{i=1}^m y_ia_{ij} \leq c_j$, so $\big(\sum_{i=1}^m y_ia_{ij}\big)x_j \leq c_jx_j$. When $j \notin \mathcal{J}$, $\sum_{i=1}^m y_ia_{ij} = c_j$, so $\big(\sum_{i=1}^m y_ia_{ij}\big)x_j = c_jx_j$. Putting both cases together, we get

$$\mathbf{y}^T A\mathbf{x} \leq \mathbf{c}^T\mathbf{x}. \tag{25.4.12}$$

Combining (25.4.11) and (25.4.12) gives the desired conclusion. □

25.4.13 Corollary. *Combining Weak Duality with Remark 25.4.7, we see that if the primal has a finite minimum (so we are in case (f)), then the dual has a finite maximum. Furthermore the point in the dual feasible set found in Theorem 25.3.3 and Theorem 25.3.17 is a maximizer for the dual problem.*

Proof. The first statement is immediate, since weak duality says that the maximum of the dual is less than the minimum of the primal. The second follows from the computation in Corollary 25.3.19, which is just a special case of weak duality applied to the minimizer of the primal and a point in the dual that satisfies complementary slackness: see §25.6. □

25.5 The Duality Theorem

We defined the notation of the primal problem and its dual problem in §25.4. Using the notation of Theorem 25.1.10, we denote the possible outcomes for the primal (e_p), (u_p), (f_p), and for the dual (e_d), (u_d), (f_d).

For two unrelated problems there are 9 possibles outcomes. But for the primal-dual pair there are only 4 outcomes, as the duality theorem states.

25.5.1 Theorem (The Duality Theorem)**.** *We assume that the primal problem is the general problem 25.1.7, and its dual is therefore 25.4.5. Then we are in one of four possible cases*

1. *If the primal or the dual problem has a bounded solution, then so does the other one. If $\mathbf{x}^*$ denotes a minimizer of the primal and $\mathbf{y}^*$ a maximizer of the dual, then*
$$(\mathbf{y}^*)^T\mathbf{b} = (\mathbf{y}^*)^T A\mathbf{x}^* = \mathbf{c}^T\mathbf{x}^* \tag{25.5.2}$$
so the minimum of $\mathbf{c}^T\mathbf{x}$ is equal to the maximum of $\mathbf{y}^T\mathbf{b}$, as $\mathbf{x}$ and $\mathbf{y}$ range over the feasible sets of the primal and the dual, respectively. This says that outcome (f_p, f_d) can occur and that (f_p, e_d), (f_p, u_d), (e_p, f_d) and (u_p, f_d) cannot occur.

2. *The feasible sets for both problems are empty. This is outcome (e_p, e_d).*

3. *The feasible set for the primal is empty, and the feasible set of the dual is unbounded and allows for arbitrary large values of the objective function of the dual, so that there is no finite maximum. This is outcome (e_p, u_d).*

4. *The feasible set for the dual is empty, and the feasible set of the primal is unbounded and allows for arbitrary small values of the objective function of the primal, so that there is no finite minimum. This is oucome (u_p, e_d).*

Note that cases (3) and (4) rule out outcome (u_p, u_d), so all possible outcomes are accounted for.

Case (1) is the only case with a solution for both the primal and dual, and in actual problems one arrives there.

25.5.3 Example. We minimize $3x_1 + 5x_2$, subject to the constraints $\mathbf{x} \succeq \mathbf{0}$ and the constraint $2x_1 + x_2 \geq 4$. Thus $n = 2$ and $m = 1$. A quick sketch will convince you that the minimum occurs at $\mathbf{x}^T = (2, 0)$, so that the minimum value $\mathbf{c}^T\mathbf{x}$ is 6. What is the dual problem? A is the 1×2 matrix $(2, 1)$, b is the number 4, y is a number too, and $\mathbf{c}^T = (3, 5)$. So the dual problem is to maximize $4y$ subject to the constraint $(2y, y) \leq (3, 5)$. Only the first of these two constraints is active, and the maximum occurs when $2y = 3$, so that the maximal cost in the dual problem is $4 \cdot \frac{3}{2} = 6$, confirming the duality theorem in this case.

Proof of the duality theorem. If the feasible sets of both primal and dual are empty (case 2), nothing more need be said. This case can occur.

Next suppose that the objective function of the primal problem takes values $\mathbf{c}^T\mathbf{x}$ that get arbitrarily negative. Then weak duality says that the dual objective function's values $\mathbf{y}^T\mathbf{b}$ are smaller than any of the primal values, so there cannot be any: thus the feasible set of the dual problem must be empty. We are in case (4) of the duality theorem.

Reversing the roles of the primal and the dual, we see by a similar argument that if the objective function of the dual problem takes values $\mathbf{y}^T\mathbf{b}$ that get arbitrarily large, the feasible set of the primal is empty: we are in case (3).

Next we assume that both feasible sets are non-empty. By weak duality, if one of the two problems has a finite optimum, then so does the other.

So all that is left to prove the full duality theorem is that $\mathbf{y}^* \cdot \mathbf{b} \geq \mathbf{c} \cdot \mathbf{x}^*$, the reverse inequality to the one in the weak duality theorem. We have already done this in Corollaries 25.3.7 and 25.3.19, since we have found a point $\mathbf{y}^*$ in the dual feasible set whose objective value $\mathbf{y}^* \cdot \mathbf{b}$ is equal to the minimum value of the primal objective function. Indeed, starting at a minimizer $\mathbf{x}^*$ for the primal, we found a feasible $\mathbf{y}^*$ for the dual with $\mathbf{y}^* \cdot \mathbf{b} = \mathbf{c} \cdot \mathbf{x}^*$.

□

We can say more for a basic minimizer $\mathbf{x}$ of the Asymmetric Problem (25.1.5). Permute the columns of the matrix A so that the columns corresponding to the basic variables come first. Then A is written in block matrix notation as $[B, N]$, where B is invertible, and $\mathbf{x}$ is written as $(\mathbf{x}_B, \mathbf{0})$. Let $\mathbf{c}_B$ be the part of $\mathbf{c}$ matching the columns of B. Then $A\mathbf{x} = B\mathbf{x}_B$, so (25.5.2) becomes

$$(\mathbf{y}^0)^T\mathbf{b} = (\mathbf{y}^0)^T B\mathbf{x}_B = \mathbf{c}_B^T\mathbf{x}_B \tag{25.5.4}$$

where $\mathbf{y}^0$ is a maximizer for the dual.

25.5.5 Example. We continue with Example 1.3.1, a symmetric minimization problem with

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \quad \text{and } \mathbf{b} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}.$$

We write the dual problem when the cost is $\mathbf{c}^T = (1,1)$. The minimum for the primal problem in this case occurs at the vertex $(2,1)$ of the feasible region, and is equal to 3.

The dual problem is to maximize $4y_1+5y_2$, subject to $\mathbf{y} \geq 0$ and $\mathbf{y}^T A \leq (1,1)$. The three vertices of the feasible set of the dual are $(0, \frac{1}{2})$, $(\frac{1}{3}, \frac{1}{3})$, and $(\frac{1}{2}, 0)$. Check the value of $4y_1 + 5y_2$ at each vertex, we get 2.5, 3 and 2, so that the maximizer is $(\frac{1}{3}, \frac{1}{3})$ and the optimal values for the primal problem and the dual problem agree.

25.6 The Equilibrium Theorem

We give two versions of the Equilibrium Theorem: first for the symmetric problem, and then for the asymmetric problem. It introduces the notion of complementary slackness, which is very useful for non-linear optimization, where it is also known as the Kuhn-Tucker condition, as we shall learn in Lecture 31.

We assume that both the primal and dual problems have finite solutions, so we are in case (1) of the duality theorem.

The idea of the Equilibrium Theorem is easy to state informally: if $\mathbf{x}$ is a minimum for the primal problem, then there exists a maximum for the dual problem satisfying complementary slackness, meaning that for any inequality constraint of the primal (resp. dual) that is slack, the corresponding inequality constraint for the dual (resp. primal) is active. In fact, the theorems of §25.3 start us off by constructing for each minimizer a dual maximizer satisfying complementary slackness.

For the symmetric problem, this gives:

25.6.1 Theorem (The Equilibrium Theorem in the symmetric case)**.** *Let* $\mathbf{x}$ *be a minimizer in the feasible set of the symmetric primal problem 25.1.6. Then any maximizer* $\mathbf{y}$ *in the feasible set of the symmetric dual problem 25.3.11, satisfies both pairs of* complementary slackness *conditions:*
For any i, $1 \leq i \leq m$,

- *If* $(A\mathbf{x})_i > b_i$, *then* $y_i = 0$.
- *If* $y_i > 0$, *then* $(A\mathbf{x})_i = b_i$.

For any j, $1 \leq j \leq n$,

- *If* $(\mathbf{y}^T A)_j < c_j$, *then* $x_j = 0$.
- *If* $x_j > 0$, *then* $(\mathbf{y}^T A)_j = c_j$.

Conversely, if $\mathbf{x}$ *and* $\mathbf{y}$ *satisfy these conditions, they are both optimizers.*

Proof. By the Duality Theorem, we know that $\mathbf{x}$ and $\mathbf{y}$ are both optimizers if and only if

$$\mathbf{y}^T\mathbf{b} = \mathbf{y}^T A\mathbf{x} = \mathbf{c}^T\mathbf{x}$$

For arbitrary feasible $\mathbf{x}$ and $\mathbf{y}$ we have

$$\mathbf{y}^T\mathbf{b} \leq \mathbf{y}^T A\mathbf{x} \leq \mathbf{c}^T\mathbf{x} \tag{25.6.2}$$

Writing this out in coordinates, the left-hand inequality gives

$$\sum_{i=1}^{m} y_i b_i \leq \sum_{i=1}^{m} y_i (A\mathbf{x})_i,$$

where $(A\mathbf{x})_i$ means the i-th row of the matrix product $A\mathbf{x}$. This can be rewritten

$$\sum_{i=1}^{m} y_i \big((A\mathbf{x})_i - b_i\big) \geq 0 \tag{25.6.3}$$

Our constraints say that $y_i \geq 0$ and that $(A\mathbf{x})_i - b_i \geq 0$, so that (25.6.3) is the sum of a product of non-negative numbers. Thus for the sum to be zero each term in the sum must be zero. If $b_i < (A\mathbf{x})_i$, the only way this strict inequality can be transformed into an equality after multiplication by the vector $\mathbf{y}$ with non-negative coordinates, is if the corresponding coordinate of $\mathbf{y}$ is 0, namely $y_i = 0$. Going the other way, if $y_i > 0$ for some i, then the nonnegative quantity $(A\mathbf{x})_i - b_i$ must be 0. So the complementary slackness equations must be satisfied.

Similarly, transforming the right-hand inequality in (25.6.2), we get

$$\sum_{j=1}^{n} \big(c_j - (\mathbf{y}^T A)_j\big) x_j \geq 0$$

with all terms non-negative. If $(\mathbf{y}^T A)_j < c_j$, we only get equality after multiplication by $\mathbf{x}$ if $x_j = 0$ and conversely. So again, complementary slackness must be satisfied. □

25.6.4 Example. We illustrate the theorem by Example 25.5.3. There we found that $\mathbf{x}^T = (2, 0)$ is the unique minimum for the primal, and $y = 3/2$ the unique maximum of the dual. The complementary slackness conditions must be satisfied. Since $m = 1$, the only value for i is 1, and $(A\mathbf{x})_1 = b_1$, since $A = (2, 1)$ and $b = 4$. Nothing to verify on this side. Since $n = 2$, there are two indices to check on the other side, and when $j = 2$ we see that $(y^T A)_2 = 3/2 < c_2 = 5$, which, according to the Equilibrium Theorem implies $x_2 = 0$, which is indeed the case. Next, at the minimum $x_1 > 0$, so we must have $(y^T A)_1 = c_1$, or $2y = 3$, as is the case.

25.6.5 Example. We continue with Example 25.5.5, setting the cost vector to $\mathbf{c}^T = (3,1)$, and keeping A and $\mathbf{b}$ as before. The primal feasible region has the vertices $(0,5)$, $(2,1)$, $(4,0)$, so, checking the cost at the three vertices, we see that the minimum cost occurs at $(0,5)$ and is 5.

The dual problem is to maximize $4y_1+5y_2$ on the set $\mathbf{y} \geq 0$, with $y_1+2y_2 \leq 3$, $2y_1 + y_2 \leq 1$. In the first quadrant the second constraint is always more stringent than the first (check this), and the vertices of the dual feasible set are $(0.5,0)$ and $(0,1)$. The maximum occurs at $(0,1)$ and is 5. Thus the duality theorem is satisfied.

We check that complementary slackness holds. On the primal, the condition given by the first row $x_1+2x_2 \geq 4$ of $Ax = b$ is slack, so complementary slackness says that the solution for the dual must have its first coordinate $y_1 = 0$, which is indeed the case. $x_2 > 0$ is slack, so we must have $2y_1 + y_2 = 1$, as is the case. Going in the other direction, we see that the condition on the dual problem given by the first column of A is slack at the maximum (in fact, it is slack everywhere), so that on the solution for the primal problem we must have $x_1 = 0$. Again, we see that this is the case.

25.6.6 Example. Continuing with Example 25.6.5, we now let the cost vector be $\mathbf{c}^T = (1,4)$. The minimum for the primal problem occurs at $\mathbf{x} = (4,0)$ and it is 4. The constraints for the dual problem are $y_1+2y_2 \leq 1$ and $2y_1+y_2 \leq 4$. The second condition is always slack in the first quadrant, and the vertices of the feasible set are $(1,0)$ and $(0,1/2)$. Since the objective function for the dual problem is $\mathbf{b} = (4,5)$, the maximum occurs at $\mathbf{y} = (1,0)$ and is also 4, as per the duality theorem. The second condition $2x_1 + x_2 \leq 5$ of the primal is slack at the minimum $\mathbf{x} = (4,0)$, so by complementary slackness we must have $y_2 = 0$ at the maximum of the dual, and indeed we do. The second condition $2y_1 + y_2 \geq 4$ is slack everywhere in the first quadrant, so it is certainly slack at the maximum for the dual problem, so by complementary slackness we must have $x_2 = 0$ at the minimum for the primal, and indeed we do. Next the first positivity constraint $x_1 > 0$ is slack, so we must have $(\mathbf{y}^T A)_1 = y_1 + 2y_2 = c_1 = 1$ at the maximum, and we do.

25.6.7 Example. We now work out a more complicated example. We take

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 2 & 2 & 4 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} 5 \\ 6 \\ 11 \end{bmatrix}$$

We first study the feasible set for $\mathbf{x} \succeq \mathbf{0}$, $A\mathbf{x} \succeq \mathbf{b}$. It is the intersection of the three half-spaces given by $A\mathbf{x} \succeq \mathbf{b}$ with the first quadrant. Thus we have 6 boundary hyperplanes, of which we need to choose 3 to get a vertex. We could, in the worst situation, get $\binom{6}{3}$, namely 20, vertices.

How many vertices do we actually have? If we label the condition given by setting $(A\mathbf{x})_i = b_i$ by (i), and the condition given by setting $x_j = 0$ by $[j]$, the 20 potential vertices can be written

$$\begin{array}{rrr}
(123) & & \\
(1)[12] & (1)[13] & (1)[23] \\
(2)[12] & (2)[13] & (2)[23] \\
(3)[12] & (3)[13] & (3)[23] \\
[1](12) & [1](13) & [1](23) \\
[2](12) & [2](13) & [2](23) \\
[3](12) & [3](13) & [3](23) \\
[123] & &
\end{array}$$

We first find the point where the three hyperplanes $(A\mathbf{x})_i = b_i$, $1 \leq i \leq 3$ intersect. It is the potential vertex written (123) above. A linear algebra computation shows it is the point with coordinates

$$(4, 1/2, 1/2)$$

which is in the first quadrant, so it satisfies all the other constraints and is feasible. Next we intersect a fixed hyperplane $(A\mathbf{x})_i = b_i$ with any two of the boundary hyperplanes of the first quadrant $x_j = 0$. There are three ways of doing this for a fixed i, so we get 9 potential vertices. For $i = 1$ we get the three potential vertices

$$(5, 0, 0),\ (0, 5, 0),\ (0, 0, 5).$$

To be feasible, they must satisfy the inequalities given by the second and the third rows of $A\mathbf{x} \succeq \mathbf{b}$. The first two fail, but the last one is feasible and therefore a vertex. For $i = 2$ we get the three potential vertices

$$(6, 0, 0),\ (0, 3, 0),\ (0, 0, 3).$$

The first one is feasible, and the last two are not. Finally for $i = 3$ we get the three potential vertices

$$(5.5, 0, 0),\ (0, 5.5, 0),\ (0, 0, 2.75).$$

Here the second one only is feasible. The pattern here is that we have three points on each coordinate axis, and only the one with largest non-zero coordinate is feasible. In terms of our notation these are $(1)[12]$, $(2)[23]$ and $(3)[13]$.

Next we turn to vertices that are the intersection of one equation $x_j = 0$ with two equations $(A\mathbf{x})_i = b_i$. For each choice of j we get one new vertex. They are $(5, 1/2, 0)$, $(4, 0, 1)$ and $(0, 4.5, 0.5)$, as you should check. These are [3](23), [2](12) and [1](13).

The last potential vertex is the intersection of all the positivity constraints, namely the origin. It clearly is not feasible so we have found all seven extreme points of the polyhedron.

Given an objective function, say $x_1 + x_2 + x_3$, so $\mathbf{c} = (1, 1, 1)$, we find the minimum by just testing the function on all seven vertices. The minimum value is 5, and it is attained at three vertices: $(0, 0, 5)$, $(4, 0, 1)$ and $(0, 4.5, 0.5)$, namely (1)[12], [2](12) and [1](13).

Given this objective function we can write the dual problem, and in particular the dual feasible set, given in $\mathbb{R}^3$ with coordinates y_1, y_2, y_3 by the positivity conditions $y_i \geq 0$ and the constraints $\mathbf{y}^T A \preceq (1, 1, 1)$. The dual feasible polyhedron is very simple: it has just four vertices: $(0, 0, 0)$, $(1, 0, 0)$, $(0, 1/2, 0)$, and $(0, 0, 1/4)$, as an analysis similar to (but much easier than) the one above shows. It is compact and the dual objective function $5y_1 + 6y_2 + 11y_3$ assumes its maximum 5 at the vertex $(1, 0, 0)$. The only slack equation for $(1, 0, 0)$ is the constraint $y_1 \geq 0$: the remaining five constraints are active. The constraint of the primal corresponding to this slack constraint for the dual is $x_1 + x_2 + x_3 \geq 5$. As required by complementary slackness, it is active for the three basic solutions of the primal.

25.6.8 Remark. In order to preserve geometric intuition in Example 25.6.7, we did not pass to the associated asymmetric problem by adding three slack variables and therefore transforming the constraint matrix A into a matrix A' with 3 rows and 6 columns:

$$A' = \begin{bmatrix} 1 & 1 & 1 & -1 & 0 & 0 \\ 1 & 2 & 2 & 0 & -1 & 0 \\ 2 & 2 & 4 & 0 & 0 & -1 \end{bmatrix}$$

Let's connect the new $(1)[23]$ notation used above into the language of basic submatrices of A', for any symmetric problem with n variables and m constraints. To specify an extreme point, we need n linear equations. A certain number $k \leq m$ of them assert that k of the constraints $\mathbf{a}^i \cdot \mathbf{x} \geq b_i$ are equalities, so that the associated slack variables z_i are 0; The remaining $n - k$ of them assert that nonnegativity equations $x_j \geq 0$ are equalities, so k of the x_j are 0. Thus in general the notation $(i_1 \dots i_k)[j_1 \dots j_{n-k}]$ gives first the indices of the slack variables z_i that are 0, followed by the indices of the original variable x_j that are 0. Thus the indices of the basic submatrix corresponding to this extreme point are all the indices that do not appear on the list: as required there are m of them. Returning to our example, the notation $(1)[23]$ means that the first slack variable is 0, and x_2 and x_3 are 0. Thus the basic submatrix associated to this extreme point consists of colums 1, 5 and 6.

To summarize: in this example, we studied extreme points by studying the n equations that vanish there, which in the method of basic submatrices we consider the m variables that do not vanish. The remark shows how one passes from one point of view to the other.

25.6.9 Exercise. In Example 25.6.7, modify $\mathbf{c} = (1, 1, 1)$ slightly so that the primal problem has only only one basic solution. Then find all the vertices of the dual feasible set, and the maximizer dual vertex. Check complementary slackness.

25.6.10 Exercise. If all the entries of A, $\mathbf{b}$ and $\mathbf{c}$ are positive in the symmetric primal problem, prove that both the primal and the dual problems have finite optimal solutions.

Hint: By the duality theorem, all you need to do is show that the feasible set of the primal is non-empty and that the objective function is bounded on it.

Finally we turn to the asymmetric form of the equilibrium theorem.

25.6.11 Theorem (The Equilibrium Theorem in the asymmetric case)**.** *Let* $\mathbf{x}$ *be a minimizer in the feasible set of the asymmetric primal problem 25.4.1. Then there exists a* $\mathbf{y}$ *in the feasible set of the asymmetric dual problem 25.4.1, satisfying both pairs of* complementary slackness *conditions:*

For any j, $1 \leq j \leq n$,

- *If* $(\mathbf{y}^T A)_j < c_j$, *then* $x_j = 0$.
- *If* $x_j > 0$, *then* $(\mathbf{y}^T A)_j = c_j$.

$\mathbf{y}$ *is then a maximizer for the dual problem. Conversely, if* $\mathbf{x}$ *and* $\mathbf{y}$ *satisfy these conditions, they are both optimizers.*

Since there are fewer inequalities, there are fewer equations in complementary slackness.

Proof. By the duality theorem, since the primal problem has a solution, the dual does too, so that there is a $\mathbf{y}$ in the feasible set $\mathbf{y}^T A \leq \mathbf{c}^T$ that maximizes $\mathbf{y}^T\mathbf{b}$. Furthermore, by duality we have $\mathbf{c}^T\mathbf{x} = \mathbf{y}^T A\mathbf{x} = \mathbf{y}^T\mathbf{b}$ as before. So

$$0 = \mathbf{c}^T\mathbf{x} - \mathbf{y}^T A\mathbf{x} = \sum_{j=1}^{n} \big(c_j - \sum_{i=1}^{m} y_i a_{ij}\big) x_j \tag{25.6.12}$$

Since $\mathbf{y}$ is feasible, $\mathbf{y}^T A \leq \mathbf{c}^T$. This, written out in equations, is just

$$c_j - \sum_{i=1}^{m} y_i a_{ij} \geq 0 \;\text{ for all } j, \; 1 \leq j \leq n \tag{25.6.13}$$

The x_j are all non-negative, so each term in the sum of the right-hand of (25.6.12) is non-negative. But it sums to 0, so each term individually must be 0, so that for each j, either $x_j = 0$ or $c_j - \sum_{i=1}^{m} y_i a_{ij} = 0$. □

25.7 Another Look at the Equilibrium Theorem

Here we see how the Duality Theorem and the Equilibrium Theorem guide us in computing the solution of a linear optimization problem. As always when one computes, we restrict to the asymmetric problem 25.1.5. We assume that A is a $m \times n$ matrix of maximal rank m.

Pick any basic submatrix B of A. Thus B is an invertible $m \times m$ matrix. For convenience in explaining what follows, we permute the columns of A so that A can be written in block notation as $[B, N]$, where N is a $m \times (n - m)$ matrix. Also write $\mathbf{c} = (\mathbf{c}_B, \mathbf{c}_N)$ and $\mathbf{x} = (\mathbf{x}_B, \mathbf{x}_N)$, so the index B denotes the first m coordinates, and the index N the last $n - m$ coordinates, of each vector.

Define the *equilibrium point* $\mathbf{x}$ associated to B by

$$\mathbf{x}_B = B^{-1}\mathbf{b} \text{ , and } \mathbf{x}_N = \mathbf{0}_{n-m}. \tag{25.7.1}$$

Then $A\mathbf{x} = [B, N](\mathbf{x}_B, \mathbf{x}_N) = BB^{-1}\mathbf{b} + N\mathbf{0}_{n-m} = \mathbf{b}$, and $\mathbf{c} \cdot \mathbf{x} = \mathbf{c}_B \cdot \mathbf{x}_B$. Thus $\mathbf{x}$ satisfies the equality constraints, and it is basic, in the sense that it has $n - m$ coordinates that are zero, but it is not necessarily feasible, which would require that $\mathbf{x}_B \succeq \mathbf{0}$.

Define the *dual equilibrium point* $\mathbf{y}$ associated to B by

$$\mathbf{y}^T = \mathbf{c}_B^T B^{-1}. \tag{25.7.2}$$

The vector $\mathbf{y}$ is not necessarily feasible for the dual problem, which requires that $\mathbf{y}^T A \preceq \mathbf{c}^T$. We have, computing with block matrices again,

$$\mathbf{y}^T A = \mathbf{c}_B^T B^{-1} A = \mathbf{c}_B^T B^{-1}[B, N] = \mathbf{c}_B^T [I, B^{-1}N] = (\mathbf{c}_B^T, \mathbf{c}_B^T B^{-1} N)$$

Thus for feasibility we need $(\mathbf{c}_B^T, \mathbf{c}_B^T B^{-1} N) \preceq \mathbf{c}^T$. This is satisfied on the first m coordinates, namely the coordinates corresponding to B (indeed, those constraints are active at $\mathbf{y}$), but not necessarily for the last $n - m$ coordinates, where we need $\mathbf{c}_B^T B^{-1} N \preceq \mathbf{c}_N$.

25.7.3 Exercise. Check that $\mathbf{c}_B^T B^{-1} N$ is a $(n - m)$-vector.

The following theorem now follows easily.

25.7.4 Theorem. *For the* $\mathbf{x}$ *and* $\mathbf{y}^T$ *defined as above, we have*

$$\mathbf{c}^T\mathbf{x} = \mathbf{y}^T\mathbf{b} \tag{25.7.5}$$

If $\mathbf{x}$ *is feasible for the primal, and* $\mathbf{y}^T$ *is feasible for the dual problem, then by the duality theorem both* $\mathbf{x}$ *and* $\mathbf{y}$ *are optimal.*

Proof. Clearly $A\mathbf{x} = B\mathbf{x}_B$. Multiply (25.7.1) on the left by $\mathbf{c}_B^T$:

$$\mathbf{c}_B^T\mathbf{x}_B = \mathbf{c}_B^T B^{-1}\mathbf{b}$$

and (25.7.2) on the right by $\mathbf{b}$:

$$\mathbf{y}^T\mathbf{b} = \mathbf{c}_B^T B^{-1}\mathbf{b}$$

We get the same value, so if the equilibrium $\mathbf{x}$ and $\mathbf{y}$ are both feasible, the value of the primal objective function at $\mathbf{x}$ is equal to that of the dual objective function at $\mathbf{y}$, so they are both optimal by the duality theorem. □

Now we connect to the Equilibrium Theorem 25.6.11. The j-th inequality constraint $x_j \geq 0$ of the primal is active at $\mathbf{x}$ for $m+1 \leq j \leq n$. The j-th inequality constraint $\mathbf{y} \cdot \mathbf{a}_j \leq c_j$ is active for $1 \leq j \leq m$. So complementary slackness will necessarily be satisfied - as expected.

This suggests the outline of an algorithm for finding the minimizer of this optimization problem. Recall the classification of all linear optimization problems given in Theorem 25.1.10.

25.7.6 Algorithm.

Step 1. Pick an $m \times m$ submatrix B of A that is invertible. This means picking m columns from n columns, so there are $\binom{n}{m}$ ways of doing this. Since we have assumed that A has rank m, we know we will be successful.

Step 2. Compute the equilibirum point $\mathbf{x}$ using (25.7.1). If it is feasible, so that $\mathbf{x}_B \succeq \mathbf{0}$, then go to Step 3. Otherwise go back back to Step 1, selecting a new B that has not be tested yet. This could fail for all B, meaning that the feasible set is empty. If it fails, we are done: we are in Case (e) of Theorem 25.1.10.

Step 3. We reach Step 3 when we have a basic submatrix B and its associated equilibrium $\mathbf{x}$ that is feasible. Then compute $\mathbf{y}$ as in (25.7.2). If the dual equilibrium $\mathbf{y}$ is not feasible for the dual problem, meaning that $\mathbf{c}_B^T B^{-1} N \succ \mathbf{c}_N$, then select a new B that has not be tested yet, one where the equilibrium

$\mathbf{x}$ is feasible. This step could fail for all B. This means that the primal function does not achieve a minimum on the feasible set, so that we are in Case (u). If it succeeds, so we have a feasible dual equilibrium $\mathbf{y}$, then for that B, $\mathbf{x}$ is feasible for the primal and $\mathbf{y}$ is feasible for the dual, so Theorem 25.7.4 tells us that $\mathbf{x}$ is a minimizer for the primal and $\mathbf{y}$ is a maximizer for the dual, and we are done. We are in Case (f).

What is missing in this outline is the order in which we choose the basic submatrices. We want to choose them systematically in order to minimize the number of basic submatrices we test. Since one of the most expensive (computationally) parts of the algorithm is the inversion of B needed to compute $\mathbf{x}$, we want the next submatrix B' we consider after B to have a determinant that is easy to compute given that of B.

This suggests the fundamental step of the simplex algorithm. Assume we have reached a submatrix B that consists of the columns $\mathbf{a}_{j_1}, \dots \mathbf{a}_{j_m}$ of A. The next submatrix B' we consider should have all the same columns as B, save one, which should be new. Thus one column leaves B and one enters. Given this strategy, the whole art of the simplex algorithm is in deciding which column enters and which column leaves. Gaussian elimination makes it easy to compute the $\det B'$ from the computation of $\det B$.

Here is an example with 2 rows and three columns, so that there at most three basic submatrices to consider.

25.7.7 Example. Consider the feasible set:

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 5 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 5 \\ 7 \end{pmatrix} \quad , \ \mathbf{x} \geq 0$$

We wish to minimize $2x_1 + x_2 + x_3$ subject to these constraints. Find all the *basic* feasible solutions to this problem.

Note that all three 2×2 submatrices of A, which is

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 5 \end{bmatrix}$$

have non-zero determinant, so they are basic.

For example, if we choose the last two columns we get

$$B = \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix} \text{ and } \det B = 1.$$

Since

$$B^{-1} = \begin{bmatrix} 5 & -3 \\ -3 & 2 \end{bmatrix},$$

we get

$$\mathbf{x}_B = \begin{bmatrix} 5 & -3 \\ -3 & 2 \end{bmatrix} \begin{bmatrix} 5 \\ 7 \end{bmatrix} = (4, -1)$$

so $\mathbf{x} = (0, 4, -1)$ and $\mathbf{x}$ is not feasible.

So we need to take another basic submatrix and repeat the computation.

It is left to you to check that the remaining two computations of $\mathbf{x}$ also give a $\mathbf{x}_B$ that is not positive, so that the feasible set of the primal is empty.

Now we modify this example by changing the bottom right entry of A, and keeping everything else the same.

25.7.8 Exercise. The feasible set satisfies:

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 5 \\ 7 \end{bmatrix} \quad , \quad \text{and } \mathbf{x} \succeq 0.$$

The goal is minimize the same objective function $2x_1 + x_2 + x_3$ on this feasible set. Find all the basic feasible vectors. Graph the constraint hyperplanes and the feasible vectors, and then determine the minimizer. Find the optimal dual vector (y_1, y_2) and verify the complementary slackness conditions.

Partial Solution As in the previous example, we pick columns two and three: We set up the usual notation:

$$B = \begin{bmatrix} 2 & 3 \\ 3 & 4 \end{bmatrix} \text{ , } \det B = -1, \text{ and therefore } B^{-1} = \begin{bmatrix} -4 & 3 \\ 3 & -2 \end{bmatrix}$$

In particular B is a basic submatrix. The solution of

$$\begin{bmatrix} 2 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \end{bmatrix}$$

is $\mathbf{x}_B = (1, 1)$, so $\mathbf{x} = (0, 1, 1)$, which is feasible.

In this example, $\mathbf{b} = (5, 7)$ and $\mathbf{c} = (2, 1, 1)$, so $\mathbf{c}_B = (1, 1)$. So

$$\mathbf{y}^T = \mathbf{c}_B^T B^{-1} = (1, 1) \begin{bmatrix} -4 & 3 \\ 3 & -2 \end{bmatrix} = (-1, 1)$$

Is this feasible? Yes, because as we saw above, the only constraint to check is

$$\mathbf{c}_B^T B^{-1} N \leq \mathbf{c}_N.$$

Now N is the unused column of A, so it is the vector $(1, 2)$, and $\mathbf{c}_N$ is the first entry of $\mathbf{c}$, which is 2. The left-hand side is therefore $(-1, 1) \cdot (1, 2) = 1$, so the inequality is satisfied and $\mathbf{y}$ is feasible. Thus Theorem 25.7.4 tells us both $\mathbf{x}$ and $\mathbf{y}$ are optimal. We check the symbolic computation that both sides of (25.7.5) are equal:

$$2x_1 + x_2 + x_3 = 5y_1 + 7y_2 = 2.$$

You should now check what happens for the remaining two basic submatrices. Before doing the computation, you should ask yourself what the feasible set is. We know it is non-empty, and it is the intersection of a line in $\mathbb{R}^3$ with the first quadrant. Show that the intersection is a segment, so its two end points are its extreme points. The computation above shows that $\mathbf{x} = (0, 1, 1)$ is the end point where the minimum occurs. Your computation for the remaining two basic submatrices will show that one of the other $\mathbf{x}$ you find is the other end point, while the remaining one will not be feasible.

25.8 Perturbations of b and Shadow Prices

In the asymmetric problem (see (25.4.1)), we let the constraint vector $\mathbf{b}$ vary. Just as in §25.2, we call the number of variables $n + m$, so A is a $m \times (n + m)$ matrix. We write F_b for the set of $\mathbf{x} \succeq \mathbf{0}$ such that $A\mathbf{x} = \mathbf{b}$. Thus for each choice of $\mathbf{b}$, F_b is the feasible set for the associated problem.

Restrict to the subset $V = \{\mathbf{b} \in \mathbb{R}^m \mid F_b \neq \emptyset\}$: the $\mathbf{b}$ such that the F_b is non-empty.

As $\mathbf{b}$ varies in V, and A and $\mathbf{c}$ remain fixed, we get a function $v(\mathbf{b})$ of $\mathbf{b}$, called the *minimum value function*, or just the *value function*, for the problem. For values of $\mathbf{b} \in V$ for which the objective function is bounded below, and therefore the minimization problem has a bounded solution, let $v(\mathbf{b})$ be the minimum of f on F_b. For values of $\mathbf{b} \in V$ for which the objective function can take arbitrarily negative values, we set $v(\mathbf{b}) = -\infty$. Let $V_0 \subset V$ be the locus of points where the value function takes on finite values. We want to understand V_0.

No matter what value is given to $\mathbf{b}$, the feasible set F_d of the dual problem remains the same. Indeed, $F_d = \{\mathbf{y} \in \mathbb{R}^m \mid \mathbf{y}^T A \preceq \mathbf{c}\}$.

25.8.1 Theorem. *The set V_0 is empty when F_d is empty. It is the closed convex cone $C_A \subset \mathbb{R}^m$ when F_d is non-empty. In the latter case the value function $v(\mathbf{b})$ is a convex function on C_A. Indeed, it is the maximum of a finite number of affine functions, so it is piecewise linear.*

Proof. The Duality Theorem 25.5.1 tells us that the primal problem has a finite solution only when the dual problem does. In particular the feasible set F_d of the dual problem must be non-empty, which we now assume. Next we write the condition that says that the feasible set of the primal is non-empty. By Definition

19.3.1 of the finite cone C_A, it is precisely that $\mathbf{b} \in C_A$. This is a closed and convex set by Proposition 19.3.4 and Corollary 19.4.5. The Duality Theorem says that when both the feasible sets of the primal and the dual are non-empty, both optimization problems have finite solutions, so we are done.

Next we prove that the value function $v(\mathbf{b})$ is convex on C_A. Pick a $\mathbf{b} \in C_A$. By the Duality Theorem, the dual linear problem at $\mathbf{b}$ has a finite solution on F_d, which is the polyhedron of $\mathbf{y} \in \mathbb{R}^m$ such that $A^T\mathbf{y} \preceq \mathbf{c}$. This polyhedron has a finite number N of extreme points $\mathbf{y}^k$, $1 \leq k \leq N$. The maximum value of the dual program, which is attained at an extreme point, is therefore $\max_k \mathbf{y}^k \cdot \mathbf{b}$. For each k, the function $\mathbf{y}^k \cdot \mathbf{b}$ is linear in the variables $\mathbf{b}$, so that the maximum value of the dual program is the maximum of a finite number of linear functions. Each linear function is convex, so by Example 22.3.11, the maximum function is convex. In fact it is *piecewise linear*, which just means that it is the maximum of a finite number of linear functions. By the Duality Theorem, the maximum of the dual function is the minimum of the primal function, so that $v(\mathbf{b})$ is convex and piecewise linear. □

We generalize this to arbitrary convex minimization in Theorem 23.6.3.

Because $v(\mathbf{b})$ is a convex function, it is continuous. It is differentiable precisely at the points $\mathbf{b}$ where the maximum of the $\mathbf{y}^k \cdot \mathbf{b}$ is attained for only one $\mathbf{y}^k$, meaning that the minimizer is unique. We can compute this derivative.

25.8.2 Theorem (Envelope Theorem)**.** *Assume the point* $\mathbf{b}$ *is chosen so that the value function* $v(\mathbf{b})$ *is differentiable at* $\mathbf{b}$*, meaning that the minimum cost is only attained for one extreme point* $\mathbf{y}^k$*. Then*

$$\frac{\partial v}{\partial b_i}(\mathbf{b}) = y_i^k \quad \textit{for } 1 \leq i \leq m,$$

where y_i^k *is the* i*-th coordinate of* $\mathbf{y}^k$*, or, equivalently*

$$\nabla v(\mathbf{b}) = \mathbf{y}^k.$$

We generalize this result to the general convex minimization problem in Theorem 23.7.7.

Thus the maximizer $\mathbf{y}^k$ for the dual problem at $\mathbf{b}$ is the gradient of the minimum cost as a function of the constraint vector $\mathbf{b}$. For this reason the dual variable y_i^k is called either the *marginal cost* or the *shadow price* of b_i.

25.8.3 Example. We continue with Example 25.5.5, where the value of $\mathbf{b}$ is $(4, 5)$. We now vary $\mathbf{b}$.

We add the slack variables x_3 and x_4 to get the problem is asymmetric form. The extended 2×4 matrix A' is

$$A' = \begin{bmatrix} 1 & 2 & -1 & 0 \\ 2 & 1 & 0 & -1 \end{bmatrix}$$

The cone C_A is generated by the four vectors $\mathbf{a}_1 = (1, -2)$, $\mathbf{a}_2 = (2, -1)$, $\mathbf{a}_3 = (-1, 0)$, and $\mathbf{a}_4 = (0, -1)$. It is not hard to see that C_A is all of $\mathbb{R}^2$. Furthermore the feasible set of the dual is non-empty by Example 25.5.5, so that the problem will have a finite minimum for all choices of $\mathbf{b}$. You should be able to see this directly by looking at the graphs in Exercise 1.3.1. Note that $\mathbf{b}^0 = 2\mathbf{a}_1 + \mathbf{a}_2$. When $\mathbf{b}^0$, we saw that the minimizer $\mathbf{x}^*$ is $(2, 1, 0, 0)$

By Theorem 25.3.9, the dual problem is to maximize $b_1 y_1 + b_2 y_2$ subject to $\mathbf{y}^T A \leq (1, 1, 0, 0)$, or

$$\begin{aligned} y_1 + 2y_2 &\leq 1 \\ 2y_1 + y_2 &\leq 1 \\ y_1 &\geq 0 \\ y_2 &\geq 0 \end{aligned}$$

This polyhedron has four vertices: $(0, 0)$, $(1/2, 0)$, $(0, 1/2)$, and $(1/3, 1/3)$ as we saw in Example 25.5.5. So for a given $\mathbf{b}$, the answer to the maximization problem is the maximum of the function evaluated at the four vertices, in other words the maximum of the four values 0, $b_1/2$, $b_2/2$, $b_1/3 + b_2/3$. As we saw in Example 25.5.5, the maximum when $\mathbf{b} = (4, 5)$ is $b_1/3 + b_2/3 = 3$. It is now easy to draw the regions where each maximum occurs.

We can go through the same analysis for the symmetric form of linear optimization.

More interestingly, let us hold A and $\mathbf{b}$ fixed in the symmetric minimization problem, and let $\mathbf{c}$ vary. We can write the dual of the symmetric problem as:

Minimize $-\mathbf{b}^T\mathbf{y}$ subject to $-A^T\mathbf{y} \geq -\mathbf{c}$ and $\mathbf{y} \geq \mathbf{0}$.

Thus $\mathbf{c}$ has become the constraint vector, so we can apply Theorem 25.8.1 to this minimization problem to conclude that its value function $w(\mathbf{c})$ is convex. But because of the introduction of minus signs, this is minus the value function we are really interested in, and that function is concave. Thus we have shown:

25.8.4 Theorem. *The set of vectors* $\mathbf{c}$ *such that Problem* (25.4.1) *has a finite solution is empty when the feasible set* F_d *of the dual problem is empty. It is a closed convex cone when the feasible set of the dual is non-empty. In the latter case, the value function* $v(\mathbf{c})$ *is a concave function. Indeed, it is piecewise linear.*

One could of course prove this directly. First notice that at a point where the minimizer $\mathbf{x}^*$ is unique, in a neighborhood of the minimizer the minimum $\mathbf{c} \cdot \mathbf{x}^*$ is a linear function with coefficients $\mathbf{x}^*$ of the variable $\mathbf{c}$. At a point where there are several minimizers $\mathbf{x}^1, \dots, \mathbf{x}^N$, since one is looking to minimize cost, in a neighborhood one chooses the minimum of the functions $\mathbf{c} \cdot \mathbf{x}^k$, $1 \le k \le N$. The linear functions $\mathbf{c} \cdot \mathbf{x}^k$ of $\mathbf{c}$ are concave and their minimum is concave, so we are done.

25.8.5 Exercise. Let

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix}$$

and let $\mathbf{b}$ and $\mathbf{c}$ be positive 2-vectors that are allowed to vary. For each fixed value of $\mathbf{b}$ write the minimum of the symmetric minimization problem for the cost vector $\mathbf{c}$.

We will revisit these issues later in the course.

Lecture 26

Applications of Linear Programming

We start by discussing the two applications of linear optimization that were introduced in the first chapter: the diet problem and the transportation problem. Then we cover two applications of the Farkas Alternative: one to probability matrices, and the other to positive matrices. We mention Brouwer's Theorem, an important result in topology that is much used in economics. Finally we discuss matrix games, another application of linear programming.

26.1 The Diet Problem

We continue the discussion of the Diet Problem, started in §1.3, in light of what we learned about linear optimization in Lecture 25. In the terminology of that lecture, we are dealing with the Symmetric Problem 25.1.6, as we now see. We use the same variable names as before.

26.1.1 Example (The Diet Problem)**.**
The indices j, $1 \le j \le n$ represent n different foods.
The indices i, $1 \le i \le m$ represent m different nutrients.
The constant c_j is the cost of one unit of the j-th food.
The constant a_{ij} is the amount of the i-th nutrient in one unit of the j-th food.
So if a person eats $\mathbf{x} = (x_1, \dots, x_n)$ units of the different foods, then

$$a_{i1}x_1 + \dots + a_{ij}x_j + \dots + a_{in}x_n$$

units of the i-th nutrient are consumed at cost

$$\mathbf{c}^T\mathbf{x} = c_1x_1 + \dots + c_jx_j + \dots + c_nx_n$$

We assume that $n \geq m$, so there are at least as many foods as nutrients. Finally we introduce additional constraints: every day, the consumer wants to consume at least b_i units of nutrient i. We refer to b_i as the minimum daily requirement of nutrient i. So we have m constraints that we write as $A\mathbf{x} \succeq \mathbf{b}$, and n further non-negativity constraints $\mathbf{x} \succeq \mathbf{0}$.

The goal is to minimize the cost, subject to the minimum daily requirement constraints and non-negativity of the x_j.

We will assume that the b_i and c_j are all positive: both perfectly reasonable assumptions. We are guaranteed that all the a_{ij} are non-negative, but we cannot say more than that. Still, by a generalization of Exercise 25.6.10, the feasible set F is non-empty, and the minimum cost always exists.

Next we add m slack variables to the constraint equations, in order to have the asymmetric problem for computations. Then the vector $\mathbf{x}$ has $n+m$ coordinates, the first n of which are the original $\mathbf{x}$, and the remaining ones the slack variables, which also must be non-negative. The matrix A gets replaced by the compound matrix $\begin{bmatrix} A - I \end{bmatrix}$ and the vector $\mathbf{c}$ by $\begin{bmatrix} \mathbf{c} & \mathbf{0} \end{bmatrix}$ where there are m trailing zeroes.

First let us consider the special case where the minimum cost hyperplane $H_{\mathbf{c},e_0}$ in $\mathbb{R}^{n+m}$ meets the feasible set F at just one point. Then the minimization problem has just one minimizer $\mathbf{x}^*$. This solution is basic, so it has at most m non-zero coordinates. We let B be the corresponding $m \times m$ submatrix of A. As always, it picks out m columns of A, and the corresponding indices j we call basic. The non-basic coordinates of $\mathbf{x}^*$ are zero, so the positivity constraints are active there. If the solution is non-degenerate (see Assumption 27.1.1), then the positivity constraints are not active for any of the basic coordinates. If we go back to considering the problem as the symmetric problem, when are the other constraints active? Only if all the basic coordinates correspond to food variables (and none to slack variables).

Let n_0 be the number of basic coordinates of $\mathbf{x}^*$ corresponding to foods, and $m - n_0$ be the number of coordinates of $\mathbf{x}^*$ corresponding to slack variables, Thus $n_0 \leq m$. Not only the consumer need not eat more different foods than the number of nutrients, but in fact, to achieve minimality, cannot eat more than m foods (because we assumed the solution was unique).

In some circumstances, the minimum price hyperplane will meet the feasible set in more than a point, so that the consumers has options in choosing their minimum cost diet.

Next the dual problem: Maximize $\mathbf{y}^T\mathbf{b}$ subject to $\mathbf{y}^T A \preceq \mathbf{c}$ and $\mathbf{y} \succeq \mathbf{0}$.

How can we interpret this? The m-vector $\mathbf{y}$ is indexed by the nutrients, and it gets compared via the dimensional-less matrix A to prices of food. Therefore it can only be the price of nutrients. The vector $\mathbf{b}$ is the daily minimum requirement for nutrients, so $\mathbf{y}^T\mathbf{b}$ is the price of the daily minimum requirement of nutrients, and so the problem is to maximize price.

Therefore this problem is being looked at from the point of view of a store owner who sells the nutrients individually, and whose goal is to maximize sales. The constraint $\mathbf{y} \succeq \mathbf{0}$ just says that the prices must be non-negative. The constraint $\mathbf{y}^T A \preceq \mathbf{c}$ is more interesting. A unit of the j-th food contains a_{ij} units of the j-th nutrient, at cost c_j. The equivalent cost for this amount of nutrients at the store is $\sum_i^m y_i a_{ij}$. The constraint says that the prices for the nutrients needed to supply the nutrient-equivalent of one unit of food j cannot be more that the cost of one unit of food j. Otherwise it would be cheaper for the consumer to buy the food directly rather than buy the nutrients.

So what do our duality results tell us? The first good news is that the 'dimension' of $\mathbf{y}^T\mathbf{b}$ and $\mathbf{c}^T\mathbf{x}$ are the same: they are both currencies, let us say dollars.

Then weak duality theorem tells us that $\mathbf{y}^T\mathbf{b} \leq \mathbf{c}^T\mathbf{x}$, meaning that the store owner cannot sell the nutrients needed to supplied the daily minimum requirement at a price higher than that of any basket of foods supplying the minimum daily requirement. If you go back and reread what the constraints on the dual problem say, this is not surprising.

More surprisingly, duality tells us that the store owner can price the nutrients so that the total price for the nutrients supplying the daily minimum requirement is the best allowed by the weak duality theorem: equality can be achieved.

Next let us think about complementary slackness. We return to the original symmetric problem, so that all the variables $\mathbf{x}$ correspond to foods. Assume we have located a minimum $\mathbf{x}^*$ for the primal and a maximum $\mathbf{y}^*$ for the dual. First we ask what complementary slackness tells us when $\mathbf{x}_j^* > 0$: the dual equation must be an equality $\mathbf{y}^* \cdot \mathbf{a}_j = c_j$, which, since the column vector $\mathbf{a}_j$ and the price c_j are known, gives us one condition on $\mathbf{y}^*$. On the other hand, if $\mathbf{a}^i \cdot \mathbf{x}^* > b_i$, which means that the basket of food producing the minimum cost produces more than enough of nutrient i, then the store owner cannot charge anything for that nutrient: $y_i^* = 0$.

On the other hand, looking at it from the dual end: if $y_i^* > 0$, then $\mathbf{a}^i \cdot \mathbf{x}^* = b_i$, so that the minimum cost diet produces exactly the right amount of nutrient i. Finally, if $\mathbf{y}^* \cdot \mathbf{a}_j < c_j$, meaning that the unit price charged for nutrients at the maximizer $\mathbf{y}^*$ is strictly less than the price of one unit of food j, the $x_j^* = 0$, so food j does not appear in the minimum diet.

Notice that if there are several minima for the primal problem, then every maximum of the dual must satisfy complementary slackness with all the minimal. You should check what this means on an example.

26.1.2 Example. Go back to Example 25.5.5, where we now change $\mathbf{c}$. Take $\mathbf{c} = (1, 2)$, so that the minimum cost level line goes through two vertices of the feasible polyhedron of the primal problem: $(2, 1)$ and $(4, 0)$. The minimal cost is 4.

Once again let us place ourselves at a minimizer $\mathbf{x}^*$ for the primal, and ask for the number n_0 of foods that are consumed. We have already seen that no more than m foods need be consumed. Can anything more be said? One cannot have $n_0 = 0$, which would mean the consumer eats nothing. Other than that, all other possibilities occur, assuming that there is at least one food containing all the nutrients: a *perfect food*. Indeed, if the cost of all the other foods is much higher than that of the perfect food, the minimum daily requirement can be fulfilled at minimum cost just by eating the perfect food. You should check that all other possibilities for n_0 can occur, depending on the relative prices of the foods. A food gets eliminated from the diet if its price is too high relative to the other prices.

Finally we consider our perturbation results from §25.8. We start at a collection of prices $\mathbf{c}$, where the diet problem admits a unique minimizer $\mathbf{x}^*$. Now we assume that the price of each food is changed by a small amount - small enough so that the minimum cost hyperplane for the problem still intersects the feasible set in just one point: the same vertex as before.

We write the new cost vector $\mathbf{c}^n$ as the old cost vector $\mathbf{c}$ plus a small change vector Δc:

$$\mathbf{c}^n = \mathbf{c} + \Delta c$$

Remember that Δc is small enough so that the vertex where the unique minimizer occurs does not change. As the prices change by Δc, the minimum cost changes by $\Delta c \cdot \mathbf{x}$.

The results of §25.8 also tell us what happens when we perturb $\mathbf{b}$.

26.1.3 Exercise. In example 1.3.1, choose prices for the two foods so that the minimum occurs at the center vertex, meaning that the consumer is able to consume the minimum amount of the two nutrients. Determine how much you can let the prices vary before the minimum jumps to another vertex, and see how much the total cost changes.

26.1.4 Exercise. Now construct a diet example with 3 foods and 2 nutrients, such that the entries of the 2×3 matrix A are all positive and A has rank 2. Choose b in the cone generated by the columns of A. As we noted earlier, the feasible set of this problem could have as many as $\binom{5}{3} = 10$ extreme points. Construct an example with as large a number of extreme points as possible.

Hint: make sure the two constraint hyperplanes each intersect the coordinate axes on their positive part. In other words, let (a_x, a_y, a_z) be the coordinate of the intersection of the first constraint hyperplane with the x, y and z axes, and similarly let (b_x, b_y, b_z) be those of the second constraint hyperplane. Assume $a_x < b_x$, $a_y < b_y$, $a_z > b_z$. Note the reversal of direction in the last inequality.

Draw this, draw the feasible set, and count the number of extreme points.

See Strang, [67], §8.3 p.412, Lax [39], p.175-6, and Luenberger [42], Example 1 p. 81 for other presentations of the dual problem to the diet problem.

26.2 The Transportation Problem

We first prove the following theorem: see Exercise 1.4.13

26.2.1 Theorem. *The $(m + n) \times mn$ matrix A giving the m-supply equations and the n-demand equations of the canonical transportation problem has rank $m + n - 1$.*

26.3 An Application of the Farkas Alternative to Probability Matrices

In §18.8 we looked at doubly stochastic matrices. Here we consider the bigger set of probability matrices, where we only require that the column sums be 1. Here is the definition:

26.3.1 Definition. A square $n \times n$ matrix $P = (p_{ij})$ is called a *probability matrix* (or a stochastic matrix) if

1. $p_{ij} \geq 0$ for all i, j.
2. The sum of the elements of each column of P is equal to 1:

$$\sum_{i=1}^{n} p_{ij} = 1 \text{ for all } j, 1 \leq j \leq n.$$

26.3.2 Example. Any permutation matrix (see Definition 6.5.1 is a probability matrix.

26.3.3 Definition. A n-vector $\mathbf{x} = (x_i)$ is a *probability vector* if

1. $x_i \geq 0$ for all i.
2. The sum of the entries is equal to 1: $\sum_{i=1}^{n} x_i = 1$.

26.3.4 Proposition. *If P is a $n \times n$ probability matrix, and $\mathbf{x}$ an n-probability vector, then $\mathbf{y} = P\mathbf{x}$ is a probability vector.*

Proof. Since $y_i = \sum_{j=1}^n p_{ij}x_j$, and all the terms in the sum are ≥ 0, it is clear that $y_i \geq 0$.

Furthermore

$$\sum_{i=1}^n y_i = \sum_{i=1}^n \sum_{j=1}^n p_{ij}x_j = \sum_{j=1}^n x_j \sum_{i=1}^n p_{ij} = \sum_{j=1}^n x_j = 1$$

□

Thus multiplication by a probability matrix transforms a probability vector into another probability vector.

26.3.5 Definition. Given a probability matrix P, a probability vector $\mathbf{x}$ is a *steady state* for P if $P\mathbf{x} = \mathbf{x}$.

So a steady state is a non-negative eigenvector of P with eigenvalue 1. Does every probability matrix have a steady state? If P is symmetric, the answer is easily seen to be yes: just take $\mathbf{x} = (1/n, 1/n, \dots, 1/n)$. Since P is symmetric, its row sums are one, so that $P\mathbf{x} = \mathbf{x}$. If P is not symmetric, the answer is still yes, as we now see

26.3.6 Theorem. *Every probability matrix P has a steady state* $\mathbf{x}$.

Proof. A matrix with column sums all equal to 1 has an eigenvalue $\lambda = 1$. Indeed the $n \times n$ matrix $P - I$ has all column sums equal to zero, so it is singular. If $\mathbf{x}$ is a non-zero element in the nullspace of $P - I$, then $\mathbf{x}$ is an eigenvector with eigenvalue 1. It is not clear that the associated eigenvector is real. If it is, by scaling its length, we can guarantee that the sum of its entries is 1, but it is harder to see why the entries should all be positive. We get this as a corollary of the Farkas Alternative.

To apply the Farkas theorem we need a matrix A and a vector $\mathbf{b}$. We choose for A the $(n+1) \times n$ matrix whose top $n \times n$ part is $P - I$, and whose bottom row is the unit vector $\mathbf{u}^T = (1, 1, \dots, 1)$. We write this in block form as

$$A = \begin{bmatrix} P - I \\ \mathbf{u}^T \end{bmatrix}$$

For $\mathbf{b}$ we take the $(n+1)$-vector with $b_1 = b_2 = \cdots = b_n = 0$ and $b_{n+1} = 1$. The Farkas alternative says that

- either there is a solution $\mathbf{x} \geq \mathbf{0}$ to the equation $A\mathbf{x} = \mathbf{b}$. This says that $P\mathbf{x} = \mathbf{x}$ and $\sum_{i=1}^{n} x_i = 1$. This is exactly what we want.
- or (the alternative) there is an $(n+1)$-vector $\mathbf{y}$ with $\mathbf{y}^T A \geq 0$ and $\mathbf{y}^T \mathbf{b} < 0$.

We must show that the alternative is impossible. Write the coordinates of $\mathbf{y}$ as $y_i = z_i$, $1 \leq i \leq n$, and $y_{n+1} = -c$. Since the only non-zero entry of $\mathbf{b}$ is $b_{n+1} = 1$, $\mathbf{y}^T\mathbf{b} < 0$ forces $c > 0$. In block form, letting $\mathbf{z}$ be the n-vector with coordinates z_i:

$$\mathbf{y} = \begin{bmatrix} \mathbf{z} \\ -c \end{bmatrix}$$

Then the alternative gives the two equations (carry out the block multiplication of matrices):

$$\mathbf{z}^T(P - I) \geq c\mathbf{u}^T, \quad \text{and } c > 0. \tag{26.3.7}$$

We need to show that these inequalities do not have a solution. The first is actually a collection of n inequalities, one for each j:

$$\sum_{i=1}^{n} z_i p_{ij} - z_j \geq c > 0, \tag{26.3.8}$$

where the rightmost inequality follows from the second inequality in 26.3.7. Assume there is a solution in $\mathbf{z}$ and c, and let $z_m = \max_i z_i$. Then we have, for each j:

$$\begin{aligned} \sum_i z_i p_{ij} &\leq \sum_i (\max_i z_i) p_{ij} && \text{since } p_{ij} \geq 0 \\ &= z_m \sum_i p_{ij} \\ &= z_m && \text{since } \sum_i p_{ij} = 1 \end{aligned}$$

Writing this in the case $j = m$, we get $\sum_i z_i p_{im} \leq z_m$. But this contradicts (26.3.8) in the case $j = m$, since $c > 0$. So the alternative cannot occur, and we are done. □

This theorem also follows from more general theorems in linear algebra concerning *positive* matrices, namely square matrices P with all entries $p_{ij} > 0$, and non-negative matrices, square matrices P with all entries $p_{ij} \geq 0$. Probability matrices are special cases of non-negative matrices. A good reference for this material is Lax [39], Chapter 16. Comprehensive results are given in Gantmacher

[25], Volume II, chapter XIII, §2. The result needed here is Frobenius's Theorem 26.4.3.

Not surprisingly, this theorem is of interest in probability theory, where it makes its appearance in the theory of Markov chains. See for example [65], §9.4 and [46], §1.7. In both these texts a probability matrix has row sums (rather than column sums) equal to 1, and what we call a steady state is called a stationary or invariant solution.

26.4 Positive Matrices

And now for a different interpretation of these results.

26.4.1 Definition. The unit simplex in $\mathbb{R}^n$ is the $n-1$ dimensional (see Definition 18.2.24) simplex in the non-negative quadrant with vertices the n unit coordinate vectors $(1, 0, \dots, 0)$, $(0, 1, 0, \dots, 0)$, $\dots$, $(0, \dots, 0, 1)$.

Thus it is the intersection of the hyperplane $\sum_{i=1}^n x_i = 1$ with the non-negative orthant. In particular probability vectors are just elements of the unit simplex, and Proposition 26.3.4 says that a probability matrix maps the unit simplex to the unit simplex.

Theorem 26.3.6 says that the linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$ given by a probability matrix has a fixed point on the unit simplex.

This is a special case of an important theorem in topology with many applications in economics: Brouwer's Theorem 26.7.1 that we prove later.

Returning to the linear algebra setup, assume that the $n \times n$ matrix A has all entries positive and consider the map f:

$$\mathbf{x} \longmapsto \frac{\sum_{j=1}^n a_{ij}x_j}{\sum_{k=1}^n \left(\sum_{l=1}^n a_{kl}x_l\right)}$$

It is well defined at all points of $\mathbb{R}^n$ except the origin, as you should check. In particular it is well defined on the unit simplex. Note that it is a *homogeneous function* of degree 0, meaning that if you replace the variables $\mathbf{x}$ by $t\mathbf{x}$, for any non zero real number, the value of the map does not change.

26.4.2 Exercise. Show that f maps the unit simplex to the unit simplex.

Thus by Brouwer's theorem, f has a fixed point $\mathbf{x}^*$ on the unit simplex. Writing $\lambda = \sum_{k=1}^n(\sum_{l=1} na_{kl}x^*_l)$, we see that $A\mathbf{x}^* = \lambda\mathbf{x}^*$, so that λ is an eigenvalue and $\mathbf{x}^*$ the corresponding eigenvector of A. By construction λ is real and positive, and $\mathbf{x}^*$ is non-negative.

This exercise also follows from Frobenius's theorem in linear algebra:

26.4.3 Theorem (Frobenius)**.** *Every non-negative $l \times l$ matrix A, not equal to 0, has an eigenvalue, denoted λ, called the dominant eigenvalue, with the following properties:*

1. *λ is real and non-negative and its associated eigenvector $\mathbf{h}$ is real and non-negative.*
2. *Every other eigenvalue κ of A (which need not be real) satisfies $|\kappa| \le \lambda$.*
3. *If $|\kappa| = \lambda$, then κ is of the form $e^{2\pi ik/m}\lambda$, where k and m are positive integers, $m \le l$.*

This is proved for example in Lax [39], chapter 16.

26.5 Matrix Games

The results of Lecture 25 have a beautiful application to two-person games.

There are two players R (the row player) and C(the column player). R has m possible moves, labeled from 1 to m, and C has n possible moves labeled from 1 to n. R and C each choose a move simultaneously without knowledge of the other player's move. That is known as a play. The payout to R when she plays i and C plays j is a_{ij}, which can be positive or negative (which means the money actually goes to C), so the $m \times n$ matrix $A = [a_{ij}]$ now represents the payouts for all possible plays (i, j). The is a zero-sum game, so the payout to C is $-a_{ij}$.

26.5.1 Example (Football: pass and run)**.** This example comes from Strang, Exercise 8.5.14 page 441. R represents the offense of a football team and C the opposing team's defense. R can choose to run or to pass, and C can choose to defend against the run or against the pass. The payoff matrix represents the number of yards gained by R, so for example if R passes and C has set up to defend against the run, R gains 8 yards; while if C has set up to defend against the pass, R loses 6 yards: quarterback sack, for example.

$$\begin{bmatrix} & \text{defend run} & \text{defend pass} \\ \text{run} & 2 & 6 \\ \text{pass} & 8 & -6 \end{bmatrix}$$

26.5.2 Example (Rock, Paper, Scissors)**.** This is a well-known children's game. R and C extend their hand simultaneously, displaying a closed fist (rock), an open hand (paper) or two fingers (scissors). If the children show the same thing, the payout is zero. Otherwise, paper covers rock, scissors cut paper, and rock breaks scissors, and the winner gets 1 cent. Thus the payout matrix, from R's point of view, is the 3×3 matrix with labels:

$$\begin{bmatrix} & \text{paper} & \text{rock} & \text{scissors} \\ \text{paper} & 0 & 1 & -1 \\ \text{rock} & -1 & 0 & 1 \\ \text{scissors} & 1 & -1 & 0 \end{bmatrix}$$

Now back to the general case. We assume the game is played many times. Each player plays randomly, but with given probability for each move. This means that player R has picked a probability vector[1]

$$\mathbf{q} = [q_1, \dots, q_m]$$

and player C a probability vector

$$\mathbf{p} = [p_1, \dots, p_n]$$

Each vector is called the *mixed strategy* of the corresponding player. If the strategy has a 1 in one position, and zeroes everwhere else, it is called a *pure* strategy.

26.5.3 Definition. The game with these rules is called the *matrix game* associated to the matrix A.

Thus R plays move i with probability q_i, and C plays move j with probability p_j. Play (i, j) therefore occurs with probability $q_i p_j$, since we are assuming the probabilites are independent.. So the expected payout (over the long run) for the game is

$$\sum_{i=1}^{m} \sum_{j=1}^{n} q_i p_j a_{ij}$$

which is just the matrix product $\mathbf{q}^T A \mathbf{p}$.[2] Thus, if we let $f(\mathbf{p}, \mathbf{q}) = \mathbf{q}^T A \mathbf{p}$, then f is linear separately in $\mathbf{p}$ and in $\mathbf{q}$, and its domain is restricted on each side to the compact set of probability vectors.

R wants to maximize the payout $f(\mathbf{p}, \mathbf{q})$, while C wants to minimize it. We will prove that there are mixed strategies $\mathbf{p}^0$ for C and $\mathbf{q}^0$ for R such that if either player changes mixed strategy their expected payout decreases. We call such a pair of strategies optimal. More formally

[1] Probability vectors are defined in 26.3.1.

[2] This is a result from probability theory. You can just take this as the definition of the expected payout if you want.

26.5.4 Definition. A pair of strategies $\mathbf{p}^0$ and $\mathbf{q}^0$ is optimal if for all strategies $\mathbf{p}$ and $\mathbf{q}$

$$\mathbf{q}^T A\mathbf{p}^0 \leq (\mathbf{q}^0)^T A\mathbf{p}^0 \leq (\mathbf{q}^0)^T A\mathbf{p} \tag{26.5.5}$$

Here is our main theorem.

26.5.6 Theorem. *Every matrix game A has an optimal strategy.*

Before proving this, here are two examples.

26.5.7 Example (Rock, Paper, Scissors, continued)**.** It is reasonable to guess $\mathbf{p} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is the optimal strategy for each player. Let us check this.

$$A\mathbf{p} = (0, 0, 0)$$

so no matter what strategy $\mathbf{q}$ player R uses, the expected payout is 0. In the same way, if R uses this strategy $\mathbf{p}$, no matter what strategy C uses the expected payout is 0. Thus the guess is correct.

26.5.8 Example. In a related[3] example let

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

Can you guess the optimal strategies?

Next we introduce a device to reduce a general matrix game to one whose matrix A has only positive entries.

26.5.9 Definition. E denotes the $m \times n$ matrix with all entries equal to 1.

For any probability vectors $\mathbf{p}$ and $\mathbf{q}$,

$$\mathbf{q}^T E = \mathbf{1}_n^T \quad \text{and} \quad E\mathbf{p} = \mathbf{1}_m,$$

where $\mathbf{1}_m$ (resp. $\mathbf{1}_n$) is the m-vector (resp. n-vector) with all entries 1.

26.5.10 Proposition. *Consider two matrix games, the first given by A and the second by $A + kE$, where k is any real number. Then a pair of strategies that are optimal for one of these games is optimal for all of them.*

[3] Related because the matrix A is also skew-symmetric, meaning $A^T = -A$: see Definition 26.5.18

Proof. Indeed. for any mixed strategies $\mathbf{p}$ and $\mathbf{q}$, and for any k, we have

$$\mathbf{q}^T(A+kE)\mathbf{p} = \mathbf{q}^T A\mathbf{p} + k \tag{26.5.11}$$

so that the change in the expected payout for any pair of strategies $(\mathbf{p}, \mathbf{q})$ is the constant k, so (26.5.5) is preserved when A is replaced by $A + kE$. □

This proposition allows us to restrict our attention to games where the matrix A has only positive entries, which we do from now on.

Consider the standard inequality problem 25.1.6, where $\mathbf{b} = \mathbf{1}_m$ and $\mathbf{c} = \mathbf{1}_n$. Thus the primal problem is to minimize $\sum_{j=1}^n x_j$ subject to $A\mathbf{x} \succeq \mathbf{1}_m$ and $\mathbf{x} \succeq \mathbf{0}$, and the dual is to maximize $\sum_{i=1}^m y_i$ subject to $\mathbf{y}^T A \preceq \mathbf{1}_n$ and $\mathbf{y} \succeq \mathbf{0}$.

Because A is positive, the feasible sets of both the primal and the dual are non-empty, as you should check. By the Duality Theorem 25.5.1, both optimization problems have (not necessarily unique) solutions that we write $\mathbf{x}^0$ and $\mathbf{y}^0$, and the values of the optima are the same:

$$\mathbf{c}^T\mathbf{x}^0 = (\mathbf{y}^0)^T A\mathbf{x}^0 = \mathbf{y}^{0^T}\mathbf{b}, \tag{26.5.12}$$

so

$$\sum_j^n x_i^0 = \sum_i^m y_i^0.$$

Call this common value w. Clearly $w > 0$, and we have:

26.5.13 Lemma. $\mathbf{p}^0 = \frac{\mathbf{x}^0}{w}$ *and* $\mathbf{q}^0 = \frac{\mathbf{y}^0}{w}$ *are probability vectors, and*

$$(\mathbf{q}^0)^T A\mathbf{p}^0 = \frac{1}{w} \tag{26.5.14}$$

Proof. All the entries of $\mathbf{x}^0$ are non-negative, and the entries of $\mathbf{c}$ and $\mathbf{b}$ are all ones, so all the entries of $\mathbf{p}^0$ are non-negative and their sum is 1. Similarly for $\mathbf{q}^0$. (26.5.14) follows by dividing (26.5.12) by w^2. □

Since $\mathbf{x}^0$ is feasible,

$$A\mathbf{x}^0 \geq \mathbf{1}_m$$

Multiply on the left by any probability m-vector $\mathbf{q}$, and divide by w:

$$\mathbf{q}^T A\mathbf{p}^0 \geq \frac{1}{w}$$

Similarly, since $\mathbf{y}^0$ is feasible,

$$(\mathbf{y}^0)^T A \leq \mathbf{1}_n$$

Multiply on the right by any probability n-vector $\mathbf{p}$, and divide by w:

$$(\mathbf{q}^0)^T A\mathbf{p} \leq \frac{1}{w}$$

Thus the pair of strategies $(\mathbf{p}^0, \mathbf{q}^0)$ constructed above is optimal for the matrix game A. This concludes the proof of the main theorem of this section, John von Neumann's minimax theorem for matrix games. It is a consequence of the duality theorem for linear programming, as we have seen.

26.5.15 Definition. The number $v = (\mathbf{q}^0)^T A\mathbf{p}^0 = \frac{1}{w}$ is called the *value* of the game. If it is zero, the game is *fair*.

In terms of the definitions in §33.1, v is the saddle value, and $(\mathbf{p}^0, \mathbf{q}^0)$ the saddle point.

26.5.16 Proposition. *For any game A with value v, the equivalent game $A - vE$ is fair.*

This is a trivial consequence of (26.5.11).

26.5.17 Example (Football: pass and run)**.** We use these results to determine the optimal strategies in Example 26.5.1. To make the matrix A positive we add $6E$ to it. By drawing the graphs and solving the systems, it is easy to see that the solution of the primal problem with $\mathbf{b} = \mathbf{c} = \mathbf{1}$ is

$$\mathbf{x} = \Big(\frac{1}{14}, \frac{1}{28}\Big)$$

and that of the dual is

$$\mathbf{y} = \Big(\frac{1}{12}, \frac{1}{42}\Big)$$

To check the computation, we compute the sum of the coordinates on both sides.

$$\frac{1}{14} + \frac{1}{28} = \frac{1}{12} + \frac{1}{42} = \frac{3}{28}$$

They are equal, confirming the computation. This is w, so the value v of the game $A + 6E$ is $\frac{28}{3}$, so by (26.5.11) the value of A is $\frac{28}{3} - 6 = \frac{10}{3}$. This means that one expects the offense to gain 3 and a third yards per play on average. The optimal strategy for the offense C is

$$\mathbf{q}^0 = \frac{28}{3}\Big(\frac{1}{12}, \frac{1}{42}\Big) = \Big(\frac{7}{9}, \frac{2}{9}\Big)$$

and for the defense R is

$$\mathbf{p}^0 = \frac{28}{3}\Big(\frac{1}{14}, \frac{1}{28}\Big) = \Big(\frac{2}{3}, \frac{1}{3}\Big)$$

so, in words, the optimum mixed strategy for the offense C is to run seven times out of nine, while the defense should defend again the run two thirds of the time, namely a slightly smaller percentage of the time.

26.5.18 Definition. A game is *symmetric* if the matrix A is skew-symmetric, meaning that it is square and $A^T = -A$.

Note that it the game is symmetric, the payoff matrix is the same from R and C's point of view.

Here is a general result about symmetric games.

26.5.19 Proposition (Symmetric Games)**.** *If A is a symmetric game, its value is 0 and any strategy that is optimal for R is optimal for C, and vice-versa.*

The proof is an amusing exercise using matrix transposition. See Franklin, [23], p.119 for details if needed.

26.5.20 Remark. Franklin [23] p.112 , Strang [67] p.437, and Lax[39] p.177 have presentations of this material at about the same level as here. Berkowitz [7] p.126 deduces it from a more comprehensive statement of the minimax theorem for those who have had more analysis.

26.6 Sperner's Lemma

We start with a n-simplex S in $\mathbb{R}^n$. Fix $n+1$ affinely independent points $\mathbf{a}^i$ in $\mathbb{R}^n$. Then, as we learned in Definition 18.3.10, we can write the points of the simplex S spanned by the $\mathbf{a}^i$ as

$$\{\sum_{i=0}^{n} \lambda_i \mathbf{a}^i \mid \lambda_i \geq 0 \, , \, \sum_{i=0}^{n} \lambda_i = 1\}.$$

The λ_i are the barycentric coordinates of the corresponding point, and are uniquely determined, as we noted in Definition 18.3.13.

Since for what we do here, we only need to work for one simplex, we may as well choose the regular simplex of Definition 18.3.23: we can take the barycenter $\mathbf{c}$ to be the origin, and vertices $\mathbf{a}^i$ of the simplex are all at the same distance from the origin, which we can take to be 1.

The simplex S has $n+1$ *faces* of dimension n-1: they are the affine span of any subcollection of n of the $(n+1)$ vertices intersected with S. More generally, for any collection of $m+1$ vertices of S, S has a m-facet: the intersection of the m-dimensional affine space spanned by $m+1$ vertices of S with S. Clearly S has $\binom{n+1}{m+1}$ facets of dimension m. The edges of the simplex are the facets of dimension 1. S has $(n+1)n/2$ edges. If S is a regular simplex, all the edges have the same length, and all the m-facets have the same m-volume.

Now for each integer N we subdivide S into a collection of smaller subsimplices, called *cells*, by allowing as possible vertices of subsimplices, points with coordinates of the form

$$\frac{1}{N}(k_0, k_1, \dots, k_n)\text{ , all } k_i \in \mathbb{Z}\text{ , } k_i \geq 0, \sum_{i=0}^{n} k_i = N. \tag{26.6.1}$$

and by allowing as faces hyperplanes that are parallel to the faces of the original simplex S and that go through collections of n of the new vertices 26.6.1.

26.6.2 Example. If S is a regular simplex, then each cell in the N-th subdivision of S is a regular simplex with edge length equal to $1/N$ the edge length of S.

26.6.3 Theorem. *For each one of the $n+1$ possible directions for the hyperplanes, there are $n+1$ different hyperplanes that intersect S. If all these hyperplanes are removed from S, there are N^n disjoint open cells left. If V denotes the n-dimensional volume of S, then each of the N-cells has n-volume V/N^n.*

26.6.4 Theorem. *The number of vertices in the N-th subdivision in dimension n is $\binom{n+N}{n}$.*

Proof. We need to count the number of vertices given by (26.6.1). This is easy. Take $n+N$ symbols all representing the number 1. Lay them out on a line. Any time you pick n of them, imagine the choice as dividing the 1's into $n+1$ groups, corresponding to k_0, k_1, ..., k_n. □

With the combinatorics of these subdivisions out of the way, we can describe the key tool of the proof of Brouwer's theorem: the Sperner labeling of the vertices of the N-subdivision.

26.6.5 Definition. On the N-th subdivision of the n-simplex, we label the vertices of the cells using the numbers 0, 1, ..., n with the following restriction:

On the vertices of the facet spanned by the vertices of S with labels l_0, l_1, ..., l_m, only the labels l_0, l_1, ..., and l_m are allowed. In particular, on the edge with labels $0, 1$, you can only use 0 and 1; on the edge with labels $1, 2$, you can only use 1 and 2, etc.

Note that there are no restrictions on the vertices that are not on the boundary of the simplex S.

26.6.6 Definition. A cell of the N-th subdivision of the n-simplex has a *complete labeling* if one of its vertices has label i, for each integer i, $0 \le i \le n$.

In particular all of its vertices have distinct labels.

Our main result is

26.6.7 Lemma (Sperner's Lemma)**.** *Label the vertices of the N-th subdivision of the n-simplex according to Definition 26.6.5. Then there are an odd number of cells with complete labels. In particular there is at least one cell with complete labels.*

Proof. For such a Sperner labeling, for $k = 0, \dots, n$, let $F_k(i_0, i_1, \dots, i_k)$ be the numbers of *elements* of dimension k in the N-th subdivision, where the $k+1$ vertices have labels $i_0, i_1, \dots, i_k$.

For example, when $k = 0$, since there are $\binom{n+N}{n}$ vertices altogether, we have

$$\sum_{i=0}^{n} F_0(i) = \binom{n+N}{n}.$$

When $k = n$, we are counting the number of cells, and we know that there are N^n of them, so

$$\sum_{i_0 \le i_1 \le \dots \le i_n} F_n(i_0, i_1, \dots, i_n) = N^n.$$

Our goal is to show that $F_n(0, 1, 2, \dots, n)$ is odd, and therefore nonzero.

The proof is by induction on n. We start with $n = 1$. The simplex is then just a line segment divided into N subsegments. The Sperber labeling requirement is simply that at one end the vertex is labeled 0, and at the other end, the vertex is labeled one. The labeling of the subsegments are arbitrary. The question is: how many cells (little segments) are labeled $[0, 1]$ (which we count as the same as $[1, 0]$): in other words, what is $F_1(0, 1)$?

Cut open the simplex at all the vertices of the N-th subdivision to produce N small simplices. How many vertices labeled 1 are there? Each simplex labeled $[0, 0]$ contributes 2 vertices while a simplex labeled $[0, 1]$ contributes 1. So altogether we get $2F_1(0, 0) + F_1(0, 1)$ vertices. Now because of our Sperner labeling rule, only one of the vertices corresponds to an outside vertex. All the others are inside vertices, therefore counted twice. So we have shown when dimension $n = 1$:

$$2F_1(0, 0) + F_1(0, 1) = 2F_{int}(0) + 1.$$

This shows that $F_1(0,1)$ is odd, the desired result.

Now assume that by induction we have shown that in dimension $n-1$,

$$F_{n-1}(0,1,2,\dots,n-1) \text{ is odd.}$$

Next we work in dimension n. Cut open the simplex into the N^n cells of the N-th subdivision. Consider all the cells that have a $n-1$- face with labels $(0,1,2,\dots,n-1)$. It is either a cell where one of the labels 0, 1, …, $n-2$, or $n-1$ is repeated, and such a cell has two such faces, or it is a cell with complete labels, and such a cell has exactly one such face.

So the number

$$2\sum_{m=0}^{n-1} F_n(0,1,2,\dots,n-1,m) + F_n(0,1,2,\dots,n-1,n)$$

represents the total number of faces of dimension $n-1$ with labels $(0,1,2,\dots,n-1)$ of the cut-up simplex.

As before we let $F_{int}(0,1,2,\dots,n-1)$ be the number of interior faces. We now let $F_{bound}(0,1,2,\dots,n-1)$ be the number of boundary faces. Because we are dealing with a Sperner labeling, all these faces can only occur on the face with labels $(0,1,2,\dots,n-1)$ of the original simplex. Therefore by induction $F_{bound}(0,1,2,\dots,n-1)$ is odd.

So we have shown that

$$\begin{aligned} 2\sum_{m=0}^{n-1} F_n(0,1,2,\dots,n-1,m) &+ F_n(0,1,2,\dots,n-1,n) \\ &= 2F_{int}(0,1,2,\dots,n-1) + F_{bound}(0,1,2,\dots,n-1), \quad (26.6.8) \end{aligned}$$

and since the last term is odd, $F_n(0,1,2,\dots,n-1,n)$ is odd, as required. □

26.7 Brouwer's Fixed Point Theorem

We will now use Sperner's Lemma to prove the following important fixed point theorem.

26.7.1 Theorem (Brouwer's Theorem)**.** *Let f be a continuous map of a simplex into itself. Then f has a fixed point, meaning a point $\mathbf{x}^*$ in the simplex such that $\mathbf{x}^* = f(\mathbf{x}^*)$.*

Proof. Here is the proof strategy: we assume by contradiction that there is a continuous map f of the n-dimensional simplex S that does not have a fixed point. Then for every integer N we use the lack of fixed point to construct a Sperner labeling 26.6.5 of the vertices of the N-th subdivision of S.

Here is the key idea of the proof.

26.7.2 Remark. Let $(\lambda_0, \lambda_1, \dots, \lambda_n)$ be the barycentric coordinates of an arbitrary point $\mathbf{x}$ of S, so that all the λ_i are non-negative, and $\sum_{i=0}^{n} \lambda_i = 1$, Let $(\mu_0, \mu_1, \dots, \mu_n)$ be the barycentric coordinates of the image $f(\mathbf{x})$. All the μ_i are non-negative, and $\sum_{i=0}^{n} \mu_i = 1$ Since $\mathbf{x}$ is not a fixed point of f , not all the μ_i are equal to the λ_i. This allows us to build a Sperner labeling.

26.7.3 Definition. Let the Sperner labeling of $\mathbf{x}$ associated to f be the smallest index i such that $\lambda_i > \mu_i$.

Because $\sum_{i=0}^{n} \lambda_i = \sum_{i=0}^{n} \mu_i$, such a smallest index exists. The next lemma shows that this labeling is indeed a Sperner labeling.

26.7.4 Lemma. *If $\mathbf{x}$ belongs to the m-facet spanned by $\mathbf{a}_{i_0}$, $\mathbf{a}_{i_1}$, ..., $\mathbf{a}_{i_m}$, then the label of $\mathbf{x}$ is one of the indices i_0, i_1, ..., i_m.*

Proof. By definition, if $\mathbf{x}$ belongs to the given m-facet, its only non-zero barycentric coordinates are among the λ_{i_0}, λ_{i_1}, ..., λ_{i_m}. Therefore these are the only ones that can be strictly greater than the corresponding barycentric coordinate for $f(\mathbf{x})$, which is non-negative. □

26.7.5 Corollary. *Assuming that f has no fixed points, the labeling of Definition 26.7.3 is a Sperner labeling of the vertices of the N-th subdivision of S. Therefore for any N we can find a cell c_N in the N-th subdivision with a complete labeling: we can number the $n+1$ vertices of the cell as $\mathbf{x}^0$, $\mathbf{x}^1$, ..., $\mathbf{x}^n$, so that the label of $\mathbf{x}^i$ is i.*

By Definition 26.7.3, if (λ^i_j), $0 \le j \le n$ are the barycentric coordinates of $\mathbf{x}^i$, and (μ^i_j), $0 \le j \le n$ those of $f(\mathbf{x}^i)$, then $\lambda^i_i > \mu^i_i$.

Now we can prove the theorem. To each N, we pick a cell c_N in the N-th subdivision with a complete labeling. If the diameter of S is d, then the diameter of c_N is d/N, so as $N \to \infty$, the distance between the vertices of c_N goes to 0.

By choosing a subsequence N_k of the N, using sequential compactness and the fact that the diameter of a cell in the N-th subdivision goes to 0, we see that the vertices $\mathbf{x}^i(N_k)$ all converge to a point $\mathbf{x}^*$ in S. This gives convergent sequences in N_k of inequalities

$$\lambda^i_i(N_k) > \mu^i_i(N_k)\ , 0 \le j \le n, \tag{26.7.6}$$

which in the limit give inequalities for the barycentric coordinates of $\mathbf{x}^*$ and $f(\mathbf{x}^*)$:

$$\lambda_i^* \geq \mu_i^* \, , 0 \leq j \leq n. \tag{26.7.7}$$

But since these are barycentric coordinates, Remark 26.7.2 shows that these inequalities imply equalities, so $\lambda_i^* \geq \mu_i^* \, , 0 \leq j \leq n$. We have found a fixed point. □

It is easy to improve the theorem by making the following definition:

26.7.8 Definition. Two closed regions X and Y in $\mathbb{R}^n$ are *topologically equivalent* if there is a continuous map $\mathbf{y} = T(\mathbf{x})$ from X to Y, with continuous inverse T^{-1} from Y back to X. T is what topologists call a homeomorphism.

26.7.9 Theorem. *If the Brouwer theorem is true for X, it is true for Y.*

Proof. Let $\mathbf{f}(\mathbf{y})$ be a continuous map from Y to itself. We want to show it has a fixed point. Consider the function $\mathbf{g}(\mathbf{x}) = \mathbf{T}^{-1}(\mathbf{f}(\mathbf{T}(\mathbf{x})))$ from X to X. As a composite of continuous functions, it is continuous. Note that $\mathbf{f}(\mathbf{y}) = \mathbf{T}(\mathbf{g}(\mathbf{T}^{-1}(\mathbf{y})))$. Since the Brouwer theorem is true for X, there is a fixed point $\mathbf{x}^*$:

$$\mathbf{g}(\mathbf{x}^*) = \mathbf{x}^*.$$

Let $\mathbf{y}^* = \mathbf{T}(\mathbf{x}^*)$, so $\mathbf{x}^* = \mathbf{T}^{-1}(\mathbf{y}^*)$. Then

$$\mathbf{y}^* = \mathbf{T}(\mathbf{x}^*) = \mathbf{T}(\mathbf{g}(\mathbf{x}^*)) = \mathbf{T}(\mathbf{g}(\mathbf{T}^{-1}(\mathbf{y}^*))) = \mathbf{f}(\mathbf{y}^*)$$

so we have found our fixed point. □

Recall Definition 18.4.11. Then

26.7.10 Theorem. *Let C be a convex body in $\mathbb{R}^n$. Then C is topologically equivalent to the unit ball in $\mathbb{R}^n$.*

Proof. Let $\mathbf{c}$ be the barycenter of C. Since C has dimension n, a small neighborhood of $\mathbf{c}$ is contained in C. By the construction of Exercise 18.4.11, we can show that every unit length ray $\mathbf{r}$ in $\mathbb{R}^n$ starting at $\mathbf{c}$ meets the boundary of C in point at distance $\lambda(\mathbf{r}) > 0$ from $\mathbf{c}$. Then the map that sends $\mathbf{r}$ to $\frac{\mathbf{r}}{\lambda(\mathbf{r})}$ maps the boundary of the unit ball to the boundary of C, and can be easily filled into a bicontinuous map from C to the unit ball. □

26.8 Kakutani's Fixed Point Theorem

In this section, for the first and only time in this course we will work with a *correspondance*: a mapping F that associates to each point $\mathbf{x}$ in a subset X of $\mathbb{R}^n$ a set $F(\mathbf{x}) \subset \mathbb{R}^n$.

The *graph* of the correspondance F is the set $\{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{2n} \mid \mathbf{y} \in F(\mathbf{x})\}$.

26.8.1 Definition. The graph of F is *closed* if any time $\mathbf{x}^n$, $n \in \mathbb{N}$, is a converging sequence of points in X, with limit $\mathbf{x}^*$, and $\mathbf{y}^n$, $n \in \mathbb{N}$, is a sequence in $F(\mathbf{x}^n)$ with limit $\mathbf{y}^*$ in $\mathbb{R}^n$, then $\mathbf{y}^*$ is in $F(\mathbf{x}^*)$.

26.8.2 Theorem (Kakutani's Theorem)**.** *Let X be a compact convex set in $\mathbb{R}^n$. For each $\mathbf{x} \in X$, let $F(\mathbf{x})$ be a non-empty convex $Y \subset X$. Assume that the graph of F is closed. Then some point $\mathbf{x}^*$ lies in $F(\mathbf{x}^*)$.*

Proof. □

26.9 Nash Equilibrium

We apply Kakutani's fixed point theorem to derive Nash's famous equilibrium theorem for n-person games. For simplicity, we only deal with the case of three players. Thus the situation is very similar to that of §26.5. Here we have three players A, B, and C. A has m possible moves, labeled from 1 to m; B has n possible moves, labeled from 1 to n; finally C has p possible moves, labeled from 1 to p.

As in §26.5 each player uses a mixed strategy: A plays i with probability p_i; B plays j with probability q_j; and C plays k with probability p_k. Of course

$$p_i \geq 0 \text{ and } \sum_{i=1}^{m} p_i = 1;$$

$$q_j \geq 0 \text{ and } \sum_{j=1}^{n} q_j = 1;$$

$$r_k \geq 0 \text{ and } \sum_{k=1}^{p} r_k = 1.$$

Our next assumption is if the three players play mixed strategies $\mathbf{p}$, $\mathbf{q}$, $\mathbf{r}$, the payoff to A is the expected value

$$a(\mathbf{p}, \mathbf{q}, \mathbf{r}) = \sum_{i,j,k} a_{ijk} p_i q_j r_k$$

where a is a *three-dimensional* matrix with pqr constant terms a_{ijk}. Similarly, the payoff to B is the expected value

$$b(\mathbf{p}, \mathbf{q}, \mathbf{r}) = \sum_{i,j,k} b_{ijk} p_i q_j r_k$$

and that to C is

$$c(\mathbf{p}, \mathbf{q}, \mathbf{r}) = \sum_{i,j,k} c_{ijk} p_i q_j r_k$$

We assume that each player knows the probabilities that the other two players will play their moves. In particular A knows $\mathbf{q}$ and $\mathbf{r}$, and therefore wants to chose $\mathbf{p}$ to maximize $a(\mathbf{p}, \mathbf{q}, \mathbf{r})$ over all probability vectors $\mathbf{p}$.

This defines a set of mixed strategies $\mathbf{p} \in P(\mathbf{q}, \mathbf{r})$, namely: given mixed strategies $\mathbf{q}$ and $\mathbf{r}$ for B and C, the set of mixed strategies in $P(\mathbf{q}, \mathbf{r})$ are optimal for A.

When $\mathbf{q}$ and $\mathbf{r}$ are given, the payoff for A can be written $\sum_i a_i p_i$, where

$$a_i = \sum_{j,k} a_{ijk} q_j r_k.$$

26.9.1 Lemma. *Let $a = max_i a_i$. To maximize $\sum_i a_i p_i$, choose those probability vectors $\mathbf{p}$ with $p_i = 0$ if $a_i < a$. This set P of probability vectors is closed bounded and convex.*

This is A's optimal payoff $P(\mathbf{q}, \mathbf{r})$.

In the same way B chooses his mixed strategy in the compact convex set of optimal mixed strategies $Q(\mathbf{p}, \mathbf{r})$, and C chooses his in the compact convex set of optimal mixed strategies $R(\mathbf{p}, \mathbf{q})$.

Lecture 27

The Simplex Method

The simplex method is the key tool for solving linear optimization problems. It is fair to say that its discovery by Dantzig in 1949 revolutionized the study of optimization. For that reason, even though these lectures do not much concern themselves with numerical methods, it is imperative to give an overview of the method in the simplest case: We take the equality (asymmetric, standard) form of the linear optimization problem:

Minimize $\mathbf{c}^T\mathbf{x}$ subject to the constraints $A\mathbf{x} = \mathbf{b}$ and $\mathbf{x} \succeq \mathbf{0}$.

Recall that A is a $m \times n$ matrix of maximal rank m. We explain the simplex algorithm in this case, focusing on the interesting mathematics it uses, and just touching on the computational methodology.

27.1 Introduction

In this section we motivate our approach to the simplex method. We develop an algorithm from the outline 25.7.6 given in §25.7. We use the terminology of that section. The computation divides into two phases with similar methodology.

In Phase 1, starting from a basic submatrix B of A, we either find by iteration a new basic submatrix such that its equilibrium $\mathbf{x}$ given by (25.7.1) is feasible, or we determine that the feasible set is empty.

In Phase 2, starting from a basic submatrix whose equilibrium $\mathbf{x}$ is feasible, we either find iteratively a new basic submatrix for which both $\mathbf{x}$ and the dual equilibrium $\mathbf{y}$ are feasible, or we determine that the objective function is unbounded negatively on the feasible set.

Then Theorem 25.7.4 tells us that $\mathbf{x}$ is a minimizer for our problem, and that $\mathbf{y}$ is a maximizer for the dual problem.

27.1.1 Non-degeneracy

The algorithm simplifies if the system satisfies the following condition.

27.1.1 Assumption (Non-degeneracy Assumption)**.** The system of equations $A\mathbf{x} = \mathbf{b}$ is *nondegenerate* if the vector $\mathbf{b}$ cannot be written as a linear combination of fewer than m columns of A.

Since A has rank m, its column rank is m, so it has m columns that are linearly independent. so any vector in $\mathbb{R}^m$, in particular the vector $\mathbf{b}$, can be written as a linear combination of those m columns. If the non-degeneracy assumption is satisfied, then $\mathbf{b}$ cannot be written as a linear combination of fewer than m of the columns of A. Thus any vector $\mathbf{x}$ that satisfies $A\mathbf{x} = \mathbf{b}$ must have m non-zero coordinates. In particular, assuming a basic submatrix B has been selected, the vector $\mathbf{x} = (\mathbf{x}_B, \mathbf{0})$ of (25.7.1) associated to it must have m non-zero coordinates, so that all the coordinates of $\mathbf{x}_B$ must be positive.

27.1.2 Example. Assume that the $\mathbf{b}$ vector in Example 1.3.1 is $(1, 2)$ so $\mathbf{b} = \mathbf{a}^1$, and the non-degeneracy assumption fails. Take the 2×2 basic submatrix corresponding to the original matrix A, so that the two slack variables are 0. The constraints are therefore

$$\begin{aligned} x_1 + 2x_2 &= 1 \\ 2x_1 + \ x_2 &= 2 \end{aligned}$$

which implies that $x_1 = 1$ and $x_2 = 0$. So three of the constraints are active: indeed, 3 of the 4 constraint hyperplanes meet at the point $(1, 0, 0, 0)$. You should draw the picture. Also show that by perturbing $\mathbf{b}$ by an arbitrarily small amount, you can get rid of degeneracy.

27.1.3 Exercise. In $\mathbb{R}^2$ consider the locus $x_1 \geq 0$, $x_2 \geq 0$, $2x_1 + x_2 \geq 6$. Draw the feasible set, and find all its extreme points . Now add a second constraint $x_1 + 3x_2 \geq 12$. Add it to your picture, and find all the extreme points in the feasible set. Add slack variables x_3 and x_4 to make this an asymmetric problem in $\mathbb{R}^4$. Is this problem non-degenerate? How many basic submatrices are there? Compute the equilibrium $\mathbf{x}$ for each basic submatrix, and indicate which ones are feasible. Indicate the projection of each one of the equilibrium points on your graph in $\mathbb{R}^2$.

Here is a useful result concerning non-degeneracy.

27.1.4 Lemma. *Assume the problem is non-degenerate, and let* $\mathbf{x}$ *be in the feasible set. Then* $\mathbf{x}$ *is basic if and only if* $\mathbf{x}$ *has exactly* $n - m$ *coordinates equal to 0.*

Proof. First we assume that $\mathbf{x}$ is basic, so that it has at least $n - m$ coordinates equal to 0. We need to eliminate the possibility that more than $n - m$ coordinates of $\mathbf{x}$ are equal to 0. If that were the case, the equation $A\mathbf{x} = \mathbf{b}$, which, written in terms of the columns of A, says after reindexing:

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \dots x_m\mathbf{a}_m = \mathbf{b}. \tag{27.1.5}$$

would express $\mathbf{b}$ as the sum of fewer than m columns of A, contradicting non-degeneracy.

For the other implication, assume that $\mathbf{x}$ has exactly $n - m$ coordinates equal to 0. By reindexing as before, (27.1.5) is satisfied with all the $x_i > 0$. We need to show that the columns $\mathbf{a}_1, \dots, \mathbf{a}_m$ are linearly independent, which then mplies $\mathbf{x}$ is basic. Suppose they are not. Then there is an equation of linear dependence that can be written

$$d_1\mathbf{a}_1 + d_2\mathbf{a}_2 + \dots d_m\mathbf{a}_m = \mathbf{0}, \tag{27.1.6}$$

where not all the real numbers d_i are zero. Subtract t times (27.1.6) from (27.1.5) to get

$$(x_1 - td_1)\mathbf{a}_1 + (x_2 - td_2)\mathbf{a}_2 + \dots (x_m - td_m)\mathbf{a}_m = \mathbf{b}.$$

By choosing t so that one of the coefficients cancels, we get a representation of $\mathbf{b}$ as a sum of a smaller number of columns of A, a contradiction. □

27.1.7 Remark. Here is what non-degeneracy means in terms of cones. By definition, the feasible set is non-empty if and only if $\mathbf{b} \in C_A$, the cone on A defined in §19.3. C_A has dimension m, and any basic subcone of C_A also has dimension m. By Theorem 19.4.1 C_A is a union of its basic subcones, so that $\mathbf{b}$ is in a basic subcone. These subcones may overlap, so $\mathbf{b}$ may belong to several of these subcones. Non-degeneracy means that $\mathbf{b}$ is not in the boundary of any basic cone. So when A is fixed, the set of degenerate $\mathbf{b}$ is a set of smaller dimension than C_A, so "most" $\mathbf{b}$ are nondegenerate. By moving by an arbitrarily small distance from a degenerate $\mathbf{b}$, one finds a nondegenerate $\mathbf{b}'$, as long as the direction one moves in is not contained in the boundary on a basic cone.

27.2 Phase 2 of the Simplex Algorithm

It is easiest to start with the second phase of the algorithm, since the first phase will turn out to be a special case of the second. Thus we assume we have found a basic submatrix B such that the associated equilibrium $\mathbf{x}$ is feasible. We will show how to improve the choice of B. We assume the system is nondegenerate.

Associated to B we have our equilibrium $\mathbf{x}$ given by (25.7.1). By assumption it is feasible, so all its coordinates are non-negative. Furthermore the non-degeneracy hypothesis means that all the coordinates of $\mathbf{x}_B$ are positive. We also have the dual equilibrium $\mathbf{y}$ defined by (25.7.2).

By Theorem 25.7.4, if $\mathbf{y}$ is feasible for the dual problem, $\mathbf{x}$ is a minimizer for the primal, and we are done. So we may assume that $\mathbf{y}$ is not feasible. Thus for some non-basic column $\mathbf{a}_s$,

$$\mathbf{y} \cdot \mathbf{a}_s > c_s. \tag{27.2.1}$$

We will use the column $\mathbf{a}_s$ to build a new basic submatrix on which the cost decreases. There may be several s for which (27.2.1) is satisfied, and we do not indicate yet how to choose one.

27.2.2 Theorem. *There is a column $\mathbf{a}_k$ in the basic submatrix B such that the submatrix B' formed by the column $\mathbf{a}_s$ and the columns of B other than $\mathbf{a}_k$ is basic. Furthermore the equilibrium $\mathbf{x}'$ associated to B' is feasible, and the cost evaluated at $\mathbf{x}'$ is less than the cost evaluated at the equilibrium point $\mathbf{x}$ associated to B.*

Proof. The entering column $\mathbf{a}_s$ can be written as a linear combination of the basic columns, so

$$\mathbf{a}_s = \sum_B t_j \mathbf{a}_j \tag{27.2.3}$$

Here $\sum_B$ means to sum over the basic columns only. If we write $\mathbf{t}_B$ for the m-vector with coordinates t_j, then $\mathbf{a}_s = B\mathbf{t}_B$. Let $\mathbf{y}$ be the dual equilibrium point for B. Since $\mathbf{y}^T = \mathbf{c}_B^T B^{-1}$ by (25.7.2), combining with (27.2.3) gives

$$\mathbf{y}^T \mathbf{a}_s = \mathbf{c}_B^T B^{-1} B \mathbf{t}_B = \mathbf{c}_B^T \mathbf{t}_B.$$

Thus we can rewrite (27.2.1), the equation that tells us that feasibility fails on the column $\mathbf{a}_s$ as

$$c_s < \mathbf{c}_B^T \mathbf{t}_B. \tag{27.2.4}$$

This allows us to formulate our first rule:

27.2.5 Algorithm (Rule 1)**.** Pick as the entering column any non-basic $\mathbf{a}_s$ satisfying (27.2.4). In words, pick a non-basic column $\mathbf{a}_s$ at which the cost c_s is less than the cost associated to the linear combination $\sum_B t_j \mathbf{a}_j$ expressing $\mathbf{a}_s$ in terms of the basic columns. If there is no such $\mathbf{a}_s$, the algorithm terminates because we are at a minimizer.

If we denote by z_s the number $\mathbf{c}_B^T \mathbf{t}_B$ appearing on the right-hand side of (27.2.4), we get

$$c_s < z_s. \tag{27.2.6}$$

Multiply (27.2.3) by a positive factor λ, and add it to the equation $B\mathbf{x}_B = \mathbf{b}$ expressing the fact that the equilibrium $\mathbf{x}$ satisfies the equality constraints. We get

$$\lambda\mathbf{a}_s + \sum_B (x_j - \lambda t_j)\mathbf{a}_j = \mathbf{b}. \tag{27.2.7}$$

Because the problem is nondegenerate, all the x_j are positive. So, by taking λ sufficiently small and positive, we can guarantee that all the coefficients in (27.2.7) are positive, so that they correspond to a point $p(\lambda)$ in the feasible set.

The cost at $p(\lambda)$ is

$$\lambda c_s + \sum_B (x_j - \lambda t_j)c_j$$

while the cost at the original $\mathbf{x}$ is of course $\sum_B x_j c_j$. Thus the increment in the cost is

$$\lambda(c_s - \sum_B t_j c_j) = \lambda(c_s - \mathbf{c}_B^T\mathbf{t}_B = \lambda(c_s - z_s). \tag{27.2.8}$$

according to the definition of z_s. The cost decreases if this number is negative, so if $c_s - z_s < 0$. Now (27.2.6) tells us that this is the case, so $p(\lambda)$ is not only feasible, but the cost at $p(\lambda)$ is less than the cost at the equilibrium $\mathbf{x}$.

It remains to find a new basic submatrix. We let λ increase from 0 in (27.2.7). If all the t_j are negative, λ can increase to infinity while we still stay in the feasible set, since all the coefficients in (27.2.7) remain positive. As λ increases, the cost goes to $-\infty$ by (27.2.8), so we are done: there is no finite solution. On the other hand, if at least one of the t_j is positive, then there is a value λ_0 for which all the coefficients in (27.2.7) are non-negative, and at least one is 0. Indeed, consider all the basic indices j such that $t_j > 0$. Among those, pick any index j such that the x_j/t_j is minimum. That is a suitable k. Then $\lambda_0 = x_k/t_k$.

Our leaving rule is:

27.2.9 Algorithm (Rule 2)**.** Given the entering column $\mathbf{a}_s = \sum_B t_j\mathbf{a}_j$, for leaving column choose any basic column $\mathbf{a}_k$ such that $t_k > 0$ and

$$x_k/t_k = \min\big(\frac{x_j}{t_j} \mid t_j > 0\big).$$

If all the $t_j \leq 0$, then the objective function takes arbitrarily negative values on the feasible set.

Actually then by the non-degeneracy assumption this determines k uniquely. If there were more than one basic index j such that $\lambda_0 = x_j/t_j$, then (27.2.7) would give a way of writing $\mathbf{b}$ as a sum of fewer than m columns, a contradiction. Thus (27.2.7) expresses $\mathbf{b}$ as a linear combination of m columns, so we have the required new basic submatrix B' of A, with its feasible equilibrium point $\mathbf{p}(\lambda_0)$. □

The proof suggests an algorithm for choosing the entering index s. Indeed, by (27.2.8) , for the indices l failing the dual constraint $\mathbf{y} \cdot \mathbf{a}_l \leq c_l$, the positive decrease in cost from the original equilibrium $\mathbf{x}$ to the one associated to the new basic submatrix is given by

$$\lambda_l(\mathbf{y}^T\mathbf{a}_l - c_l), \tag{27.2.10}$$

where λ_l stands for the value λ_0 in the proof that brings us to the equilibrium value for the new basis. So a good choice for s might be any index l for which (27.2.10) is maximal. A simpler choice would be to take an l for which the factor $\mathbf{y}^T\mathbf{a}_l - c_l$ is maximal. This is a good surrogate for the maximal decrease in cost. It might still be necessary to break ties in indices with the same decrease in cost.

The proof shows that under the non-degeneracy assumption, the leaving column is unique.

27.2.11 Example. Here is a simple example where n = 4 and $m = 2$. The constraint equation $A\mathbf{x} = \mathbf{b}$ is

$$\begin{bmatrix} 1 & 0 & 3 & 4 \\ 0 & 1 & 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

When there are two equations, non-degeneracy means that $\mathbf{b}$ is not a multiple of a column. In this example it is not. Let the cost vector be $\mathbf{c} = (1, 2, 3, 3)$. Take as basic submatrix B the first two columns of A. B is the identity matrix, therefore B^{-1} is also, and our formulas simplify a great deal. The equilibrium $\mathbf{x}$ for B is $(2, 1, 0, 0)$, which is feasible. The cost at $\mathbf{x}$ is $\mathbf{c} \cdot \mathbf{x} = 4$ The dual equilibrium point for B is $\mathbf{y} = (1, 2)$ by (25.7.2), since $\mathbf{c}_B = (1, 2)$. The feasibility test is $\mathbf{y}^T N \preceq \mathbf{c}_B^T$, which would say that

$$\begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 3 & 4 \\ 1 & 3 \end{bmatrix} \preceq \begin{bmatrix} 3 & 3 \end{bmatrix} \text{ , or } \quad \begin{bmatrix} 5 & 10 \end{bmatrix} \preceq \begin{bmatrix} 3 & 3 \end{bmatrix}.$$

which is false, so $\mathbf{y}$ is not feasible. In fact the test fails on both coordinates, so we can pick either $\mathbf{a}_3$ or $\mathbf{a}_4$ as the entering column. The failure is worse on the $\mathbf{a}_4$ column, so we will chose it to enter the basis. It is written $\mathbf{a}_4 = 4\mathbf{a}_1 + 3\mathbf{a}_2$, , so $t_1 = 4$ and $t_2 = 3$. They are both positive, so we compare the two expressions $x_1/t_1 = 1/2$ and $x_2/t_2 = 1/3$, so the leaving column is $\mathbf{a}_2$. Our new basic B' is

$$\begin{bmatrix} 1 & 4 \\ 0 & 3 \end{bmatrix},$$

its inverse is

$$\begin{bmatrix} 1 & -4/3 \\ 0 & 1/3 \end{bmatrix},$$

and its feasible equilibrium $\mathbf{x}'$ is $(2/3, 0, 0, 1/3)$. The cost there is $5/3$, so it has gone down from the first feasible point. Now

$$\mathbf{y}^T = \begin{bmatrix} 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & -4/3 \\ 0 & 1/3 \end{bmatrix} = \begin{bmatrix} 1 & -1/3 \end{bmatrix},$$

and the feasibility test is

$$\begin{bmatrix} 1 & -1/3 \end{bmatrix} \begin{bmatrix} 0 & 3 \\ 1 & 1 \end{bmatrix} \preceq \begin{bmatrix} 1 & 3 \end{bmatrix}, \text{or} \quad \begin{bmatrix} -1/3 & 8/3 \end{bmatrix} \preceq \begin{bmatrix} 3 & 3 \end{bmatrix}.$$

which is true, so we have achieved our goal: .

27.2.12 Exercise. In the example above, use as basic columns $\mathbf{a}_3$ and $\mathbf{a}_4$. Determine if the associated equilibrium $\mathbf{x}$ is feasible, and if so, compute the cost at that vertex. Finally, how many vertices does the feasible set have? We know that there are a maximum of $\binom{4}{2} = 6$. Is the feasible set bounded?

This theorem and the example exhibits the fundamental strategy of the simplex method. We move from one basic submatrix to a new one where the associated $\mathbf{x}$ is feasible and the cost is lower. Because there are only a finite number of submatrices, and because each one is visited at most once since the cost decreases, this algorithm terminates in a finite number of steps. We either learn that the solution is unbounded negatively, or we arrive at a minimizer.

It is worth asking what can happens when the system is degenerate. If the dual equilibrium is not feasible, we can still find an index s where (27.2.1) holds, so we choose $\mathbf{a}_s$ as the entering column. Because we no longer have the non-degeneracy hypothesis, it could happen that one of the coordinates of $\mathbf{x}_B$, say x_k is 0. In that case the biggest λ one can take in the proof above is $\lambda = 0$, in which case the cost on the new basic submatrix does not decrease. Thus it is possible that the algorithm could revisit certain extreme points repeatedly and not terminate. There are more refined algorithms that prevent this phenomenon, called cycling. The simplest is perhaps the one due to Bland: see [8]. Another interesting approach is to use Remark 27.1.7 in the following way: solve the optimization problem by perturbing $\mathbf{b}$ slightly so that it is not in the boundary of a basic cone. This is due to Charnes [17].

27.3 Phase 1 of the Simplex Algorithm

Finally we deal with the issue of finding a basic submatrix B with a feasible equilibrium $\mathbf{x}$. First we rewrite our system of equations $A\mathbf{x} = \mathbf{b}$ by multiplying any row where b_i is negative by -1. The new constraint matrix is obviously equivalent to the old one. Thus we may assume that all the b_i are non-negative.

The standard method for accomplishing Phase 1 is to setting up an auxiliary asymmetric optimization problem to be solved by the method of Phase 2.

27.3.1 The Auxiliary Problem

Starting from our original constraint set $A\mathbf{x} = \mathbf{b}$, we add m new variables z_i, $1 \le i \le m$ constrained to be non-negative to our set of n variables x_j, still constrained to be non-negative. We form the $m \times (n+m)$ matrix written in block notation as $\begin{bmatrix} A & I \end{bmatrix}$ and replace the original constraints by the m constraints

$$\begin{bmatrix} A & I \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix} = \mathbf{b}$$

Thus the i-th constraint is written

$$\sum_{j=1}^{n} a_{ij} x_j + z_i = b_i.$$

The auxiliary problem is to minimize $\sum z_i$ subject to these constraints. Thus the cost vector has coordinates 0 for all the x variables, and 1 for all the z variables. We assume the problem is non-degenerate.

The key remark is that any feasible point $\mathbf{x}^*$ for our original problem produces a feasible point $[\mathbf{x}^*, \mathbf{0}]$ for the auxiliary problem, at which the cost is 0. Thus unless the feasible set of our problem is empty, the minimum cost for the auxiliary problem is at most 0.

There is an obvious basic submatrix B to start with: the m columns corresponding to the z_i. It is the identity matrix, so it has maximum rank as required. Thus we have a basic solution given by setting all the x_j to zero, and $z_i = b_i$ for all i. It is feasible because all the b_i are non-negative. Furthermore since we are assume the problem is non-degenerate, all the b_i are positive. So we can proceed with the Phase 2, using the submatrix B to start.

Next notice that the minimum cost of the auxiliary problem is bounded below by 0, since it is $\sum z_i$, and all the z_i are non-negative. Thus we know in advance that the Phase 2 algorithm will produce a basic solution.

There are two possible outcomes: either the minimal cost is 0 or not.

Assume the minimal cost is 0. Since the cost associated to each of the z_i is one, and the feasible z_i are non-negative, this shows that all the z_i are 0 in the solution of our auxiliary problem. Thus the x_j appearing in the solution of the auxiliary problem satisfy $A\mathbf{x} = \mathbf{b}$, so we have found a point in the feasible set of the original problem and we are done.

The only other possibility is that the minimal cost $\sum z_i$ is positive. As noticed in the key remark, this shows that the feasible set of our problem is empty.

27.4 Phase 2 in the Degenerate Case

In this section we present Bland's method for preventing *cycling* in the degenerate case.

First we replace the original asymmetric problem by an new problem, much in the spirit of the auxiliary problem for the first phase of the simplex method. This replacement is interesting in its own right.

27.4.1 Definition. Given our asymmetric minimization problem with n variables and m equalities constraints $A\mathbf{x} = \mathbf{b}$, and objective function $c_1x_1 + \cdots + c_nx_n$, we define a new minimization problem, called the *associated* problem as follows. Add a coordinate x_0 to the vector $\mathbf{x}$, and one new constraint

$$x_0 - c_1x_1 - \cdots - c_nx_n = 0.$$

The objective function is the function x_0. So the constraint matrix A' is now a $(m+1) \times (n+1)$ matrix, with the new equation as its 0-th row, so $b_0 = 0$. All the constraints are equality constraints.

This is not quite an asymmetric problem because we do not require the x_0 variable to be non-negative.

27.4.2 Example. Example 27.2.11 has the associated problem: minimize x_0 subject to

$$\begin{bmatrix} 1 & -1 & -2 & -3 & -3 \\ 0 & 1 & 0 & 3 & 4 \\ 0 & 0 & 1 & 1 & 3 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$$

and $x_j \geq 0$ for $1 \leq j \leq n$.

The following theorem is an analog of Theorem 25.2.2.

27.4.3 Theorem. . *The associated problem and the original problem are equivalent, in the following sense:*

1. *There is a one-to-one mapping π between the feasible sets. If*

   $$\mathbf{x}' = (x_0, x_1, \dots, x_n) \in \mathbb{R}^{n+1}$$

   is in the feasible set F' of the associated problem, then its projection

   $$\pi(\mathbf{x}') = (x_1, \dots, x_n) \in \mathbb{R}^n$$

   is in the feasible set F of the original problem. Conversely, if $(x_1, \dots, x_n) \in F$, then $(c_1x_1 + \dots + c_nx_n, x_1, \dots, x_n) \in F'$. Thus F is empty if and only if F' is.

2. *The mapping π and its inverse π^{-1} map the extreme point of F' to the extreme points of F, and conversely.*

3. *Assume that one of the two problems has a finite minimum. Then the other one does: indeed the minimizers are related as follows: If the minimum of the original problem is attained at $\mathbf{x}^* \in \mathbb{R}^n$, then the minimum of the associated problem is attained at $\mathbf{x}' = \pi^{-1}(\mathbf{x}^*)$, which in block notation is written $\begin{bmatrix}\mathbf{c}\cdot\mathbf{x}^* & \mathbf{x}^*\end{bmatrix}$. The minimum value is the same for both problems. The same relationship holds for all the extreme points.*

4. *If the feasible set for the dual problem is $F_d \subset \mathbb{R}^{m+1}$, and the feasible set of the dual of the the associated problem is $F'_d \subset \mathbb{R}^m$, then the projection map π_d that omits the 0-th coordinate gives a one-to-one map from F'_d to F_d. The inverse map π_d^{-1} takes $(y_1, \dots, y_m)$ to $(1, y_1, \dots, y_m)$. The value of the dual objective function $\mathbf{y}\cdot\mathbf{b}$ is equal to the value of the dual objection function for the associated problem at $\pi_d^{-1}(\mathbf{y})$ As a special case, the maximum of the associated problem is attained at $\mathbf{y}' = (1, \mathbf{y}^*)$, where $\mathbf{y}^*)$ is a maximizer for the dual of the original problem.*

Proof. The proof is similar to proofs we have seen before, except for the last item. Since it gives us the opportunity to compute the dual of a problem that is neither in asymmetric or symmetric form, it is worth writing down.

The associated problem, is in the form 25.1.7, so its dual is in form 25.4.5. Referring back to the notation in these problems, in our case $\mathcal{I}$ is empty, $\mathcal{J}$ is the set $\{1, \dots, m\}$, so its complement is just the index 0.

Write the dual variable as $(y_0, \mathbf{y})$, where $\mathbf{y}$ is an m-vector with indices running from 1 to m. Thus the dual problem is to maximize

$$b_0y_0 + b_1y_1 + \dots + b_my_m = \mathbf{b}\cdot\mathbf{y},$$

since $b_0 = 0$. The constraints are as follows. Write $\mathbf{a}'_j$ for the j-th column of A'. Then by Definition 25.4.5, the constraints are, recalling that the vector $\mathbf{c}'$ is $(1, 0, \dots, 0)$,

$$(y_0, \mathbf{y}) \cdot \mathbf{a}'_0 = 1, \text{ and } (y_0, \mathbf{y}) \cdot \mathbf{a}'_s \leq 0, \text{ for } s \geq 1.$$

Since $\mathbf{a}'_0 = (1, 0, \dots, 0)$, this forces $y_0 = 1$. Then the remaining m equations simplify to

$$-c_s + \mathbf{y} \cdot \mathbf{a}_s \leq 0, \text{ for } s \geq 1.$$

These are the same constraints as those for the dual feasible set of the original problem by the Duality Theorem 25.6.11, proving the result. □

The question is: why introduce the associated problem at all? The reason is simply that the constants $c_1, \dots c_n$ get absorbed into the matrix A, making it easier to formulated and prove theorems.

So now we forget about the original problem and study the the associated problem, which we write in the form:

Minimize x_0 subject to $A\mathbf{x} = \mathbf{b}$ and $x_j \geq 0$ for $j \geq 1$, where A is a $m + 1 \times n + 1$ matrix and $\mathbf{x} = (x_0, x_1, \dots, x_n)$, $\mathbf{b} = (b_0, b_1, \dots, b_m)$. As always we assume that A has maximal rank. We assume that the feasible set of this problem is non-empty, so that we can reduce to the case where

$$A = \begin{bmatrix} I_{m+1} N \end{bmatrix}, \text{ and } b_i \geq 0 \text{ for } 1 \leq i \leq m.$$

We no longer assume that the problem is nondegenerate. The goal is to produce an algorithm that solves the problem in a finite number of steps.

We first need to reexamine Rules 1 and 2 concerning the selection of the entering variable and the leaving variable for this problem, where we use as our basic submatrix the submatrix B formed by the first $m + 1$ columns of A, namely the identity matrix.

Because B is the identity matrix, for a non-basic column $\mathbf{a}_s$ we have

$$\mathbf{a}_s = a_{0s}\mathbf{a}_0 + a_{1s}\mathbf{a}_1 + \dots + a_{ms}\mathbf{a}_m,$$

so that what we were calling t_j in §27.2 (see (27.2.3)) is now a_{js}. If all the dual constraints are satisfied, the dual equilibrium vector is feasible and we are done. So assume the dual constraint for the non-basic column $\mathbf{a}_s$ is violated, so $0 < \mathbf{y} \cdot \mathbf{a}_s$. The 0 on the right-hand side comes from the fact that the cost vector is $(1, 0, \dots, 0)$. This equation is the translation of (27.2.4) in our current notation. Then the column $\mathbf{a}_s$ can enter the basis.

Next we turn to the leaving basic variable j . As before we require $a_{js} \neq 0$, so that the new set of basic variables still forms a basis. Because some of the coefficents x_k in (27.2.7) could be negative, the best we can do is to take $\lambda = 0$, so the cost at the new basic submatrix has not changed.

As we repeat this process we could get stuck without ever decreasing the cost, so the algorithm could *cycle* without terminating.

27.4.4 Example. This example, the simplest I know of, comes from [29], §2. We rewrite it as a standard symmetric minimization problem with $n = 4$, $m = 2$. Then $\mathbf{c} = (-2.3, -2.15, 13.55, 0.4)$, $\mathbf{b} = (0, 0)$ and

$$A = \begin{bmatrix} -0.4 & -0.2 & 1.4 & 0.2 \\ 7.8 & 1.4 & -7.8 & -1.4 \end{bmatrix}$$

so as always we are minimizing $\mathbf{c} \cdot \mathbf{x}$ subject to $A\mathbf{x} \succeq \mathbf{0}$ and $\mathbf{x} \succeq \mathbf{0}$. We add two slack variables x_5 and x_6, and get the asymmetric minimization problem with matrix (which we still call A)

$$\begin{bmatrix} -0.4 & -0.2 & 1.4 & 0.2 & -1 & 0 \\ 7.8 & 1.4 & -7.8 & -1.4 & 0 & -1 \end{bmatrix}$$

and $\mathbf{c} = (-2.3, -2.15, 13.55, 0.4, 0, 0)$. It is worth thinking about the cone C_A in the plane associated to A. It has four generators and $\mathbf{a}_4 = -\mathbf{a}_2$.

Then $\mathbf{x} = \mathbf{0}$ is feasible for the constraints, with cost 0. We use as basic submatrix columns 5 and 6. The first column $\mathbf{a}_1 = 0.4\mathbf{a}_1 - 7.8\mathbf{a}_2$. Since $c_1 = -2.3$, (27.2.4) is satisfied, so the first column can be used as the entering column.

The equilibrium $\mathbf{x}$ for our current basis in the origin, so the problem is degenerate. Note that $B = -I_2$, and that $\mathbf{c}_B = (0, 0)$, so that the dual equilibrium $\mathbf{y}$ is $(0, 0)$ by (25.7.2). So we choose $\mathbf{a}_5$ as the leaving column, since there is no apparent difference between $\mathbf{a}_5$ and $\mathbf{a}_6$: neither decreases cost.

You should enter this example into the simplex tool at

```
 http://www.zweigmedia.com/RealWorld/simplex.html
```

Enter

```
Minimize p = -2.3 u  -2.15 v + 13.55 w + 0.4 x subject to
-0.4 u -0.2v + 1.4w + 0.2 x >= 0
7.8 u + 1.4v -7.8w -1.4 x >=  0
```

The tool understands that all the variables are constrained to be non-negative without having to enter the constraints. The tool first adds slack variables, just as we have done. It will tell you that no optimal solution exists: indeed, the objective function goes to $-\infty$ on the feasible set. Once you have studied the anti-cycling algorithm below, you should compare

We now describe Bland's refinement of the simplex algorithm, designed to prevent cycling.

Part VIII

Nonlinear Optimization

Lecture 28

Equality Constrained Optimization and Lagrange Multipliers

We consider the problem of equality-constrained optimization in $\mathbb{R}^n$. We have a non-linear objective function $f(x_1, x_2, \dots, x_n)$ which we want to minimize subject to the equality constraints 28.3.2.

Using Lagrange multipliers, we prove the Lagrange Multiplier Theorem (28.3.9), the famous first-order necessary condition for this problem to have a solution at a point x^* that is regular (28.3.3).

We start with a simple special case in §28.1. The main theorem is stated in §28.3 and some examples given in §28.4. First we discuss the technical condition called constraint qualification, imposed by regularity, which we will meet again in other optimization problems. This condition is needed to apply the Implicit Function Theorem.

Finally the theorem is proved in §28.6. A second proof, useful for the next lecture, is given in §28.5.

The proof of the Lagrange Multiplier Theorem uses two tools:

1. The chain rule in several variables. This is covered in §12.2. You can also consult Stewart [63], §14.5, for example.

2. The Implicit Function Theorem (IFT) in §17.6. In a neighborhood of a regular point where the Lagrange condition holds, it allows us to transform the constrained problem to an unconstrained problem (studied in §13.1) that we know how to solve, at least in principle.

28.1 Lagrange Multipliers: the Simplest Case

In this section, we give the statement and the proof of the Lagrange Multiplier Theorem in a simple case: three variables and one constraint:

Minimize the $\mathcal{C}^1$ function $f(x, y, z)$ subject to the $\mathcal{C}^1$ constraint $h(x, y, z) = 0$.

Pick a feasible point $\mathbf{p} = (x^*, y^*, z^*)$, so that

$$h(x^*, y^*, z^*) = 0. \tag{28.1.1}$$

We assume that $\nabla h(\mathbf{p})$ is not the zero vector, so that one of the three partial derivatives of h at the point $\mathbf{p}$ is not zero. We assume without loss of generality that $\partial h/\partial x(\mathbf{p}) \neq 0$. Our fundamental tool, the implicit function theorem, then tells us that we can write x as a $\mathcal{C}^1$ function $g(y, z)$ in a small enough neighborhood of $\mathbf{p}$.

With these hypotheses, the Lagrange multiplier theorem says that a necessary condition for $\mathbf{p}$ to be a (local) minimizer for the constrained optimization problem is that there exist a number λ, called the Lagrange multiplier, such that

$$\nabla f(\mathbf{p}) + \lambda \nabla h(\mathbf{p}) = \mathbf{0}. \tag{28.1.2}$$

Furthermore the Lagrange multiplier is unique. This gives a system of four equations in four variables x, y, z and λ : the equation 28.1.1 saying that the point is on the constraint set and the three scalar Lagrange equations (28.1.2).

Here is the proof. On a small neighborhood of $\mathbf{p}$ in the feasible set, we can write the objective function as the composite function $F(y, z) = f(g(y, z), y, z)$. The constraint has disappeared, so that we are dealing with an unconstrained optimization problem.

Step 1. By our IFT construction, the function $h(g(y, z), y, z)$ is identically zero, so its gradient is identically 0. We compute it via the chain rule and get

$$\Big(\frac{\partial h}{\partial x}\frac{\partial g}{\partial y} + \frac{\partial h}{\partial y}, \frac{\partial h}{\partial x}\frac{\partial g}{\partial z} + \frac{\partial h}{\partial z}\Big) = \mathbf{0},$$

where we can evaluate the partials at any point on the constraint set close to $\mathbf{p}$. In particular, we can evaluate at $\mathbf{p} = (x^*, y^*, z^*)$, getting

$$\Big(\frac{\partial h}{\partial x}(\mathbf{p})\frac{\partial g}{\partial y}(y^*, z^*) + \frac{\partial h}{\partial y}(\mathbf{p}), \quad \frac{\partial h}{\partial x}(\mathbf{p})\frac{\partial g}{\partial z}(y^*, z^*) + \frac{\partial h}{\partial z}(\mathbf{p})\Big) = \mathbf{0}$$

Step 2. The composite $F(y, z)$ has an extremum at (y^*, z^*), so we compute its gradient at (y^*, z^*) using the chain rule. We get, completely analogously,

$$\Big(\frac{\partial f}{\partial x}(\mathbf{p})\frac{\partial g}{\partial y}(y^*, z^*) + \frac{\partial f}{\partial y}(\mathbf{p}), \quad \frac{\partial f}{\partial x}(\mathbf{p})\frac{\partial g}{\partial z}(y^*, z^*) + \frac{\partial f}{\partial z}(\mathbf{p})\Big) = \mathbf{0}$$

Step 3. To simplify the notation, let

$$\begin{array}{lll} f_1 = \dfrac{\partial f}{\partial x}(\mathbf{p}), & f_2 = \dfrac{\partial f}{\partial y}(\mathbf{p}), & f_3 = \dfrac{\partial f}{\partial z}(\mathbf{p}), \\ h_1 = \dfrac{\partial h}{\partial x}(\mathbf{p}), & h_2 = \dfrac{\partial h}{\partial y}(\mathbf{p}), & h_3 = \dfrac{\partial h}{\partial z}(\mathbf{p}), \\ & g_2 = \dfrac{\partial g}{\partial y}(\mathbf{p}), & g_3 = \dfrac{\partial g}{\partial z}(\mathbf{p}). \end{array}$$

Then the equations from steps 1 and 2 become:

$$\begin{aligned} h_1 g_2 + h_2 = 0, \quad h_1 g_3 + h_3 = 0, \\ f_1 g_2 + f_2 = 0, \quad f_1 g_3 + f_3 = 0. \end{aligned}$$

Since $h_1 \neq 0$ by hypothesis, divide the first line by h_1 in order to solve for g_2 and g_3:

$$g_2 = -h_2/h_1, \quad g_3 = -h_3/h_1.$$

Inserting these values into the second line, get:

$$f_1 h_2 = f_2 h_1, \quad f_1 h_3 = f_3 h_1.$$

Thus the vectors (f_1, f_2, f_3) and (h_1, h_2, h_3) are proportional with proportionality factor f_1/h_1. Then $\lambda = -f_1/h_1$ is the Lagrange multiplier. Thus we have established, in this simple case, that the Lagrange equations (28.1.2) hold at a critical point $\mathbf{p}$, if the gradient to the constraint set at $\mathbf{p}$ is not the zero vector.

28.1.3 Example. Let $f(x, y, z) = x^2 + y^2 + z^2$ be the objective function, and let the constraint be $z - x^2 - y^2 - 2 = 0$. Here the level sets of f are spheres centered at the origin, and we are looking for the level sphere of smallest radius meeting the constraint set, which is a paraboloid of rotation around the z axis. By drawing the graph, it is easy to see what the answer is, but let us proceed by a general argument. In the feasible set, z is obviously a function of x and y, and we can substitute it into $f(x, y, z)$, getting a new function:

$$F(x, y) = f(x, y, x^2 + y^2 + 2) = x^2 + y^2 + \bigl(x^2 + y^2 + 2\bigr)^2.$$

This is an unconstrained minimization problem, so we simply set the gradient of F to zero to find the critical points. We get

$$\begin{aligned} 2x + 2(x^2 + y^2 + 2)2x = 2x\bigl(1 + 2(x^2 + y^2 + 2)\bigr) = 0, \\ 2y + 2(x^2 + y^2 + 2)2y = 2y\bigl(1 + 2(x^2 + y^2 + 2)\bigr) = 0. \end{aligned}$$

and these only vanish when $x = y = 0$. Then $z = 2$. So as expected, the minimizer is at the bottom-most point of the paraboloid of revolution, which is the closest point on the constraint set to the origin.

Because we were able to solve for z explicitly in terms of x and y - which usually does not happen - we did not have to introduce Lagrange multipliers. You should work through the argument again, as in the proof, to see how the partials all appear.

The proof in the general case is more involved for two reasons. The first is the book-keeping necessary in keeping track of a larger number of variables and equations. More importantly, regularity (the condition that replaces the non-vanishing of the gradient in the case of one constraint) is more complicated and harder to work with when there is more than one constraint. We study the constraint qualification of regularity first.

28.2 Constraint Qualification

28.2.1 Definition. Let $\mathbf{x}^*$ be a feasible point for an optimization problem involving a $\mathcal{C}^1$ function $f(\mathbf{x})$. Assume that $\mathbf{x}^*$ is a local solution to the optimization problem. Then a *constraint qualification* on the constraints at $\mathbf{x}^*$ is

> a sufficient condition on the constraints at $\mathbf{x}^*$ for $\mathbf{x}^*$ to be a solution of the appropriate Lagrange equations.

In today's lecture, the constraint qualification is that $\mathbf{x}^*$ be *regular* for the constraints, as per Definition 17.1.4, which we restate in our current context in Definition 28.3.3. Then, assuming $\mathbf{x}^*$ is regular for the constraints, we establish that a necessary condition for the objective function f to have a critical point at $\mathbf{x}^*$ is that $\mathbf{x}^*$ be a solution of the Lagrange equations.

In subsequent lectures we discuss other constraint qualifications which allow us to establish similar first-order necessary conditions: see Definitions 31.2.1 and 23.7.2.

First an example satisfying regularity.

28.2.2 Example. Let $f(x, y) = x^2 + y^2$, with the linear constraint $ax + by + c = 0$. So we want to minimize this simple quadratic function on the affine line L given by the constraint. If the line goes through the origin (so $c = 0$), then the minimum occurs at the origin, the minimizer of the unconstrained function. So assume L does not do through the origin. Let's think about how L meets the level curves of f, which are circles through the origin. The minimum will occur on the level curve meeting L corresponding to the smallest value of f, meaning the circle with the smallest radius. The line could meet such a level curve in two points, but then it would meet a level curve of smaller radius. So L must meet the smallest level circle in just one point, so it must be tangent to the level circle at that point. Now the gradient $(2x, 2y)$ of f is perpendicular to the level curves, and the fixed vector (a, b) is perpendicular to the line. To say the line L is tangent, is to say that $(2x, 2y)$ points along the same line at (a, b): in other words, they are proportional with coefficient of proportionality we call λ. So we must have the vector equation $(2x, 2y) = \lambda(a, b)$ satisfied at the minimum, so $x = \lambda a/2$ and $y = \lambda b/2$. Furthermore the minimizer must be on the constraint set, so substituting these values into the equation of L, we get

$$\frac{\lambda a^2}{2} + \frac{\lambda b^2}{2} + c = 0$$

which allows us to solve for λ:

$$\lambda = \frac{-2c}{a^2 + b^2}$$

so $x = -ac/(a^2 + b^2)$ and $y = -bc/(a^2 + b^2)$.

Let us check this computation on a simple line, say the vertical line $x + 1 = 0$, so $a = 1$, $b = 0$, $c = 1$. Then the minimizer is at $(-1, 0)$, and the line is indeed tangent to the level curve $x^2 + y^2 = 1$ at the left-most point on the circle.

Note that we found the equation of the tangent line to the unit circle, which corresponds to taking $c = 1$ above, in Example 17.3.5.

Now an example in two variables where constraint qualification fails.

28.2.3 Example (The cuspidal cubic curve)**.** Let $f(x, y) = y$, so the objective is to make the y variable as small as possible. The unconstrained problem has no solution, so we introduce the constraint: $h(x, y) = x^2 - y^3 = 0$. Graph the solution set of the constraint $h(\mathbf{x}) = 0$ in the plane. This curve C is called the cuspidal cubic. Because $y^3 = x^2 \geq 0$ on C, the minimizer is a point with $y = 0$, and there is only one such point on C, the origin, which is therefore the minimizer. At the origin the gradient $\nabla h = (2x, -3y^2)$ vanishes, so that the origin is not regular for the constraints. We will see later (Example 28.4.3) that this lack of regularity, prevents us some solving this problem using Lagrange multipliers.

28.3 The First-Order Necessary Conditions: Statement

First we set up our notation. Our problem is:

28.3.1 Definition (Equality Constrained Minimization)**.** Minimize $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, subject to

$$h_1(\mathbf{x}) = h_2(\mathbf{x}) = \cdots = h_m(\mathbf{x}) = 0, \qquad (28.3.2)$$

where f and all the h_i are $\mathcal{C}^1$ functions. We assume $m < n$.

Recall the vector notation (17.1.2) for the constraints, and the notation (17.1.3) for the vector gradient of the constraints. We write S for the set where the functions f and all the h_i are defined. We usually only consider interior points of S.

We now restate Definition 17.1.4 in the optimization context.

28.3.3 Definition. We say that the point $\mathbf{x}^*$ is *regular* for the constraints 28.3.2 if the matrix $\nabla\mathbf{h}(\mathbf{x}^*)$ has maximal rank m. This is the same thing as saying that the vectors $\nabla h_i(\mathbf{x}^*)$, $1 \le i \le m$, are linearly independent in $\mathbb{R}^n$.

28.3.4 Definition (Lagrange Multipliers)**.** To each constraint h_i we attach a variable λ_i, $1 \le i \le m$, called a *Lagrange multiplier*. Then the *Lagrangian* $\mathcal{L}(\mathbf{x}, \lambda)$ associated to the problem 28.3.1 is the function

$$\begin{aligned} \mathcal{L}(\mathbf{x}, \lambda_1, \dots, \lambda_m) &= f(\mathbf{x}) + \lambda_1 h_1(\mathbf{x}) + \lambda_2 h_2(\mathbf{x}) + \cdots + \lambda_m h_m(\mathbf{x}) \qquad (28.3.5) \\ &= f(\mathbf{x}) + \lambda^T \mathbf{h}(\mathbf{x}). \end{aligned}$$

where $\mathbf{h}$ is the m-column vector of constraints, and λ is the m-column vector of Lagrange multipliers (λ_i), so $\lambda^T\mathbf{h}(\mathbf{x})$ denotes matrix multiplication.

28.3.6 Definition (Lagrange Equations)**.** The *Lagrange equations* are the n equations obtained by setting the gradient with respect to the $\mathbf{x}$ variables of the Lagrangian 28.3.5 equal to 0:

$$\nabla f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}) = \mathbf{0}. \qquad (28.3.7)$$

Individually, the j-th equation, the partial with respect to x_j, is

$$\frac{\partial f}{\partial x_j}(\mathbf{x}) + \lambda_1 \frac{\partial h_1}{\partial x_j}(\mathbf{x}) + \cdots + \lambda_m \frac{\partial h_m}{\partial x_j}(\mathbf{x}) = \mathbf{0}. \qquad (28.3.8)$$

28.3.9 Theorem (The Lagrange Multiplier Theorem)**.** *Let $\mathbf{x}^*$ be an extremum for the constrained minimization problem 28.3.1. Assume that $\mathbf{x}^*$ is regular for the constraints. Then there are unique numbers λ_1^*, ..., λ_m^*, such that the Lagrange equations* (28.3.7) *are satisfied at $\mathbf{x}^*$ and λ^*:*

$$\nabla_x \mathcal{L}(\mathbf{x}^*, \lambda^*) = \nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(\mathbf{x}^*) = 0, \quad \textit{for } 1 \le j \le n.$$

The theorem says that $\nabla f(\mathbf{x}^*)$ is a linear combination of the m vectors $\nabla h_i(\mathbf{x}^*)$. Note that we have a system of $n + m$ equations:

- the m constraints (28.3.2);
- and the n Lagrangian equations (28.3.7).

in the $n + m$ variables

- $x_1, \dots, x_n$;
- $\lambda_1, \dots \lambda_m$.

Since we have the same number of equations as we have variables, we can hope to have a finite number of solutions in $\mathbf{x}$ and in λ.

Also notice that

$$\frac{\partial \mathcal{L}}{\partial \lambda_j} = h_j,$$

so that setting the partials of the Lagrangian $\mathcal{L}$ with respect to the multipliers λ_j equal to 0 yields the constraints.

28.3.10 Corollary. *Assume $\mathbf{x}^*$ is regular for the constraints. If $\mathbf{x}^*$ is an extremum for Problem 28.3.1, it is a critical point of the Lagrangian with respect to all its variables $\mathbf{x}$ and λ.*

The Lagrange multiplier theorem is proved in §28.5. The same result applies without change if one starts from a local maximum, since all the theorem allows us to do is identify the critical points

28.4 Examples

The goal of these exercises and examples is to show you how to solve minimization problems using Lagrange multipliers. Try out:

28.4.1 Exercise. [1]

$$\text{Minimize } f(x_1, x_2, x_3) = x_1^3 + x_2^3 + x_3^3,$$
$$\text{subject to } h_1(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 - 4 = 0$$

Then add the constraint $h_2(x_1, x_2, x_3) = x_1 + x_2 + x_3 - 1 = 0$.

[1] Examples 3.5 and 3.6 in [50], p.70.

28.4.2 Remark. One way of doing the last part of this exercise uses the following interesting idea. Assume $n = m + 1$, so that the number of constraints is one less than the number of variables. Since the Lagrange Theorem says that the n-vectors ∇f, ∇h_1, $\dots \nabla h_m$ are linearly dependent, the $n \times n$ matrix whose rows are these vectors has determinant 0. Call this determinant $D(\mathbf{x})$. Note that it does not involve the λ.

If we replace the n Lagrange equation by the equation $D(\mathbf{x}) = 0$, we have replaced the system of $n + m$ equations in $n + m$ variables by a system of $1 + m = n$ equations in n variables $\mathbf{x}$. A moment's thought will convince you that you get all the solutions of the original Lagrange equations by this trick. The determinant equation is of very high degree, so it is not clear that this is worthwhile, except when the determinant, like the Vandermonde, which appears in Exercise 28.4.1 is known to factor. For more on the Vandermonde determinant see Lax [39], Appendix 1.

28.4.3 Example. We now go back to Example 28.2.3. The two Lagrange equation are

$$\begin{aligned} 0 - 2\lambda x &= 0; \\ 1 + 3\lambda y^2 &= 0. \end{aligned}$$

Obvious either $\lambda = 0$ or $x = 0$ is imposed by the first equation. But $\lambda = 0$ is impossible for the second equation, and $x = 0$ forces $y = 0$ because of the constraint equation, and we again have impossibilty. So we find no solution using Lagrange multipliers, even though the minimizer for the problem is clearly at the point $(0, 0)$, as we already noted. What went wrong? The point $(0, 0)$ is not regular. It is called a *cusp*.

Our curve, the cuspidal cubic, can be parametrized by $x(t) = t^3$ and $y = t^2$. It passes through the point of interest $(0, 0)$ at $t = 0$. Now $h(x(t), y(t)) = 0$, so our problem is to minimize $f(x, y) = y$ subject to $x(t) = t^3$ and $y = t^2$, so we get the unconstrained problem: minimize $F(t) = t^2$. So the answer is 0, and in this simple case we can bypass the Lagrange multipliers method.

28.4.4 Exercise. A more complicated example of a constraint set failing regularity at a point is given by the intersection of two spheres

$$\begin{aligned} h_1(x_1, x_2, x_3) &= x_1^2 + x_2^2 + x_3^2 = 1; \\ h_2(x_1, x_2, x_3) &= (x_1 + 1)^2 + x_2^2 + x_3^2 = 4. \end{aligned}$$

Determine the feasible set using elementary algebra. Write down $\nabla \mathbf{h}$ and compute its rank at all points of intersection. What is wrong? The intersection of the two spheres has smaller dimension than expected.

28.4.5 Remark. A word of warning. Regularity depends on how the constraint locus is described. Here is a simple example. The unit circle is given by $h(x_1, x_2) = x_1^2 + x_2^2 - 1 = 0$. With this description, every point on the circle is regular, because one of the two partials is nonzero. The unit circle can also be described by $h(x_1, x_2)^2 = 0$, since an equation vanishes if and only if its square vanishes. Now the gradient of $h(x_1, x_2)^2$ vanishes as every point where $h(x_1, x_2)$ vanishes, by the product rule for differentiation.

28.4.6 Exercise. For any set of real constants $a_1, \dots, a_n$ minimize

$$f(\mathbf{x}) = \sum_{j=1}^{n} x_j^2 \quad \text{subject to} \quad a_1x_1 + a_2x_2 + \cdots + a_nx_n = 0.$$

28.4.7 Exercise (MacLaurin)**.** For any positive integer k minimize

$$x_1^k + x_2^k + \cdots + x_n^k \quad \text{subject to} \quad x_1 + x_2 + \cdots + x_n = a,$$

for any real number a. According to Hancock [30], p. 22, this example is due to MacLaurin. Hint: Notice the symmetry when you interchange the indices.

28.5 The Proof of the Lagrange Multiplier Theorem

We prove the main theorem. A second, shorter proof is given in the next section, but this one shows explicitly how the IFT works, and generalizes to the second-order Lagrange Theorems treated in Lecture 29.

28.5.1 Notation. To simplify the notation, we write $\mathbf{x}_b$ of the first m coordinates of $\mathbf{x}$, and $\mathbf{x}_f$ for the remaining ones, so

$$\mathbf{x}_b = (x_1, \dots, x_m)\,, \mathbf{x}_f = (x_{m+1}, \dots, x_n) \text{ and } \mathbf{x} = [\mathbf{x}_b, \mathbf{x}_f]$$

We write $\nabla_b f$ for the m-vector of partials of f with respect to the first m variables, and $\nabla_f f$ for the $(n-m)$-vector of partials of f with respect to the remaining variables. Thus

$$\nabla f = [\nabla_b f, \nabla_f f]\,.$$

We let $\nabla_b \mathbf{h}$ be the square submatrix of $\nabla \mathbf{h}(\mathbf{x})$ formed by the first m columns, namely

$$\nabla_b \mathbf{h} = \begin{bmatrix} \frac{\partial h_1}{\partial x_1}(\mathbf{x}) & \dots & \frac{\partial h_1}{\partial x_m}(\mathbf{x}) \\ \vdots & \vdots & \vdots \\ \frac{\partial h_m}{\partial x_1}(\mathbf{x}) & \dots & \frac{\partial h_m}{\partial x_m}(\mathbf{x}) \end{bmatrix} \tag{28.5.2}$$

We often write $\nabla_b \mathbf{h}^*$ for the evaluation of $\nabla_b \mathbf{h}$ at $\mathbf{x}^*$

We now prove the general case of the theorem. Regularity at $\mathbf{x}^*$ means that the matrix $\nabla \mathbf{h}(\mathbf{x}^*)$ has maximal rank m, so that by reordering the variables x_j, we may assume that the first m columns of $\nabla \mathbf{h}(\mathbf{x}^*)$ are linearly independent. So $\nabla_b \mathbf{h}^*$ is invertible.

According to the implicit function theorem, the regularity at $\mathbf{x}^*$ using the first m columns of the m constraints means that x_1 through x_m can be written as C^1 functions g_1 through g_m of the remaining variables $x_{m+1} \dots x_n$, in a sufficiently small neighborhood of $\mathbf{x}^*$.

28.5.3 Definition. In order to use the chain rule computation from Theorem 12.2.3, we will extend the collection of implicit functions g_i, $1 \le i \le m$, by defining functions $g_{m+i}(\mathbf{x}_f) = x_{m+i}$, for $1 \le i \le n - m$. So

$$x_j = g_j(\mathbf{x}_f) \text{ for } 1 \le j \le n, \quad \text{or} \quad \mathbf{x} = \mathbf{g}(\mathbf{x}_f) \tag{28.5.4}$$

where $\mathbf{g}$ is now the n-vector of functions $(g_1, \dots, g_n)$.

Step 1.

With this notation, for each i, $1 \le i \le m$, we consider the composite functions

$$\psi_i(\mathbf{x}_f) = h_i(\mathbf{g}(\mathbf{x}_f) = 0 \tag{28.5.5}$$

By the IFT there are values $(x^*_{m+1}, \dots, x^*_n)$ such that

$$x^*_j = g_j(x^*_{m+1}, \dots, x^*_n) \text{ for } 1 \le j \le n, \text{ or} \quad \mathbf{x}^* = \mathbf{g}(\mathbf{x}^*_f).$$

We take the gradient of the composite function ψ_i (28.5.5) with respect to $\mathbf{x}_f$, using the chain rule (12.2.3), and get

$$\nabla \psi_i(\mathbf{x}_f) = \nabla h_i(\mathbf{g}(\mathbf{x}_f))\nabla \mathbf{g}(\mathbf{x}_f) = 0. \tag{28.5.6}$$

28.5.7 Remark. If we write $\breve{\mathbf{g}}$ for the vector function formed by the first m of the g_i, namely the ones that really come from the implicit function theorem, and also noting that the gradient of the remaining g_i, $m + 1 \le i \le n$ is the identity matrix, we get

$$(\nabla_b \mathbf{h})(\nabla \breve{\mathbf{g}}) + \nabla_f \mathbf{h},$$

so we have recovered the derivative computation in the Implicit Function Theorem, which says that this matrix is $\mathbf{0}$. Thus we could have completed Step 1 without computation by appealing to the IFT.

Step 2.

Now we turn to the objective function. We define a new function φ of the $(n-m)$ variables $\mathbf{x}_f$ by

$$\varphi(\mathbf{x}_f) = f(\mathbf{g}(\mathbf{x}_f)).$$

$\varphi(\mathbf{x}_f)$ has an unconstrained minimum at the value $\mathbf{x}^*_f$, since f has a minimum at

$$\mathbf{x}^* = \mathbf{g}(\mathbf{x}^*_f). \tag{28.5.8}$$

We have used the implicit function $\mathbf{g}$ to eliminate the variables $\mathbf{x}_b$ from f. Then by Theorem 13.1.1 on unconstrained optimization, a necessary condition for an unconstrained extremum is that the partials with respect to the remaining variables (in our case $\mathbf{x}_f$) vanish.

By the chain rule, just as in (28.5.6), we have

$$\nabla\varphi(\mathbf{x}_f) = \nabla f(\mathbf{g}(\mathbf{x}_f))\nabla\mathbf{g}(\mathbf{x}_f) = \mathbf{0}. \tag{28.5.9}$$

Step 3.

Add to (28.5.9) the sum over i, $1 \le i \le m$, of the i-th equation in (28.5.6) multiplied by the variable λ_i. Since all these equations are equal to 0, the sum is too:

$$\nabla f(\mathbf{g}(\mathbf{x}_f))\nabla\mathbf{g}(\mathbf{x}_f) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{g}(\mathbf{x}_f))\nabla\mathbf{g}(\mathbf{x}_f) = \mathbf{0}.$$

This factors as

$$\Big(\nabla f(\mathbf{g}(\mathbf{x}_f)) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{g}(\mathbf{x}_f)\Big)\nabla\mathbf{g}(\mathbf{x}_f) = \mathbf{0}.$$

Finally we evaluate at $\mathbf{x}^*$, using (28.5.8)

$$\Big(\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}^*)\Big)\nabla\mathbf{g}(\mathbf{x}^*_f) = \mathbf{0}. \tag{28.5.10}$$

This is true for any choice of λ. We will now see that we can choose a unique set of λ_i in order to satisfy the theorem. Consider the terms inside the big parentheses of (28.5.10), but only including the derivatives with respect to $\mathbf{x}_b$.

$$\nabla_b f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla_b h_i(\mathbf{x}^*) = \nabla_b f(\mathbf{x}^*) + \lambda^T \nabla_b \mathbf{h}(\mathbf{x}^*)$$

by definition of the $m \times m$ matrix $\nabla_b \mathbf{h}(\mathbf{x}^*)$: see (28.5.2).

Consider the equation

$$\nabla_b f(\mathbf{x}^*) + \lambda^T \nabla_b \mathbf{h}(\mathbf{x}^*) = \mathbf{0}. \tag{28.5.11}$$

This is a system of m linear equations in the m variables λ. The matrix of coefficients of the variables is $\nabla_b \mathbf{h}(\mathbf{x}^*)$, and it is invertible by the regularity hypothesis, Therefore there is a unique solution λ^* to the linear system (28.5.11):

$$\nabla_b f(\mathbf{x}^*) + \lambda^T \nabla_b \mathbf{h}(\mathbf{x}^*) = \mathbf{0}, \tag{28.5.12}$$

namely $\lambda^{*T} = -\nabla_b f(\mathbf{x}^*)(\nabla_b \mathbf{h}(\mathbf{x}^*))^{-1}$.

With this value for λ^*, (28.5.10) reduces to

$$\Bigl(\nabla_f f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla_f h_i(\mathbf{x}^*)\Bigr) \nabla_f \mathbf{g}(\mathbf{x}_f^*) = 0$$

By construction (see Definition 28.5.3) $\nabla_f \mathbf{g}(\mathbf{x}_f^*)$ is the identity matrix, so

$$\nabla_f f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla_f h_i(\mathbf{x}^*) = \mathbf{0}. \tag{28.5.13}$$

Now (28.5.12) and (28.5.13) show that all the partials of the Lagrangian vanish at λ^*, concluding the proof of the theorem.

28.6 A Second Proof of the Lagrange Multiplier Theorem

This second proof applies the implicit function theorem to a basis of the tangent space of the feasible set at the minimizer $\mathbf{x}^*$, one basis vector at a time.

28.6.1 Notation. We write A for the $m \times n$ matrix $\nabla \mathbf{h}(\mathbf{x}^*)$ and $\mathbf{b}$ for the row vector $\nabla f(\mathbf{x}^*)$. By Corollary 17.9.2 of the implicit function theorem applied to the constraint equations $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ at the point $\mathbf{x}^*$, for any non-zero vector $\mathbf{v}$ in the tangent space of the constraints at $\mathbf{x}^*$, there is a curve $\mathbf{x}(t)$ in the feasible set that passes through $\mathbf{x}^*$ at $t = 0$, such that $\dot{\mathbf{x}}(0) = \mathbf{v}$. The dot in $\dot{\mathbf{x}}(t)$ denotes differentiation with respect to t.

Step 1.

By the choice of $\mathbf{v}$, we have

$$A\mathbf{v} = \mathbf{0}. \tag{28.6.2}$$

Since $\mathbf{v}$ is in the tangent space of the feasible set at $\mathbf{x}^*$, this follows from Definition 17.3.1.

Step 2.

On the other hand, since the constrained function $f(\mathbf{x})$ has an extremum at $\mathbf{x}^*$, the composite function $\varphi(t) = f(\mathbf{x}(t))$ has an extremum at $t = 0$. Thus $\dot{\varphi}(0)$, its derivative at 0, vanishes. The chain rule gives

$$\dot{\varphi}(0) = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\mathbf{x}(0))\dot{x}_j(0) = 0,$$

or, in vector notation,

$$\nabla f(\mathbf{x}^*)\dot{\mathbf{x}}(0) = 0.$$

In Notation 28.6.1 this can be written via vector multiplication (since $\mathbf{b}$, a gradient, is a row vector) as:

$$\mathbf{b}\mathbf{v} = 0 \tag{28.6.3}$$

Step 3.

The regularity of $\mathbf{x}^*$ for the constraints means that the matrix A has maximum rank m. Now we use the Four Subspaces Theorem 7.2.3 applied to A. The nullspace $\mathcal{N}(A)$ of the linear map $A\mathbf{x} : \mathbb{R}^n \to \mathbb{R}^m$ has dimension $n - m$, since A has rank m. This nullspace is the tangent space to the feasible set at $\mathbf{x}^*$. By Corollary 17.9.2 of the implicit function theorem, we repeat the construction above to $n - m$ tangent vectors $\mathbf{v}_1, \dots, \mathbf{v}_{n-m}$ forming a basis for $\mathcal{N}(A)$.

So (28.6.2) and (28.6.3) hold for all $\mathbf{v}_i$, $1 \le i \le n - m$.

By the Four Subspaces Theorem, $\mathcal{R}(A^T)$ is the orthogonal of $\mathcal{N}(A)$.

The row space $\mathcal{R}(A^T)$ of the linear transformation T_{A^T}, associated to the transpose A^T of A, has as basis the columns of A^T, therefore the rows of A. By the Four Subspaces Theorem, it consists in the linear space of vectors perpendicular to $\mathcal{N}(A)$. Step 2 says that $\mathbf{b}$ is in $\mathcal{R}(A^T)$, so it can be written uniquely as a linear combination of the basis of $\mathcal{R}(A^T)$ formed by the rows of A: thus we have unique λ_i^*, $1 \le i \le m$, such that $\lambda = (\lambda_1, \dots, \lambda_m)$ satisfies

$$\mathbf{b} = -\lambda^{*T} A.$$

Using the definitions of $\mathbf{b}$ and A from Notation 28.6.1, this gives

$$\nabla f(\mathbf{x}^*) + \lambda^{*T}\nabla \mathbf{h}(\mathbf{x}^*) = \mathbf{0}, \tag{28.6.4}$$

which is the content of the Lagrange multiplier theorem, so we are done.

Lecture 29

Second Order Lagrangian Conditions

We continue to consider the problem of minimization of $f(\mathbf{x})$ in $\mathbb{R}^n$ subject to equality constraints 28.3.2. Using Lagrange multipliers, we develop the usual necessary and sufficient second-order conditions for the standard problem to have a strict minimum and just a minimum at a point x^* that is *regular*. The approach is the same as in Lecture 28: use the implicit function theorem to reduce the constrained problem to an unconstrained problem. The other main tool is the same, too: the chain rule 12.2.9. A good understanding of the tangent space of the constraint set at a regular point is required: this is covered in §17.3, and in Remark 17.6.13.

Both second-order theorems require that a certain Hessian be positive semidefinite or positive definite: this is the Hessian of the objective function restricted to the tangent space of the feasible set at the minimizer $\mathbf{x}^*$. It is difficult to determine this directly. The second part of this lecture deals with the algebra necessary to establish an easier test of positive definiteness. The key tool is the *bordered Hessian*. We give three different ways of determining positive semidefiniteness, of which the easiest to use in the bordered Hessian criterion 29.6.6. A different criterion is given in Theorem 29.8.4.

29.1 An Expression for the Hessian

Today we assume that the objective function f and the m constraints $\mathbf{h}$ are $\mathcal{C}^2$, since we need second derivatives in the key Theorems 29.3.2 and 29.4.1.

We work at a regular point $\mathbf{x}^*$ for the constraints $\mathbf{h}(\mathbf{x}) = \mathbf{0}$, so by Definition

(28.3.3) the rank of the matrix

$$A = \nabla \mathbf{h}(\mathbf{x}^*) \tag{29.1.1}$$

is m. By changing the order of the variables, we will assume that the first m columns of A form a $m \times m$ matrix A_b of rank b. and write A in block form as $\begin{bmatrix} A_b & A_f \end{bmatrix}$. We write F for the Hessian of the objective function f evaluated at $\mathbf{x}^*$, and H_i for the Hessian of the i-th constraint h_i evaluated at $\mathbf{x}^*$.

We gave two proofs of the Lagrange Multiplier Theorem 28.3.9. In §28.5, we gave a proof of this theorem that reduces the problem to an unconstrained optimization problem. We continue in the same vein today.

Let $\mathbf{g}$ denotes the extended implicit function given by Definition 28.5.3 for the constraints at $\mathbf{x}^*$. So $\mathbf{g}$ is a function of the $n-m$ variables collectively denoted $\mathbf{x}_f$, and has n coordinates. The last $n-m$ coordinate functions are the identity function:

$$g_{m+i}(\mathbf{x}_f) = x_{m+i} \text{ for } 1 \le i \le n-m$$

so that when we take the gradient of $\mathbf{g}$, that part of the matrix is the $(n-m) \times (n-m)$ identity matrix. Recall (28.5.8)

$$\mathbf{x}^* = \mathbf{g}(\mathbf{x}^*_f).$$

By the Implicit Function Theorem (see Remark 17.6.13),

$$\nabla \mathbf{g}(\mathbf{x}^*_f) = \begin{bmatrix} -A_b^{-1}A_f \\ I_{n-m} \end{bmatrix} \tag{29.1.2}$$

The matrix A and its submatrices A_b and A_f are defined in (29.1.1) The right-hand side of (29.1.2) is written as two blocks: the top block $-A_b^{-1}A_f$ is a $m \times (n-m)$ matrix, and the bottom block is the identity matrix of size $(n-m)$.

In Lecture 12, Theorem 12.2.9 we computed the Hessian Φ of any composite of the form $\varphi(\mathbf{x}_f) = f(\mathbf{g}(\mathbf{x}_f))$ at $\mathbf{x}^*$, and found that

$$\Phi = \nabla \mathbf{g}(\mathbf{x}^*_f)^T F \nabla \mathbf{g}(\mathbf{x}^*_f) + \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\mathbf{x}^*) G_j \tag{29.1.3}$$

where F is the $n \times n$ Hessian matrix of f at $\mathbf{x}^*$ and G_j is the $p \times p$ Hessian of g_j at $\mathbf{x}^*$.

The same computation works for each of the h_i. Write

$$\psi_i(\mathbf{x}_f) = h_i(\mathbf{g}(\mathbf{x}_f)) = 0 \tag{29.1.4}$$

so $\psi_i(\mathbf{x}_f)$ is the composite for the i-th constraint, Then if Ψ_i is the Hessian matrix of $\psi_i(\mathbf{x}_f)$ evaluated at $\mathbf{x}^*_f$, we have

$$\Psi_i = \nabla\mathbf{g}(\mathbf{x}^*_f)^T H_i \nabla\mathbf{g}(\mathbf{x}^*_f) + \sum_{j=1}^{n} \frac{\partial h_i}{\partial x_j}(\mathbf{x}^*) G_j = 0 \tag{29.1.5}$$

where H_i is the Hessian of h_i at $\mathbf{x}^*$.

29.2 The Hessian at a Critical Point

Next we rewrite the Hessian at a critical point of the Lagrangian.

29.2.1 Theorem. *Let f be our objective function, $\mathbf{g}$ the function given by Definition 28.5.3, and $\varphi(\mathbf{x}_f)$ the composite $f(\mathbf{g}(\mathbf{x}_f))$. At a regular point $\mathbf{x}^*$ satisfying the Lagrange equation 28.3.6*

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda^*_i \nabla h_i(\mathbf{x}^*) = 0$$

for a unique collection of Lagrange multipliers λ^*_i, the Hessian Φ of $\varphi(\mathbf{x}_f)$ at $\mathbf{x}^*_f$ can be written

$$\Phi = M^T\Big(F + \sum_{i=1}^{m} \lambda^*_i H_i\Big)M = M^T\big(F + \lambda^{*T}\mathbf{H}\big)M,$$

where M is the $n \times (n-m)$ matrix $\nabla\mathbf{g}(\mathbf{x}^*_f)$ whose columns form a basis for the tangent space of the constraint set at $\mathbf{x}^*$.

Proof. We computed the Hessian of $\varphi(\mathbf{x}_f)$ at $\mathbf{x}^*$ in (29.1.3). We add to it the following linear combination, which does not change its value since all terms Ψ_i are 0 by (29.1.5):

$$\sum_{i=1}^{m} \lambda^*_i \Psi_i = 0$$

This gives a new expression for the Hessian of φ:

$$\nabla\mathbf{g}(\mathbf{x}^*_f)^T\Big(F + \sum_{i=1}^{m} \lambda^*_i H_i\Big)\nabla\mathbf{g}(\mathbf{x}^*_f) + \sum_{j=1}^{n}\Big(\frac{\partial f}{\partial x_j}(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda^*_i \frac{\partial h_i}{\partial x_j}(\mathbf{x}^*)\Big)G_j(\mathbf{x}^*_f) \tag{29.2.2}$$

Consider the expressions indexed by j, $1 \leq j \leq n$, in the last sum in (29.2.2):

$$\frac{\partial f}{\partial x_j}(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* \frac{\partial h_i}{\partial x_j}(\mathbf{x}^*).$$

They all vanish because the Lagrange equations (28.3.6) are satisfied.

This shows that the Hessian of $\varphi(\mathbf{x}_f)$ is

$$\nabla \mathbf{g}(\mathbf{x}_f^*)^T \Big(F + \sum_{i=1}^m \lambda_i^* H_i \Big) \nabla \mathbf{g}(\mathbf{x}_f^*) \tag{29.2.3}$$

Note that $F + \sum_{i=1}^m \lambda_i^* H_i$ is a sum of symmetric $n \times n$ matrices, and so is itself a symmetric $n \times n$ matrix. The matrix $\nabla \mathbf{g}(\mathbf{x}_f^*)$ is an $n \times (n-m)$ matrix, so the matrix appearing in (29.2.3) is a symmetric $(n-m) \times (n-m)$ matrix, a reasonable candidate for the Hessian of an unconstrained function in $n-m$ variables $\mathbf{x}_f$. The effect of the multiplication by the matrix $\nabla \mathbf{g}(\mathbf{x}_f^*)$ and its transpose is to restrict the Hessian to the tangent space M at $\mathbf{x}^*$ of the constraints, since for each constraint h_i

$$\nabla h_i(\mathbf{x}^*) \nabla \mathbf{g}(\mathbf{x}_f^*) = \mathbf{0}.$$

This equation is just Corollary 17.6.8: remember that the last $n-m$ coordinates of $\mathbf{g}$ give the corresponding variable, so their gradients form the identity matrix: see Definition 28.5.3. Compare with Definition 17.3.1. □

Now just use the IFT (see (29.1.2)) to transform the expression (29.2.3) for the Hessian of $\varphi(\mathbf{x}_f)$ to

$$\begin{bmatrix} -(A_b^{-1}A_f)^T & I_{n-m} \end{bmatrix} \Big(F + \sum_{i=1}^m \lambda_i^* H_i \Big) \begin{bmatrix} -A_b^{-1}A_f \\ I_{n-m} \end{bmatrix}. \tag{29.2.4}$$

29.3 The Necessary Second Order Condition for a Minimum

If the point $\mathbf{x}^*$ is regular for the constraints, the results of Lecture 28 have reduced the problem of showing that $\mathbf{x}^*$ is a local minimum for the constrained problem to that of showing that $\mathbf{x}_f^*$ is a local minimum for the unconstrained function $\phi(\mathbf{x}_f) = f(\mathbf{g}(\mathbf{x}_f))$, as we have already noted.

The results of §13.1 give the results for unconstrained optimization: the necessary conditions for $\phi(\mathbf{x}_f)$ to have a minimum at $\mathbf{x}_f^*$ are

1. its gradient at $\mathbf{x}_f^*$ must be 0;

2. its Hessian Φ at $\mathbf{x}^*_f$ must be positive semidefinite. See Theorem 13.1.1.

We just apply this to the current situation. Write

$$L = F + \sum_{i=1}^{m} \lambda_i^* H_i, \tag{29.3.1}$$

where F and H_i are the Hessians of f and h_i respectively, and the λ_i are the Lagrange multipliers. So L is a symmetric $n \times n$ matrix.

29.3.2 Theorem. *Suppose that the feasible vector $\mathbf{x}^*$ is a local minimizer of $f(\mathbf{x})$ subject to m constraints $\mathbf{h}(\mathbf{x}) = \mathbf{0}$, and that $\mathbf{x}^*$ is regular for the constraints. Then there is a m-vector λ^* of Lagrange multipliers satisfying the Lagrange equations*

$$\nabla f(\mathbf{x}^*) + \lambda^* \nabla \mathbf{h}(\mathbf{x}^*) = \mathbf{0}$$

and the matrix

$$\begin{bmatrix} -(A_b^{-1}A_f)^T & I_{n-m} \end{bmatrix} L \begin{bmatrix} -A_b^{-1}A_f \\ I_{n-m} \end{bmatrix} \tag{29.3.3}$$

is positive semidefinite.

Proof. We know that the Hessian for the unconstrained problem must be positive semidefinite by Theorem 13.1.2. Theorem 29.2.1 and (29.2.4) show that the Hessian for the unconstrained problem is (29.3.3), as claimed. □

To repeat, the matrix in (29.3.3) is the restriction of the Hessian of f to the tangent space of the constraints at $\mathbf{x}^*$.

29.3.4 Remark. It is worth noting that if L is positive semidefinite on all $\mathbb{R}^n$, then its restriction to any linear subspace is still positive semidefinite. So it is always worth checking if this is the case. On the other hand, L could have negative eigenvalues on $\mathbb{R}^n$, while its restriction to a subspace has only non-negative eigenvalues. So in some cases we will need to compute the matrix in (29.3.3).

29.4 The Sufficient Second Order Condition for a Minimum

For unconstrained optimization, a sufficient condition for having a minimum is, besides the usual condition on the gradient, that the Hessian of the objective function be positive definite. See Theorem 13.1.2. Here is how this translates.

29.4.1 Theorem. *Suppose that $\mathbf{x}^*$ is regular for the constraints and that there is a m-vector λ^* of Lagrange multipliers satisfying the Lagrange equations*

$$\nabla f(\mathbf{x}^*) + \lambda^* \nabla \mathbf{h}(\mathbf{x}^*) = \mathbf{0}$$

Then if the matrix (29.3.3) *is positive definite, $\mathbf{x}^*$ is a strict local minimizer for f constrained to $\mathbf{h} = \mathbf{0}$.*

Proof. This follows in exactly the same way as the previous theorem from the results of §13.1 on unconstrained minima. See Theorem 13.1.3. □

29.5 Examples

Here are two examples from economics that use the second order sufficient condition, as well as the implicit function theorem and the characterization of quasiconvex functions we give in Theorem 24.4.1.

29.5.1 Cost Minimization

The objective function is linear in n variables x_j, $1 \le j \le n$. We write it

$$f(\mathbf{x}) = \sum_{j=1}^{n} p_j x_j.$$

The feasible set S is the open first quadrant: $x_j > 0$. We have one constraint (so $m = 1$), which we write

$$z - g(\mathbf{x}) = 0, \tag{29.5.1}$$

where z is a positive constant. We assume g is $\mathcal{C}^2$.

Here is how we interpret this. An industrial producer is making one product (the output) using n inputs labeled by j, $1 \le j \le n$. The unit price of the j-th input is p_j, another constant. We assume $p_j > 0$.

The function $g(\mathbf{x})$ describes how much output the producer can make from inputs $\mathbf{x}$. The constraint therefore fixes the output level at z, and the problem is to minimize the cost of producing z units of output.

Assume $\mathbf{x}^* \in S$ is a minimizer for the problem. In order to use Lagrange multipliers, we assume that the constraint is regular at $\mathbf{x}^*$, namely $\nabla g(\mathbf{x}^*) \ne \mathbf{0}$.In economics these partials are called the *marginal productivities*.

The Lagrangian is $\mathcal{L} = \sum_{j=1}^{n} p_j x_j + \lambda(z - g(\mathbf{x}))$, and the Lagrange equations are therefore the constraint (29.5.1) and the n partial derivatives of the Lagrangian with respect to the x_j:

$$p_j - \lambda \frac{\partial g}{\partial x_j}(\mathbf{x}^*) = 0.$$

The Lagrange Theorem 28.3.9 tells us that at a minimizer $\mathbf{x}^*$, there is a unique λ^* solving these equations. Solving for λ^*, assuming that all the marginal productivities are non-zero, we get

$$\frac{p_i}{p_j} = \frac{\partial g/\partial x_i(\mathbf{x}^*)}{\partial g/\partial x_j(\mathbf{x}^*)},$$

so the ratios of the marginal productivities are equal to the ratios of the unit price of the corresponding inputs.

Next we write down a second-order condition that guarantees that we are at a *strict* minimum. We want to use Theorem 29.4.1, so we assume that the Hessian of the constraint is positive definite at $\mathbf{x}^*$ when restricted to the tangent space of the constraint: the set of $\mathbf{v}$ such that $\langle \mathbf{v}, \nabla g(\mathbf{x}^*)\rangle = 0$.

The sufficient condition of Theorem 24.4.1 tells us that $-g$ is quasiconvex, so that g is quasiconcave. As we will see in Theorem 29.6.6, this implies that the $(n+1) \times (n+1)$ bordered Hessian matrix, where G is the Hessian of g,

$$\begin{bmatrix} 0 & -\nabla g(\mathbf{x}^*) \\ -\nabla g(\mathbf{x}^*)^T & -G(\mathbf{x}^*) \end{bmatrix} \tag{29.5.2}$$

has non-zero determinant.

Because g is quasiconcave, its superlevel sets S^c are convex. Our constraint says that we are working on the level set of g of level z, and $\mathbf{x}^*$ is a point of the boundary of S^z. Since S is open, we have a non-vertical supporting hyperplane at this point, and because g is $\mathcal{C}^2$, this hyperplane is given by the gradient of g at $\mathbf{x}^*$, as we saw in Theorem 22.1.2.

Now we allow the p_j and z to vary, so that the $n+1$ Lagrange equations are a system of equations not only in the $n+1$ variables x_j and λ, but also in the additional $n+1$ variables p_j and z. Matrix (29.5.2) is the matrix of partials of the Lagrange equations with respect to the variables λ, $x_1, \dots, x_n$, in that order, as you can easily check. The fact that its determinant is non-zero allows us to apply the implicit function theorem: In a small enough neighborhood of the original point $\mathbf{p}$, z, the coordinate functions of the unique minimum are $\mathcal{C}^1$ functions $\lambda(\mathbf{p}, z)$ and $x_j(\mathbf{p}, z)$. The objective function f evaluated at this family of unique minima is called the *value function*. We write it

$$v(\mathbf{p}, z) = \sum_{j=1}^{n} p_j x_j(\mathbf{p}, z) = \langle \mathbf{p}, \mathbf{x}(\mathbf{p}, z)\rangle,$$

a function of the $n+1$ variables $\mathbf{p}$ and z in a neighborhood of our starting point. The functions $\lambda(\mathbf{p}, z)$ and $x_i(\mathbf{p}, z)$ are only known implicitly but we do know their partials at the start point. So we can compute the partials of v using the chain rule. In particular, the partial with respect to z is

$$\frac{\partial v}{\partial z} = \sum_{j=1}^{n} p_j \frac{\partial x_j}{\partial z}(\mathbf{p}, z).$$

In particular, the partial derivative of the cost function with respect to z is λ, and the gradient of the cost function with respect to $\mathbf{p}$ is $\mathbf{x}(\mathbf{p}, y)$. This result is known as Shephard's lemma.

The implicit function theorem only guarantees that this is true is a small neighborhood of our start point of price and output. Because g is quasiconcave, if we make the assumptions above for all $\mathbf{p}$ and all output z, we get a result over the entire first quadrant.

29.5.2 Utility Maximization

Exceptionally in this example we write the objective function as $u(x_1, \dots, x_n)$, where u measures the utility of a basket of n goods for a household. There is one linear constraint

$$r - \sum_{i=1}^{n} p_i x_i = 0,$$

where r is a parameter measuring the income of the household, and the vector $\mathbf{p}$ gives the unit prices of the n goods. The problem is to maximize u subject to this constraint. Because the constraint is linear, all points are regular. So by the Lagrange multiplier theorem, if $\mathbf{x}^*$ is a local maximum, there exists a unique λ^* such that the Lagrange equations

$$\frac{\partial u}{\partial x_i}(\mathbf{x}^*) - \lambda^* p_i = 0, 1 \le i \le n,$$

have a solution. As in the previous example, we assume that the Lagrangian

$$\mathcal{L} = u(\mathbf{x}) + \lambda(r - \sum_{i=1}^{n} p_i x_i)$$

is negative definite when restricted to the tangent space of the constraint space: the set of $\mathbf{v}$ such that $\langle \mathbf{v}, \mathbf{p} \rangle = 0$. This means that the bordered Hessian

$$\begin{bmatrix} 0 & -\mathbf{p} \\ -\mathbf{p}^T & U(\mathbf{x}^*) \end{bmatrix} \tag{29.5.3}$$

has non-zero determinant. Here U is the Hessian of u. The invertibility of this matrix means that the Implicit Function Theorem can be used to write λ and x_i as $\mathcal{C}^1$ functions of the parameters $\mathbf{p}$ and r, on a sufficiently small neighborhood of the point we started from.

The composite function $v(\mathbf{p}, r) = u(x_1(\mathbf{p}, r), \dots, x_n(\mathbf{p}, r))$ is called the value function. Then $\partial v/\partial r(\mathbf{p}, r) = \lambda$.

29.6 Bordered Hessians: an Elementary Approach

For the rest of this lecture, we do some algebra. F stands for the Hessian of the function of interest at $\mathbf{x}^*$. In applications this is usually the Hessian of the Lagrangian. In any case it is a symmetric $n \times n$ matrix. A is the matrix of gradients of the constraints at $\mathbf{x}^*$, so it is an $m \times n$ matrix, $m < n$.

Goal : To develop a criterion for recognizing when the symmetric matrix F is positive definite (or positive semidefinite) when restricted to the linear subspace $A\mathbf{x} = 0$.

As usual we assume that A has maximum rank m, so that we may assume that the square matrix consisting of the first m columns of A is invertible. We usually call this matrix A_b, where b stands for bound, but occasionally we write it A_m to denote its size. For $r > m$ we write the submatrix of the first r columns of A as A_r, and the leading principal submatrix of F of size r as F_r.

So we forget about the optimization problem, and solve this linear algebra problem. First the direct approach.

Let D be the following $n \times n$ matrix, written in blocks:

$$D = \begin{bmatrix} A_b^{-1} & -A_b^{-1}A_f \\ 0 & I_{n-m} \end{bmatrix}. \tag{29.6.1}$$

Then

$$D^{-1} = \begin{bmatrix} A_b & A_f \\ 0 & I_{n-m} \end{bmatrix}.$$

Note that A_b^{-1} is a $m \times m$ matrix, and $-A_b^{-1}A_f$ a $m \times n$ matrix: D is invertible with determinant $\det A_b^{-1}$, so it can be used as a change of basis matrix as in §8.3.

By design, we get

$$AD = \begin{bmatrix} A_b & A_f \end{bmatrix} \begin{bmatrix} A_b^{-1} & -A_b^{-1}A_f \\ 0 & I_{n-m} \end{bmatrix} = \begin{bmatrix} I_m & 0_{m,n-m} \end{bmatrix}$$

so that the equations $A\mathbf{x} = \mathbf{0}$ become

$$z_1 = z_2 = \cdots = z_m = 0 \tag{29.6.2}$$

if we define $\mathbf{z}$ by $\mathbf{z} = D^{-1}\mathbf{x}$.

In this new coordinate system, the Hessian F is written $D^T FD$. Call this Hessian F_D. Then the restriction of F to the subspace $A\mathbf{x} = \mathbf{0}$ is simply the bottom right square submatrix of F_D of size $n - m$. Call this matrix F_f.

So we just need to determine when F_f is positive definite or positive semidefinite. For that we have the tests of §9.4 and §9.5. Problem solved.

Why do anything more than that? If there are many bound variables, meaning that there are many equations, it can be annoying to compute D because it is necessary to invert A_b. Everything that follows is done to avoid that computation. As you will see, the only computations done below are row and column operations, until we get to the test itself, which involves determining the sign of the determinant of matrices. Note that when there are many equations, so m is large, there are few minors to consider: only $n - m$. So the bordered Hessian test is easiest when the direct approach is hardest.

So first a definition.

29.6.3 Definition. Given a symmetric $n \times n$ symmetric matrix F, and a $m \times n$ matrix A of rank m, $m < n$, the associated *bordered Hessian* matrix of size $(n + m) \times (n + m)$ is the matrix

$$C = \begin{bmatrix} 0_{m\times m} & A \\ A^T & F \end{bmatrix}$$

where $0_{m\times m}$ is the zero matrix of size $m \times m$. Since F is symmetric, so is C.

29.6.4 Remark. As we noticed in Corollary 28.3.10 for the gradient, this is the Hessian of the Lagrangian with respect to λ and $\mathbf{x}$.

By definition, we denote the leading principal submatrix of size $m+r$ of C by C_r (notice the shift in indexing), so:

$$C_r = \begin{bmatrix} 0_{m\times m} & A_r \\ A_r^T & F_r \end{bmatrix} \tag{29.6.5}$$

Because of the submatrix of zeros in the upper left-hand corner of C, it is obvious that $\det C_r = 0$ when $r < m$. When $r = m$, so C_m is a $2m \times 2m$ matrix, the A-block that appears in $\det C_m$ is A_m, a square matrix, and an easy computation shows that $\det C_m = (-1)^m(\det A_m)^2$. We will establish this in the course of the proof of our main theorem:

29.6.6 Theorem (Bordered Hessian Test)**.** *For the symmetric matrix F to be positive definite when restricted to the subspace $A\mathbf{x} = 0$, it is necessary and sufficient that*

$$(-1)^m \det C_r > 0 \text{ for } \quad m+1 \le r \le n$$

In other words, we are testing the leading principal minors of C of size $2m+1$, $2m+2, \dots, m+n$.

Proof. As in §8.3 we make a linear change of basis in $\mathbb{R}^n$, using an $n \times n$ invertible matrix D giving the change of basis $\mathbf{x} = D\mathbf{z}$. Then the linear equations, in the $\mathbf{z}$-coordinate system become

$$AD\mathbf{z} = \mathbf{0},$$

and the matrix F becomes

$$\mathbf{z}^T D^T F D\mathbf{z}.$$

So let's define

$$A_D = AD \quad \text{ and } \; F_D = D^T F D.$$

A_D has rank m, just like A, since D is invertible. F_D is symmetric, and is congruent (see Definition 8.4.1) to F In the $\mathbf{z}$-basis, the bordered Hessian is written:

$$C_D = \begin{bmatrix} 0_{m\times m} & A_D \\ A_D^T & F_D \end{bmatrix}$$

For details on this change of basis process, see §8.3.

Let S be the $(n+m) \times (n+m)$ matrix

$$S = \begin{bmatrix} I_m & 0_{m\times n} \\ 0_{n\times m} & D \end{bmatrix}$$

and let S_r be the leading principal matrix of size $m+r$ of S, so that the indexing is the same as that for C. Let D_r be the leading principal matrix of size r of D. Then we have:

$$S_r = \begin{bmatrix} I_m & 0 \\ 0 & D_r \end{bmatrix}$$

S and S_r are invertible because D, and therefore D_r, is.

Here is the key step of the proof of the theorem.

29.6.7 Proposition. *The matrix S transforms the bordered Hessian in the $\mathbf{x}$-basis into the bordered Hessian in the $\mathbf{z}$-basis, as follows:*

$$S^T C S = C_D \;\; \textit{and} \;\; S_r^T C_r S_r = (C_D)_r$$

The sign of $\det C_D$ is the same as that of $\det C$, and the sign of $\det\, (C_D)_r$ is the same as that of $\det C_r$.

Proof. We first multiply out S^TCS in blocks:

$$\begin{aligned}
S^TCS &= \begin{bmatrix} I_m & 0 \\ 0 & D^T \end{bmatrix} \begin{bmatrix} 0 & A \\ A^T & F \end{bmatrix} \begin{bmatrix} I_m & 0 \\ 0 & D \end{bmatrix} \\
&= \begin{bmatrix} I_m & 0 \\ 0 & D^T \end{bmatrix} \begin{bmatrix} 0 & AD \\ A^T & FD \end{bmatrix} \\
&= \begin{bmatrix} 0 & AD \\ D^TA^T & D^TFD \end{bmatrix} \\
&= \begin{bmatrix} 0 & A_D \\ (A_D)^T & F_D \end{bmatrix} \\
&= C_D
\end{aligned}$$

as required. The same proof works for the leading principal submatrices. This establishes that C and C_D are congruent, so by the Law of Inertia 8.5.5, the sign of their determinants are the same. We can also see this directly since $\det S^TCS = (\det S)^2 \det C$. C_r and $(C_D)_r$ are also congruent, so the same argument holds. □

The Proposition shows that to prove Theorem 29.6.6 it is enough to prove it in one convenient basis for F, in other words, for one choice of change of basis matrix D. Of course we use the D from (29.6.1), for the reasons described there. Thus, for the new coordinate system we have

$$C_D = \begin{bmatrix} 0_{m\times m} & I_m & 0 \\ I_m & F_b & F_{bf} \\ 0 & F_{bf}^T & F_f \end{bmatrix} \tag{29.6.8}$$

where F_D is partitioned into the four blocks that appear in the matrix: F_b is the leading $m \times m$ matrix, F_f the lowest diagonal $(n-m)\times(n-m)$ block, etc. By choice of D, we see that the restriction of the Hessian F to the subspace $A\mathbf{x} = \mathbf{0}$ is F_f. This is the matrix of interest.

By symmetric row and column operations on C_D (as in §8.6), we can reduce the bordered Hessian (29.6.8) to

$$\begin{bmatrix} 0_{m\times m} & I_m & 0_{n-m\times n-m} \\ I_m & 0 & 0 \\ 0_{n-m\times n-m} & 0 & F_f \end{bmatrix}. \tag{29.6.9}$$

The key point is that these row and column operations do not affect the bottom right corner, so we still have F_f there. These operations do not affect the sign of the determinant of the matrix or its leading principal minors, by Sylvester's Law of Inertia 8.5.5: indeed, the matrix C and the matrix in (29.6.9) are congruent. Finally, by m row exchanges we can transform the matrix in Theorem 29.6.9 to:

$$\begin{bmatrix} I_m & 0_{m\times m} & 0 \\ 0 & I_m & 0 \\ 0 & 0 & F_f \end{bmatrix} \tag{29.6.10}$$

Each row exchange multiplies the determinant by -1 (see Corollary 6.6.11), so m rows exchanges multiply the determinant by $(-1)^m$. The determinant of the matrix (29.6.10) is clearly $\det F_f$, so the determinant of the matrix (29.6.9) is $(-1)^m \det F_f$. Working backwards, this shows that the determinant of C has the sign of $(-1)^m \det F_f$.

We want F_f to be positive definite, so by the leading principal minor test of Theorem 9.4.1 its determinant and all its leading principal minors must be positive. For this to be true, we see that C and its leading principal minors must have sign $(-1)^m$, which is what the theorem says. □

29.7 A Generalization of the Rayleigh Quotient

In this section and the next, we take a different approach to the bordered Hessian. Indeed, we get an intermediate result that is of interest. Furthermore the proof uses both the Weierstrass Theorem 16.2.2 and a generalization of the Rayleigh Quotient 9.1.1.

Recall that the ordinary Rayleigh quotient associated to F is

$$R(\mathbf{x}) = \frac{\langle \mathbf{x}, F\mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle}$$

as we learned in §9.1. We replace it by a generalized Rayleigh quotient:

$$S(\mathbf{x}) = -\frac{\langle \mathbf{x}, F\mathbf{x}\rangle}{\langle A\mathbf{x}, A\mathbf{x}\rangle} \tag{29.7.1}$$

The ordinary Rayleigh quotient corresponds to the case where A is the identity matrix, with the minus sign removed. Note that S is not well-defined on the linear subspace $A\mathbf{x} = \mathbf{0}$, but like the Rayleigh quotient is constant on rays through the origin where it is defined, because it too is homogenous of degree 0. Thus it takes all its values on the unit sphere U, just like the ordinary Rayleigh quotient. We view S as a function on all of U by assigning to it the value $-\infty$ at all points of $M = (A\mathbf{x} = \mathbf{0}) \cap U$.

29.7.2 Theorem (Debreu [19]). *Assume that $\mathbf{x}^T F\mathbf{x} > 0$ for all $\mathbf{x} \in M$. Then S attains its maximum value s^* on U at a point $\mathbf{x}^*$ not in M.*

Proof. Because $\mathbf{x}^T F\mathbf{x} > 0$ for all $\mathbf{x} \in M$, it is easy to see that S is 'continuous' in the sense that for any sequence of points $\{\mathbf{x}_k\}$ in U approaching M,

$$\lim_{k\to\infty} S(\mathbf{x}_k) = -\infty$$

Thus by an extension of the Maximum Theorem 16.2.2, $S(\mathbf{x})$ attains its maximum s^* at a point $\mathbf{x}^*$ which is not in M. □

So $s^* = S(\mathbf{x}^*) \geq S(\mathbf{x})$ for all $\mathbf{x} \in U$. Clearing the denominator of S in (29.7.1), and rearranging, we get

$$\langle \mathbf{x}, F\mathbf{x}\rangle + s^*\langle A\mathbf{x}, A\mathbf{x}\rangle \geq 0 \quad \text{for all } \mathbf{x} \in U \tag{29.7.3}$$

Now $\langle A\mathbf{x}, A\mathbf{x}\rangle = \mathbf{x}^T A^T A\mathbf{x}$, where the matrix $A^T A$ is an $n \times n$ symmetric matrix. Consider the family of $n \times n$ symmetric matrices $F + s(A^T A)$ parametrized by the variable s, and evaluate at an arbitrary $\mathbf{x}$:

$$\mathbf{x}^T(F + s(A^T A))\mathbf{x} = \mathbf{x}^T F\mathbf{x} + s\mathbf{x}^T A^T A\mathbf{x} \tag{29.7.4}$$

The term $\mathbf{x}^T A^T A\mathbf{x}$ is non-negative, since it is the square of the length of a vector. When $s = s^*$ in (29.7.4), you get an equality in (29.7.3) when $\mathbf{x} = \mathbf{x}^*$. When $s > s^*$, the expression on the left in (29.7.3) is strictly positive, so we have proved:

29.7.5 Corollary. *For $s > s^*$ the symmetric matrix $F+s(A^T A)$ is positive definite.*

29.7.6 Corollary. *The following two statements are equivalent:*

- *For all nonzero $\mathbf{x}$ satisfying $A\mathbf{x} = 0$, $\mathbf{x}^T F\mathbf{x}$ is positive.*
- *There exists a $s \in \mathbb{R}$ such that $F + s(A^T A)$ is a positive definite matrix.*

Proof. First we assume that we have found a s such that $F + s(A^T A)$ is positive definite. Then by definition, for all non-zero $\mathbf{x}$ we have

$$\mathbf{x}^T(F + s(A^T A))\mathbf{x} = \mathbf{x}^T F\mathbf{x} + s\mathbf{x}^T A^T A\mathbf{x} > 0$$

For $\mathbf{x}$ satisfying $A\mathbf{x} = \mathbf{0}$, the second term vanishes, so $\mathbf{x}^T F\mathbf{x} > 0$, and we have proved the ⇐ implication.

Now we turn to the ⇒ implication. If $\mathbf{x}^T F\mathbf{x} > 0$ for all non-zero $\mathbf{x}$ such that $A\mathbf{x} = 0$, the hypothesis of Theorem 29.7.2 is satisfied, so Corollary 29.7.5 applies and we are done. Any s greater that s^* works.

□

29.8 A Criterion for Positive Definiteness on a Subspace

We establish a new criterion 29.8.5 for positiveness definiteness, and then, using Lemma 29.8.6 we get a second proof of the bordered Hessian test.

We apply the results of the previous section to A_r, the matrix formed by the first r columns of A, and F_r, the leading principal submatrix of size r of F, for any r between m and n. Recall our assumption that the matrix formed by the first m columns of A is invertible, so each A_r has rank m in the range of r indicated.

The linear transformation T_{A_r} associated to A_r, studied in §7.2, goes from the subspace $\mathbb{R}^r$ formed by the first r coordinates of $\mathbb{R}^n$ to $\mathbb{R}^m$. The nullspace $\mathcal{N}(A_r)$ of T_{A_r} has dimension $r-m$, and it is the subspace of elements of $\mathcal{N}(A_{r+1})$ with last entry equal to 0. So F_r is positive definite on $\mathcal{N}(A_r)$ if F is positive definite on $\mathcal{N}(A)$.

Thus by Corollary 29.7.6, for s sufficiently large, $F_r + sA_r^T A_r$ is positive definite on $\mathbb{R}^r$.

Note the identity

$$\begin{bmatrix} -I_m & A_r \\ sA_r^T & F_r \end{bmatrix} \begin{bmatrix} I_m & A_r \\ 0_{rm} & I_r \end{bmatrix} = \begin{bmatrix} -I_m & 0 \\ -sA_r^T & F_r + sA_r^T A_r \end{bmatrix} \tag{29.8.1}$$

29.8.2 Lemma. *Assume that F_r is positive definite on $A_r x = 0$. For s sufficient large, the sign of the determinant of the matrix in* (29.8.1)*:*

$$\begin{bmatrix} -I_m & A_r \\ sA_r^T & F_r \end{bmatrix} \tag{29.8.3}$$

is $(-1)^m$.

Proof. This is simply because the determinant of the middle matrix in (29.8.1) is 1 and the right-hand matrix has determinant of the same sign as

$$(-1)^m(\det F_r + sA_r^T A_r) = (-1)^m$$

since $F_r + sA_r^T A_r$ is positive definite by Corollary 29.7.6. □

Now we want an implication in the other direction, in order to establish a criterion for F to be positive definite on $A\mathbf{x} = \mathbf{0}$. The following theorem follows from the leading principal minor test for positive definiteness (9.4.1):

29.8.4 Theorem. *For an arbitrary $n \times n$ symmetric matrix F and an $m \times n$ matrix A such that A_m has rank m, assume that there is a s^* such that for all $s > s^*$, $F_r + sA_r^T A_r$ has positive determinant for $r = m+1, \dots, n$. Then F is positive definite on the locus M given by $Ax = 0$.*

Thus we get the

29.8.5 Corollary. *F is positive definite on $A\mathbf{x} = \mathbf{0}$ if and only if the sign of the determinant of the matrix in* (29.8.3)*) is $(-1)^m$, for $r = m+1, \dots, n$ and s sufficiently large.*

The test for positive definiteness of F on $A\mathbf{x} = \mathbf{0}$ given by Corollary 29.8.5, while beautiful, is not immediately useful because the matrix is complicated and involves a parameter s. We now replace it by a simpler matrix, the bordered Hessian: see Definition 29.6.3

29.8.6 Lemma. *The sign of the determinant of the matrix in (29.8.3), namely*

$$D_r(s) = \begin{bmatrix} -I_m & A_r \\ sA_r^T & F_r \end{bmatrix}$$

is for large enough s, the same as the sign of the determinant of C_r.

Proof. $\det D_r(s)$ is a polynomial in s of degree at most m, since A_r^T has m columns. For s large enough, the sign of the determinant is determined by the sign of the coefficient of s^m, unless it happens to be 0. The key remark is that the determinant of the matrix

$$E_r(s) = \begin{bmatrix} 0_{mm} & A_r \\ sA_r^T & F_r \end{bmatrix} \tag{29.8.7}$$

has the same leading coefficient as a polynomial in s. Finally, the determinant of C_r is the leading coefficient of this polynomial, which finishes the proof. □

Corollary 29.8.5 combined with this lemma give us a second proof of the bordered Hessian Test 29.6.6.

Lecture 30

Quadratic Optimization

By quadratic optimization we mean an optimization problem where the objective function is a quadratic polynomial in the variables and the constraints are all linear functions, which here we will take to be equality constraints. This means that these notes cover a special case of nonlinear optimization started in Lecture 28 and 29. This case is worth doing because the hypotheses mean that all points on the constraint set are regular, and because the implicit function given by the implicit function theorem can be found explicitly. In consequence, the standard problem in quadratic optimization can be solved explicitly using linear algebra. A key tool is block matrix multiplication. We cover two aspects of the solution for quadratic optimization today: unconstrained optimization, and constrained with equality constraints. Subsequently we will look at constrained optimization with both equality and inequality constraints.

30.1 Unconstrained Quadratic Optimization

We have already studied linear optimization, namely optimization problems where both the objective function and the constraints are linear. These involve different tools from ones we are currently using, because the extremum always occur on the boundary, since linear functions have no critical points. The next simplest case, where results still follow explicitly using linear algebra techniques, occurs when the objective function is a second degree polynomial. Here is the standard set up.

30.1.1 Definition. The *standard objective function* in quadratic optimization is written:

$$f(\mathbf{x}) = \frac{\mathbf{x}^T Q \mathbf{x}}{2} + \mathbf{p}^T \mathbf{x} \tag{30.1.2}$$

where Q is a symmetric $n \times n$ matrix of real numbers, $\mathbf{p}$ an n-vector of real numbers, and $\mathbf{x}$ a vector in $\mathbb{R}^n$.

As always we focus on minimization.

In this section we analyze the case where there are no constraints. We have already studied all this in a more general context. The point in doing it again is to get explicit formulas involving matrices when the objective function has this simple form.

To find a minimum or a maximum, we set the gradient at $\mathbf{x}$ to zero:

$$\nabla f|_{\mathbf{x}} = \mathbf{x}^T Q + \mathbf{p}^T = \mathbf{0} \tag{30.1.3}$$

or transposing, using the fact that Q is symmetric, we get:

$$Q\mathbf{x} + \mathbf{p} = \mathbf{0} \tag{30.1.4}$$

If Q is invertible, we can solve for $\mathbf{x}$:

$$\mathbf{x}^* = -Q^{-1}\mathbf{p} \tag{30.1.5}$$

so there is a unique candidate $\mathbf{x}^*$ for a solution.

Then, as we have learned from our study of unconstrained optimization: Theorem 13.1.3, we have:

30.1.6 Theorem. *Assuming that Q is invertible, we have three cases at the $\mathbf{x}^*$ given by* (30.1.5)*:*

- *If Q is positive definite, then x^* is the unique global minimum for f.*
- *If Q is negative definite, then x^* is the unique global maximum for f.*
- *If Q is neither, so it has both positive and negative eigenvalues, then x^* is a saddle point for f.*

Let us now focus on the case where Q is not invertible. We view Q as giving a linear map: $\mathbb{R}^n \to \mathbb{R}^n$ by right multiplication: the linear map (which we call Q) sends $\mathbf{x}$ to $Q\mathbf{x}$. If the matrix Q is not invertible, then the range of the linear map Q is not all of $\mathbb{R}^n$ and there are non-zero vectors $\mathbf{n}$ in its nullspace $\mathcal{N}(Q)$. There are two cases to consider:

1. Assume $\mathbf{p}$ is not in the range of Q. Then (30.1.4) has no solution, since as $\mathbf{x}$ varies $Q\mathbf{x}$ describes the range of Q. Therefore the minimization has no solution. Let us now examine why. Since $\mathbf{p}$ is not in the column space of Q, by symmetry of Q it is not in the row space of Q. Add $\mathbf{p}$ as an extra row

to Q, getting a new matrix $\tilde{Q}$. By construction the rank of $\tilde{Q}$ is greater than that of Q, so we can find a $\mathbf{n}$ in the nullspace of Q and not in that of $\tilde{Q}$. This means that $\mathbf{p} \cdot \mathbf{n} \neq 0$. Evaluate f on $c\mathbf{n}$, $c \in \mathbb{R} \neq 0$: $f(c\mathbf{n}) = c\mathbf{p}^T\mathbf{n} \neq 0$. By choosing c very large, either positively or negatively depending on the sign of $\mathbf{p}^T\mathbf{n}$, we can make f go to $-\infty$, showing there is no minimum.

2. Assume $\mathbf{p}$ is in the range of f, so we can write $\mathbf{p} = Q\mathbf{q}$ for a suitable $\mathbf{q}$. Then (30.1.4) can be written $Q\mathbf{x} + Q\mathbf{q} = 0$. So any $\mathbf{x}$ such that $\mathbf{x} + \mathbf{q}$ is in the nullspace of Q is a solution.

   - If Q is positive semidefinite, then any solution is a global minimum;
   - if Q is negative semidefinite, then any solution is a global maximum;
   - if if Q is neither, so it has both positive and negative eigenvalues, any solutiion is a saddle point;

Thus we have the following theorems, special cases of theorems we have seen before:

30.1.7 Theorem. *The function $f(\mathbf{x})$ given by* (30.1.2) *has a strict minimum if and only if the matrix Q is positive definite, and the unique minimizer $\mathbf{x}^*$ is then given by* (30.1.5).

30.1.8 Theorem. *The function $f(\mathbf{x})$ given by* (30.1.2) *has a minimum if and only if the matrix Q is positive semidefinite and $\mathbf{p}$ is in the range of the linear map $L : \mathbb{R}^n \to \mathbb{R}^n$ given by $L(\mathbf{x}) = Q\mathbf{x}$, in which case the function is minimized at any $\mathbf{x} \in L^{-1}(\mathbf{p})$. Then the set of minimizers is a linear space of dimension equal to that of the nullspace of L.*

30.1.9 Exercise. Let $n = 3$, and

$$Q = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

We will select $\mathbf{p}$ later.

Show that Q is positive semidefinite, with nullspace of dimension 1 generated by the vector $\mathbf{n} = (1, -1, 1)$. Show that the other eigenvalues are 1 and 3, with associated eigenvectors $(1, 0, -1)$ and $(1, 2, 1)$. As expected, the eigenvectors are orthogonal.

1. Let $\mathbf{p} = (1, 0, -1)$. Show that $\mathbf{p}$ is in the range of Q. Indeed, show that $\mathbf{p} = Q\mathbf{q}$, where $\mathbf{q} = (1, 0, -1)$. Solve the minimization problem 30.1.2. by writing

$$Q\mathbf{x} + \mathbf{p} = 0 \tag{30.1.10}$$

so

$$Q(\mathbf{x}+\mathbf{q})=0$$

which implies that $\mathbf{x}=-\mathbf{q}+t\mathbf{n}$, where t is an arbitrary real number, and $\mathbf{n}$ is the chosen generator of the nullspace. Show the minimizer is

$$\mathbf{x}^* = -\begin{bmatrix}1\\0\\-1\end{bmatrix} + t\begin{bmatrix}1\\-1\\1\end{bmatrix} = \begin{bmatrix}-1+t\\-t\\1+t\end{bmatrix}$$

Check that this $\mathbf{x}$ satisfies (30.1.10) and check that $f(\mathbf{x}^*)=0$.

2. Let $\mathbf{p}=(2,-1,0)$. Show that $\mathbf{p}$ is not in the range of Q. Compute $f(t\mathbf{n})$, for the generator $\mathbf{n}$ of the nullspace. You should get $f(t\mathbf{n})=3t$, so by letting $t\to-\infty$ we can make this arbitrarily negative, so there is no minimum.

30.2 Equality Constrained Quadratic Minimization

Next we consider the case where our objective function is the same as before (30.1.2), and we add linear equality constraints:

$$A\mathbf{x}=\mathbf{b} \tag{30.2.1}$$

where A is a $m\times n$ matrix, with $m<n$, of maximal rank m.[1] This gives m constraints, one for each row of A, written:

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \text{ for} 1 \le i \le m$$

We will typically think of the b_i as parameters.

30.2.2 Definition. The *Standard Problem* is:

$$\text{Minimize } \frac{\mathbf{x}^T Q\mathbf{x}}{2} + \mathbf{p}^T\mathbf{x} \;\text{ subject to } A\mathbf{x}=\mathbf{b} \tag{30.2.3}$$

The Lagrangian function for this problem can be written:

$$\mathcal{L}(\mathbf{x},\mu) = \frac{\mathbf{x}^T Q\mathbf{x}}{2} + \mathbf{p}^T x + \mu^T(A\mathbf{x}-\mathbf{b}) \tag{30.2.4}$$

[1] If the rank of A is less than m we remove all the dependent constraints until we get to maximal rank.

where $\mu = -\lambda$, the usual m-vector of Lagrange multipliers, with $\mu_i = -\lambda_i$ associated to the i-th constraint (thus the i-th row of A).[2]

We apply our first-order conditions to find the extrema for our optimization problem. Any solution we find will be regular, because the constraint equations are linear and linearly independent. So a minimum $\mathbf{x}^*$ is one of the solutions $(\mathbf{x}^*, \mu^*)$ to the usual system of $n+m$ equations in our $n+m$ variables $x_1, \dots, x_n, \mu_1, \dots, \mu_m$:

$$A\mathbf{x} = \mathbf{b} \tag{30.2.5}$$

$$\nabla\mathcal{L} = \mathbf{x}^T Q + \mathbf{p}^T + \mu^T A = 0 \tag{30.2.6}$$

where ∇ denotes the gradient with respect to the x-variables.[3] The $\nabla\mathcal{L}$ equations (30.2.6) form a row of equations. We transpose them to get a column of equations, remembering that $Q^T = Q$:

$$Q\mathbf{x} + \mathbf{p} + A^T\mu = 0 \tag{30.2.7}$$

If we let $\nabla_{\mu,x}$ be the gradient with respect to the $m + n$ variables μ and x, (30.2.5) and (30.2.6) can be subsumed into one equation:

$$\nabla_{\mu,x}\mathcal{L}(\mu, \mathbf{x}) = 0$$

In block matrix notation this is

$$\begin{bmatrix} 0 & A \\ A^T & Q \end{bmatrix} \begin{bmatrix} \mu \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ -\mathbf{p} \end{bmatrix} \tag{30.2.8}$$

The matrix on the left is symmetric, since Q is. It is to get symmetry here and later that we changed the sign in the Lagrangian.

Next we establish a necessary and sufficient second order condition for the Standard Problem 30.2.3 to have a strict minimum, and also to have a minimum.

The Hessian of the objective function with respect to the x variables is just Q, so it is a matrix of constants, which is what makes this case so pleasant to analyze. Furthermore the constraint locus is an intersection of hyperplanes, so that to get all tangent directions one can take lines, a fact we exploit in the proof of the next theorem.

30.2.9 Definition. The linear space $M = \{\mathbf{x} \mid A\mathbf{x} = 0\}$ of dimension $n - m$ is the *tangent space* to the affine feasible set $\{\mathbf{x} \mid A\mathbf{x} = \mathbf{b}\}$.

[2] Replacing λ by $\mu = -\lambda$, improves some formulas: see (30.2.8)

[3] It is a good exercise to check that the derivative computation is correct.

Here is why it is called the tangent space. If $\mathbf{x}^0$ is a fixed feasible point, and $\mathbf{x} + \mathbf{x}^0$ a point in the feasible set, we have

$$\mathbf{b} = A(\mathbf{x} + \mathbf{x}^0) = A\mathbf{x} + A\mathbf{x}^0 = A\mathbf{x} + \mathbf{b}$$

so $A\mathbf{x} = 0$. This is a derivative computation: write $\mathbf{x} = \sum_{i=1}^{n} \alpha_i t$ for a fixed vector $\alpha = (a_1, \dots, \alpha_n)$. Then

$$\lim_{t \to 0} \frac{A(\mathbf{x} + \mathbf{x}^0) - A(\mathbf{x}^0)}{t} = A\alpha$$

so that the linear map $A : \mathbf{x} \to A\mathbf{x}$ is the derivative of the feasible set at $\mathbf{x}^0$, and at all other points of the feasible set, for that matter.

30.2.10 Theorem (Condition for a strict minimum when the objective function is quadratic and the constraint equations are linear)**.** *A necessary and sufficient condition for the Standard Problem 30.2.3 to have a strict minimum $\mathbf{x}^*$ is that the Hessian Q be positive definite when restricted to the linear subspace M of Definition 30.2.9.*

Proof. Let $\mathbf{x}(t)$ be a line in the feasible set $A\mathbf{x} = \mathbf{b}$, parametrized by t and passing through the point $\mathbf{x}^*$ at time $t = 0$. We write it

$$x_j(t) = \alpha_j t + x_j^* \quad, 1 \le j \le n$$

where α is not the zero vector.

30.2.11 Proposition. *The line $\mathbf{x}(t)$ lies in the feasible set $\{\mathbf{x} | A\mathbf{x} = \mathbf{b}\}$ iff $\mathbf{x}^*$ is in the feasible set and $A\alpha = 0$, so that $\alpha \in M$.*

If $\mathbf{x}^*$ is a strict minimum of the objective function, then the objective function restricted to the curve $(\mathbf{x}(t), f(\mathbf{x}(t)))$, has a strict minimum at $t = 0$. Since $f(\mathbf{x}(t))$ is a degree 2 polynomial in t, this will happen if and only if the second derivative of $f(\mathbf{x}(t))$ with respect to t is positive at $t = 0$. Let us compute it.

We denote the derivative of $\mathbf{x}(t)$ with respect to t by $\dot{\mathbf{x}}$. Note that $\dot{\mathbf{x}} = \alpha$ in our notation above. By the chain rule,

$$\frac{d}{dt} f(\mathbf{x}(t)) = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j} \dot{x_j}$$

Rewriting this in matrix notation, we get

$$\frac{d}{dt} f(\mathbf{x}(t)) = \mathbf{x}^T Q \dot{\mathbf{x}} + \mathbf{p}^T \dot{\mathbf{x}}$$

Differentiating with respect to t a second time, since $\ddot{\mathbf{x}} = 0$, we get[4]

$$\frac{d^2}{dt^2} f(\mathbf{x}(t)) = \dot{\mathbf{x}}^T Q \dot{\mathbf{x}} = \alpha^T Q \alpha \tag{30.2.12}$$

If Q is positive definite on M this is strictly positive, since $\alpha \neq 0$.

As we noted in Proposition 30.2.11, we may choose the line $\mathbf{x}(t)$ so that its tangent vector $\dot{\mathbf{x}}(0) = \alpha$ is an arbitrary non-zero element of M. Thus (30.2.12) says that Q is positive definite on all of M, so this is, as claimed, a necessary and sufficient condition. □

Next we solve the Standard Problem (30.2.3) with the hypothesis from Theorem 30.2.10 that Q restricted to the constraint set is positive definite. Since the $m \times n$ matrix A is assumed to have rank m, we can assume that its leading submatrix of size $m \times m$ has rank m , and is therefore invertible. We can achieve this by reordering the variables x_i. Thus we rewrite the constraint matrix in block matrix form as

$$\begin{bmatrix} A_x & A_y \end{bmatrix}$$

with A_x an invertible $m \times m$ matrix, and A_y a $m \times (n - m)$ matrix. We rename the last $n - m$ variables:

$$y_i = x_{i+m} \text{ for } 0 < i \leq n - m$$

so we can write the constraints in block matrix form

$$\begin{bmatrix} A_x & A_y \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \mathbf{b}.$$

Multiply by A_x^{-1} on the left:

$$\begin{bmatrix} I & A_x^{-1} A_y \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = A_x^{-1} \mathbf{b} \tag{30.2.13}$$

For convenience we write

$$\mathbf{A} = A_x^{-1} A_y$$

not to be confused with the full constraint matrix, called A. Our new $\mathbf{A}$ is a $m \times (n - m)$ matrix. Also note that $\mathbf{x}$ is now a m-vector, and $\mathbf{y}$ a $(n - m)$-vector.

Multiplying out (30.2.13) gives:

$$\mathbf{x} = -\mathbf{A}\mathbf{y} + A_x^{-1} \mathbf{b} \tag{30.2.14}$$

[4] The terms $\mathbf{x}^t Q \ddot{\mathbf{x}} + \mathbf{p}^t \ddot{\mathbf{x}}$ that would usually appear vanish.

so that the matrix of partials of the functions $\mathbf{x}$ with respect to the variables $\mathbf{y}$ is

$$\nabla_y \mathbf{x} = -\mathbf{A}. \tag{30.2.15}$$

This allows us to eliminate the $\mathbf{x}$ explicitly without appealing to the implicit function theorem.

Indeed, the composite function is

$$g(\mathbf{y}) = f(-\mathbf{A}\mathbf{y} + A_x^{-1}\mathbf{b}), \tag{30.2.16}$$

which we could write out explicitly using the expression for f. By the chain rule

$$\frac{\partial g}{\partial y_k} = \sum_{i=1}^{m} \frac{\partial f}{\partial x_i}\frac{\partial x_i}{\partial y_k} + \frac{\partial f}{\partial y_k} \text{ , for } 1 \leq k \leq n-m,$$

which can be written

$$\nabla_y g = \nabla_{x,y} f \begin{bmatrix} -\mathbf{A} \\ I_{n-m} \end{bmatrix}. \tag{30.2.17}$$

Here I_{n-m} is the $(n-m) \times (n-m)$ identity matrix.

We write this out in block matrix notation. First write Q as

$$Q = \begin{bmatrix} Q_1 & Q_{12} \\ Q_{12}^T & Q_2 \end{bmatrix} \tag{30.2.18}$$

where Q_1 is a symmetric $m \times m$ matrix, Q_2 is a symmetric $(n-m) \times (n-m)$ matrix, and Q_{12} is $m \times (n-m)$. Then

$$\nabla_{x,y} f = \begin{bmatrix} \mathbf{x}^T & \mathbf{y}^T \end{bmatrix} \begin{bmatrix} Q_1 & Q_{12} \\ Q_{12}^T & Q_2 \end{bmatrix} + \mathbf{p}^T.$$

Substitute this into ((30.2.17):

$$\nabla_y g = \begin{bmatrix} \mathbf{x}^T & \mathbf{y}^T \end{bmatrix} \begin{bmatrix} Q_1 & Q_{12} \\ Q_{12}^T & Q_2 \end{bmatrix} \begin{bmatrix} -\mathbf{A} \\ I_{n-m} \end{bmatrix} + \mathbf{p}^T \begin{bmatrix} -\mathbf{A} \\ I_{n-m} \end{bmatrix}$$

or

$$\nabla_y g = -\mathbf{x}^T Q_1 \mathbf{A} - \mathbf{y}^T Q_{12}^T \mathbf{A} + \mathbf{x}^T Q_{12} + \mathbf{y}^T Q_2 + \mathbf{p}^T \begin{bmatrix} -\mathbf{A} \\ I_{n-m} \end{bmatrix} \tag{30.2.19}$$

We differentiate a second time with respect to $\mathbf{y}$ to get the Hessian G of g. Using (30.2.15), we get

$$G = \mathbf{A}^T Q_1 \mathbf{A} - Q_{12}^T \mathbf{A} - \mathbf{A}^T Q_{12} + Q_2 \tag{30.2.20}$$

We have established the following variant of Theorem 30.2.10.

30.2.21 Theorem. *A necessary and sufficient condition for the Standard Problem 30.2.3 to have a strict minimum x^* is that the Hessian G be positive definite, so that*

$$G = \mathbf{A}^T Q_1 \mathbf{A} - Q_{12}^T \mathbf{A} - \mathbf{A}^T Q_{12} + Q_2$$

is positive definite, or, equivalently, substituting $A_x^{-1}A_y$ in for $\mathbf{A}$

$$G = A_y^T (A_x^{-1})^T Q_1 A_x^{-1} A_y - Q_{12}^T A_x^{-1} A_y - A_y^T (A_x^{-1})^T Q_{12} + Q_2$$

is positive definite.

Proof. Indeed, by solving for the $\mathbf{x}$ using the $\mathbf{y}$ we have transformed our constrained minimization problem to an unconstrained one, so we just apply Theorem 30.1.7. □

Now that we have established explicitly the necessary and sufficient condition for having a strict minimum, we want to find that minimum. We go back to (30.2.19), and substitute in the value of $\mathbf{x}^T$ computed from (30.2.13) to get

$$\begin{aligned}\nabla_y g = &- \mathbf{b}^T (A_x^{-1})^T (Q_1 \mathbf{A} - Q_{12}) \\ &+ \mathbf{y}^T (\mathbf{A}^T Q_1 \mathbf{A} - Q_{12}^T \mathbf{A} - \mathbf{A}^T Q_{12} + Q_2) + \mathbf{p}^T \begin{bmatrix} -\mathbf{A} \\ I_{n-m} \end{bmatrix}\end{aligned} \tag{30.2.22}$$

Now at a minimum this gradient must be zero. Notice that the matrix that multiplies $\mathbf{y}^T$ on the right is the Hessian G appearing in Theorem 30.2.21, which by assumption is positive definite and therefore invertible. It is a symmetric matrix of size $(n-m) \times (n-m)$.

So, rearranging (30.2.22), we get

$$\mathbf{y}^T G = \mathbf{b}^T (A_x^{-1})^T (Q_1 \mathbf{A} - Q_{12}) - \mathbf{p}^T \begin{bmatrix} -\mathbf{A} \\ I_{n-m} \end{bmatrix} \tag{30.2.23}$$

Since G is invertible, we can solve for $\mathbf{y}$ by multiplying on the right by G^{-1}:

$$\mathbf{y}^T = \mathbf{b}^T (A_x^{-1})^T (Q_1 A - Q_{12}) G^{-1} - \mathbf{p}^T \begin{bmatrix} -\mathbf{A} \\ I_{n-m} \end{bmatrix} G^{-1} \tag{30.2.24}$$

Using Equation (30.2.14) and plugging in the value of $\mathbf{y}$ from (30.2.24), we get

$$\mathbf{x}^T = \mathbf{b}^T (A_x^{-1})^T (I - (Q_1 \mathbf{A} - Q_{12}) G^{-1} \mathbf{A}^T) + \mathbf{p}^T \begin{bmatrix} -\mathbf{A} \\ I_{n-m} \end{bmatrix} G^{-1} \mathbf{A}^T \tag{30.2.25}$$

and we have our explicit and unique solution $\mathbf{x}^*$. While this looks formidable, let us now see what happens in a numerical example.

30.2.26 Exercise. Use the positive semidefinite matrix Q of Exercise 30.1.9, and adjoin the constraint $x_1 - x_2 + x_3 = b$, so $m = 1$ and the new matrix $\mathbf{A}$ is $(-1, 1)$. Write y_1 and y_2 for x_2 and x_3, and write the objective function $f(x)$ (given by (30.1.2)) as a function $g(\mathbf{y})$ ((30.2.16)). Compute the Hessian G of g: see (30.2.15). You should get

$$\begin{aligned} G &= \begin{bmatrix} -1 \\ 1 \end{bmatrix} \begin{bmatrix} -1 & 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} -1 & 1 \end{bmatrix} \\ &\quad - \begin{bmatrix} -1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} + \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 5 & -1 \\ -1 & 2 \end{bmatrix} \end{aligned}$$

Show that it is positive definite.

Compute the minimizer $\mathbf{y}$. (30.2.13) tells you that

$$x = b + y_1 - y_2 \tag{30.2.27}$$

and (30.2.23) that

$$\mathbf{y}^T G = \begin{bmatrix} -2b - p_1 - p_2 & b + p_1 - p_3 \end{bmatrix}$$

To compute this you need the inverse of the Hessian G. Show that

$$G^{-1} = \frac{1}{9} \begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix}$$

Solve for $\mathbf{y}$

$$\begin{aligned} \begin{bmatrix} y_1^* & y_2^* \end{bmatrix} &= \frac{1}{9} \begin{bmatrix} -2b - p_1 - p_2 & b + p_1 - p_2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix} \\ &= \frac{1}{9} \begin{bmatrix} -3b - p_1 - 2p_2 - p_3 & 3b + 4p_1 - p_2 - 5p_3 \end{bmatrix} \end{aligned}$$

Finally substitute the value for $\mathbf{y}^*$ into (30.2.27) to get

$$x^* = b + \frac{1}{9}(-6b - 5p_1 - p_2 + 4p_3)$$

Check that the triple (x^*, y_1^*, y_2^*) is correct by checking it satisfies the constraint and by substituting it into the gradient of the Lagrangian, and seeing that we get 0 for suitable choice of the Lagrange multiplier μ:

$$\begin{pmatrix} x^* & y_1^* & y_2^* \end{pmatrix} Q + \begin{pmatrix} p_1 & p_2 & p_3 \end{pmatrix} + \mu \begin{pmatrix} 1 & -1 & 1 \end{pmatrix} = 0$$

An easy computation shows that $\mu = -1/3$ does the trick.

30.3 Simplifications if Q is Invertible

It is sometimes known in advance that Q is positive definite on the whole space, and this assumption simplifies the computations. Q is positive definite in Examples 30.4.1 and 30.4.4, for instance.

With this assumption, we can solve for $\mathbf{x}$ in (30.2.7) and get

$$\mathbf{x} = -Q^{-1}(A^T\mu + \mathbf{p}) \tag{30.3.1}$$

30.3.2 Proposition. *If Q is a positive definite $n \times n$ matrix, and A is a rectangular ($m \times n$) matrix with independent rows, then $AQ^{-1}A^T$ is positive definite.*

Proof. We need to show that for all non-zero vectors $\mathbf{x}$,

$$\mathbf{x}^T AQ^{-1}A^T\mathbf{x} > 0$$

If $A^T\mathbf{x} \neq \mathbf{0}$, this follows since Q is positive definite. On the other hand, if $A^T\mathbf{x} = \mathbf{0}$, since A^T has independent columns, then $\mathbf{x} = \mathbf{0}$. □

Writing Q_A for the positive definite $m \times m$ matrix $AQ^{-1}A^T$, we substitute the value of $\mathbf{x}$ found in (30.3.1)) into $A\mathbf{x} = b$,

$$A\mathbf{x} = -AQ^{-1}A^T\mu - AQ^{-1}\mathbf{p} = -Q_A\mu - AQ^{-1}\mathbf{p} = \mathbf{b}$$

Since Q_A is invertible, we can solve for μ:

$$\mu^* = -Q_A^{-1}\mathbf{b} - Q_A AQ^{-1}\mathbf{p} \tag{30.3.3}$$

and get the unique solution μ^* for μ. Substituting this into (30.3.1), we get

$$\mathbf{x}^* = Q^{-1}A^T Q_A^{-1}\mathbf{b} + Q^{-1}A^T Q_A AQ^{-1}\mathbf{p} - Q^{-1}\mathbf{p} \tag{30.3.4}$$

(30.3.3) and (30.3.4) give an explicit solution $(\mathbf{x}^*, \mu^*)$ in terms of $\mathbf{b}$ and $\mathbf{p}$, using only matrix algebra. Because Q_A is positive definite, the second-order test for unconstrained optimization (Theorem 13.1.3) insures that we are at a strict minimum.

Next we compute the *value function* $f^*(\mathbf{b})$ of the parameter $\mathbf{b}$. For each value of the parameter $\mathbf{b}$ near 0, the value function picks out the minimum value of the function f. In Lecture 29 we could only do this implicitly using the implicit function theorem (see §29.5). Here we can do it explicitly. This is

$$\frac{(\mathbf{x}^*)^T Q\mathbf{x}^*}{2} + \mathbf{p}^T\mathbf{x}^*$$

We substitute for $\mathbf{x}^*$ the value found in (30.3.1), getting

$$f^*(\mathbf{b}) = \frac{\mathbf{p}^t + \mu^T A)Q^{-1}QQ^{-1}(A^T\mu + \mathbf{p})}{2} - \mathbf{p}^T Q^{-1}(A^T\mu + \mathbf{p})$$

which simplifies to

$$f^*(\mathbf{b}) = \frac{(-\mathbf{p}^T + \mu^T A)Q^{-1}(A^T\mu + \mathbf{p})}{2}$$

and then to

$$f^*(\mathbf{b}) = \frac{-\mathbf{p}^T Q^{-1}\mathbf{p}}{2} + \frac{\mu^T A Q^{-1} A^T \mu}{2}$$

remembering that μ stands for the $\mu^*(\mathbf{b})$ found in 30.3.3. Recall that $Q_A = AQ^{-1}A^T$. Since the only dependence on $\mathbf{b}$ is through μ, the Hessian of $f^*(b)$ is

$$Q_A^{-1} A Q^{-1} A^T Q_A^{-1} = Q_A^{-1}.$$

By Proposition 30.3.2 we know that the symmetric matrix Q_A is positive definite: thus its inverse is too. Thus we have established directly in this case that the value function is convex: a special case of Theorem 23.6.3 that we will come to later.

Next we compute $\nabla_b f^*$ using the chain rule.

$$\nabla_b f^* = -\mu^T A Q^{-1} A^T Q_A^{-1} = -\mu^T Q_A Q_A^{-1} = -\mu^T = \lambda^T.$$

This establishes the Envelope Theorem 23.7.7 in this special case.

30.4 Exercises

30.4.1 Exercise. Let $n = 2$, $m = 1$, $Ax = x_1 + 2x_2$, and

$$Q = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$$

Show Q is positive definite.

We leave b and $\mathbf{p}$ as parameters. Then show that

$$\begin{aligned}
\mathcal{L}(x,\mu) &= x_1^2 + x_1x_2 + \frac{3}{2}x_2^2 + p_1x_1 + p_2x_2 + \mu(x_1 + 2x_2 - b) \\
\nabla\mathcal{L}(x,\mu) &= (2x_1 + x_2, x_1 + 3x_2) + (p_1, p_2) + \mu(1,2) \\
&= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} + (p_1, p_2) + \mu \begin{pmatrix} 1 & 2 \end{pmatrix}
\end{aligned}$$

in agreement with our general formula. The version of (30.2.8) for this example is

$$\begin{bmatrix} 2 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \mu \end{bmatrix} = \begin{bmatrix} -p_1 \\ -p_2 \\ b \end{bmatrix} \tag{30.4.2}$$

Show the 3×3 matrix in (30.4.2) is invertible, and compute its inverse (given below), which allows you to solve for x_1, x_2 and μ in terms of b:

$$\begin{bmatrix} x_1 \\ x_2 \\ \mu \end{bmatrix} = \frac{1}{7} \begin{bmatrix} 4 & -2 & 1 \\ -2 & 1 & 3 \\ 1 & 3 & -5 \end{bmatrix} \begin{bmatrix} -p_1 \\ -p_2 \\ b \end{bmatrix} = \frac{1}{7} \begin{bmatrix} -4p_1 + 2p_2 + b \\ 2p_1 - p_2 + 3b \\ -p_1 - 3p_2 - 5b \end{bmatrix}$$

30.4.3 Exercise (Exercise 30.4.1 continued, with $\mathbf{p} = 0$)**.** You should get the three equations in x_1, x_1, μ and the parameter b:

$$\begin{aligned} x_1 + 2x_2 &= b \\ 2x_1 + x_2 + \mu &= 0 \\ x_1 + 3x_2 + 2\mu &= 0 \end{aligned}$$

Show the inverse of the matrix Q is

$$\frac{1}{5} \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix}$$

so

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{5} \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} -\mu \\ -2\mu \end{bmatrix} = \frac{1}{5} \begin{bmatrix} -1 \\ -3 \end{bmatrix} \mu$$

Plug these values of x into the constraint equation. You should get $b = -7\mu/5$, so $\mu = -5b/7$, and

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{7} \begin{bmatrix} 1 \\ 3 \end{bmatrix} b$$

The quadratic form associated to the value function is

$$\frac{1}{2 \cdot 7^2} \begin{bmatrix} 1 & 3 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \frac{5}{2 \cdot 7}$$

so the value function itself is $f^*(b) = \frac{5}{2 \cdot 7} b^2$, its derivative df^*/db is $\frac{5}{7} b = -\mu$ as expected. Note that the last computation from Example 30.4.1 is confirmed.

30.4.4 Exercise. Let $n = 3$, $m = 2$,

$$Ax = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{30.4.5}$$

and

$$Q = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 4 \end{bmatrix} \tag{30.4.6}$$

Show Q is positive definite.

As in Exercise 30.4.1, we leave $\mathbf{b}$ and $\mathbf{p}$ as parameters. Then

$$\begin{aligned} \mathcal{L}(x,\mu) =& x_1^2 + x_1x_2 + \frac{3}{2}x_2^2 + x_2x_3 + 2x_3^2 + p_1x_1 + p_2x_2 + p_3x_3 \\ &+ \mu_1(x_1 + 2x_2 - b_1) + \mu_2(x_1 + x_2 + x_3 - b_2) \\ \nabla\mathcal{L}(x,\mu) =& (2x_1 + x_2, x_1 + 3x_2 + x_3, x_2 + 4x_3) + (p_1, p_2, p_3) \\ &+ \mu_1(1,2,0) + \mu_2(1,1,1) \\ =& \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 4 \end{bmatrix} + (p_1, p_2, p_3) \\ &+ \begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 1 \end{bmatrix} \\ =& \mathbf{x}^t Q + \mu^t A \end{aligned}$$

Show the version of 30.2.8 for this example is

$$\begin{bmatrix} 2 & 1 & 0 & 1 & 1 \\ 1 & 3 & 1 & 2 & 1 \\ 0 & 1 & 4 & 0 & 1 \\ 1 & 2 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} -p_1 \\ -p_2 \\ -p_3 \\ b_1 \\ b_2 \end{bmatrix} \tag{30.4.7}$$

Show the 5×5 matrix in (30.4.7) is invertible, so solve for $\mathbf{x}$ and μ in terms of $\mathbf{b}$ by inverting it, or, if you prefer (this is computationally preferable) by Gaussian elimination:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \mu_1 \\ \mu_2 \end{bmatrix} = \frac{1}{13} \begin{bmatrix} 4 & -2 & 2 & -3 & 10 \\ -2 & 1 & 1 & 8 & -5 \\ -2 & 1 & 1 & -5 & 8 \\ -3 & 8 & -5 & -14 & 12 \\ 10 & -5 & 8 & 12 & -27 \end{bmatrix} \begin{bmatrix} -p_1 \\ -p_2 \\ -p_3 \\ b_1 \\ b_2 \end{bmatrix}$$

$$= \frac{1}{13} \begin{bmatrix} -4p_1 + 2p_2 - 2p_3 - 3b_1 + 10b_2 \\ 2p_1 - p_2 - p_3 + 8b_1 - 5b_2 \\ 2p_1 - p_2 - p_3 - 5b_1 + 8b_2 \\ 3p_1 - 8p_2 + 5p_3 - 14b_1 + 12b_2 \\ -10p_1 + 5p_2 - 8p_3 + 12b_1 - 27b_2 \end{bmatrix}$$

This is the only candidate for a solution. It remains to check that it is a strict minimum.

30.4.8 Exercise (Exercise 30.4.4 continued, with $\mathbf{p} = 0$)**.** Carry out the general computations in this special case. Mathematica code for doing this is in the appendix.

$$Q^{-1} = \frac{1}{18} \begin{bmatrix} 11 & -4 & 1 \\ -4 & 8 & -2 \\ 1 & -2 & 5 \end{bmatrix}$$

$$Q_A = \frac{1}{18} \begin{bmatrix} 27 & 12 \\ 12 & 14 \end{bmatrix}$$

$$Q_A^{-1} = \frac{1}{13} \begin{bmatrix} 14 & -12 \\ -12 & 27 \end{bmatrix}$$

Using (30.3.3) you will get

$$\mu = \frac{-1}{13} \begin{bmatrix} 14 & -12 \\ -12 & 27 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{30.4.9}$$

So using (30.3.4)) you will get

$$\mathbf{x} = \frac{1}{13} \begin{bmatrix} -3 & 10 \\ 8 & -5 \\ -5 & 8 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{30.4.10}$$

Check the computation by multiplying (30.4.10) on the left by A: you should get the identity $\mathbf{b} = \mathbf{b}$. Show the quadratic form associated to the value function $f^*(\mathbf{b})$ is

$$\frac{1}{2 \cdot 13^2} \begin{bmatrix} -3 & 8 & -5 \\ 10 & -5 & 8 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} -3 & 10 \\ 8 & -5 \\ -5 & 8 \end{bmatrix}$$

and this multiplies out the the matrix in (30.4.9), as expected. Note that the matrix inversion computation from Example 30.4.4 is confirmed. You can decide how you prefer to carry out the computation. Since Q is positive definite, use the second-order test to show that this is a minimum.

30.4.11 Exercise (Exercise 30.4.8 continued)**.** Compute the vectors $\dot{\mathbf{x}}$ for this example. Since the feasible set is the line given by the two equations (30.4.5), $\dot{\mathbf{x}}$ satisfies the equations

$$A\dot{\mathbf{x}} = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

you can solve for $\dot{x}_2$ and $\dot{x}_1$ in terms of $\dot{x}_3$:

$$\dot{x}_1 = -2\dot{x}_3 \quad \text{and } \dot{x}_2 = \dot{x}_3$$

Use these values to evaluate (30.2.12).

$$x_1^2 + x_1x_2 + \frac{3}{2}x_2^2 + x_2x_3 + 2x_3^2$$

and get the quadratic form in one variable $\frac{13}{2}x_3^2$ which is the restriction of the quadratic form to the line, and positive definite.

We now illustrate this approach.

30.4.12 Exercise (Exercise 30.4.11 continued)**.** Compute the matrix

$$\mathbf{A}^T Q_1 \mathbf{A} - Q_{12}^T \mathbf{A} - \mathbf{A}^T Q_{12} + Q_2$$

from 30.2.20) to check that it is positive definite. Because $m = n - 1$, it is just a number, and we need to show it is positive. The original A is given by (30.4.5), and our new $\mathbf{A}$ is $A_x^{-1}A_y$. Finally Q is given by (30.4.6). From these you can read off the submatrices needed, and carry out the simple matrix multiplication, ending up with the 1×1 matrix (13) , which indeed is positive. Find $\mathbf{y}$ using (30.2.24). You should get

$$\begin{aligned} \mathbf{y}^T &= \frac{1}{13}\begin{bmatrix} b_1 & b_2 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 2 & -1 \end{bmatrix} \left(\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right) \\ &= \frac{1}{13}\begin{bmatrix} b_1 & b_2 \end{bmatrix} \begin{bmatrix} -5 \\ 8 \end{bmatrix} \\ &= \frac{1}{13}(-5b_1 + 8b_2) \end{aligned}$$

Finally, (30.2.13)) gives

$$\begin{aligned}\mathbf{x}^T &= \frac{1}{13}\begin{bmatrix} b_1 & b_2 \end{bmatrix}\left(13\begin{bmatrix} -1 & 1 \\ 2 & -1 \end{bmatrix} - \begin{bmatrix} -5 \\ 8 \end{bmatrix}\begin{bmatrix} 2 & -1 \end{bmatrix}\right) \\ &= \frac{1}{13}\begin{bmatrix} b_1 & b_2 \end{bmatrix}\begin{bmatrix} -3 & 8 \\ 10 & -5 \end{bmatrix}\end{aligned}$$

and you are done. Note that you get the same result as in (30.4.10).

30.5 Appendix

Mathematica code for the matrix computations in Example 30.4.8

$Q = \{\{2,1,0\},\{1,3,1\},\{0,1,4\}\}$	input Q
$Qinv = Inverse[Q]$	compute Q^{-1}
$A = \{\{1,2,0\},\{1,1,1\}\}$	input A
$At = Transpose[A]$	compute A^t
$QA = A.(Qinv.At)$	compute Q_A
$QAinv = Inverse[QA]$	compute $(Q_A)^{-1}$
$TripProd = Qinv.(At.QAinv)$	compute $Q^{-1}A^t(Q_A)^{-1}$
$A.TripProd$	check: identity matrix?
$xx = TripProd.bb$	answer
$Simplify[A.xx]$	check:b?

Lecture 31

The Karush-Kuhn-Tucker Conditions

This chapter studies constrained optimization with both equality and inequality constraints. We prove that the famous first-order conditions, called the Kuhn-Tucker or Karush-Kuhn-Tucker conditions, are necessary for a point $\mathbf{x}^*$ to be a minimizer of the objective function, subject as usual to some form of constraint qualification. We will look at several. The easiest one, called regularity, is a generalization of the regularity constraint we used in the Lagrange Theorem. It is too restrictive for many problems, so we also introduce a more flexible constraint qualification called CQ, and then an even more flexible tangential constraints: Definitions 31.2.3 and 31.2.7. We also examine what needs to be changed in our theorem to get a condition for $\mathbf{x}^*$ to be a local maximizer: only the sign of the Lagrange multipliers associated to the inequalities need to be reversed: they go from non-negative to non-positive.

In the next lecture we prove the usual second-order necessary and sufficient conditions.

Two simple examples in the plane are worked out.

31.1 The Standard Problem

We always assume that our problem is in the following form.

31.1.1 Definition (Inequality Constrained Minimization)**.** Minimize $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, subject to

$$h_i(\mathbf{x}) = 0 \, , 1 \leq i \leq m \text{ and } g_k(\mathbf{x}) \leq 0, 1 \leq k \leq p, \tag{31.1.2}$$

where f, h_i and g_k are $\mathcal{C}^1$ functions.

It is essential that the inequalities be written with the inequality sign as stated: $g_k(\mathbf{x}) \le 0$.

We write $\mathbf{h}(\mathbf{x})$ for the vector function of constraints from $\mathbb{R}^n \to \mathbb{R}^m$ whose coordinate functions are the h_i, and $\mathbf{g}(\mathbf{x})$ for the vector function of constraints from $\mathbb{R}^n \to \mathbb{R}^p$ whose coordinate functions are the g_k. Then, as usual,

$$\nabla\mathbf{h} \text{ is the } m \times n \text{ matrix } \left[\frac{\partial h_i}{\partial x_j}\right] \text{ and } \nabla\mathbf{g} \text{ is the } p \times n \text{ matrix } \left[\frac{\partial g_k}{\partial x_j}\right].$$

We generalize the Lagrange multipliers of Definition 28.3.4 to our new situation.

31.1.3 Definition. The *Lagrangian* $\mathcal{L}(\mathbf{x}, \lambda, \mu)$ is the function

$$\begin{aligned}\mathcal{L}(\mathbf{x}, \lambda_1, \dots, \lambda_m, \mu_1, \dots \mu_p) &= f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i h_i + \sum_{k=1}^{p} \mu_k g_k \qquad (31.1.4)\\ &= f(\mathbf{x}) + \lambda^T \mathbf{h}(\mathbf{x}) + \mu^T \mathbf{g}(\mathbf{x})\end{aligned}$$

where the λ_i and the μ_k are real variables known as the *Lagrange multipliers*. Furthermore the μ_k are non-negative. Write λ is the m-vector (λ_i), μ is the p-vector (μ_k). λ and μ are column vectors.

Let $\mathbf{x}^*$ be a point in the feasible set, so $h_i(\mathbf{x}^*) = 0$ and $g_k(\mathbf{x}^*) \le 0$ for all indices i and k.

31.1.5 Definition. If $g_k(\mathbf{x}^*) = 0$, then g_k is an *active* constraint at $\mathbf{x}^*$. Let K be the set of indices corresponding to active constraints at $\mathbf{x}^*$. We write $\mathbf{g}_K(\mathbf{x}^*)$, or just $\mathbf{g}_K$, for the collection of all active constraints at $\mathbf{x}^*$. Let p_a be the number of active constraints at $\mathbf{x}^*$, so $0 \le p_a \le p$.

The important necessary conditions for $\mathbf{x}^*$ to be a local minimizer are:

31.1.6 Definition (The KKT conditions)**.** These are the following three conditions pertaining to the Lagrangian 31.1.4.

1. There exist numbers $\lambda_1^*, \dots, \lambda_m^*$ and $\mu_1^*, \dots, \mu_p^*$ such that for each j, $1 \le j \le n$,

$$\frac{\partial f}{\partial x_j}(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \frac{\partial h_i}{\partial x_j}(\mathbf{x}^*) + \sum_{k=1}^{m} \mu_k^* \frac{\partial g_k}{\partial x_j}(\mathbf{x}^*) = 0. \qquad (31.1.7)$$

We could rewrite this as one equation of column vectors:

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(\mathbf{x}^*) + \sum_{k=1}^{m} \mu_k^* \nabla g_k(\mathbf{x}^*) = \mathbf{0},$$

or even

$$\nabla f(\mathbf{x}^*) + \lambda^{*T}\nabla\mathbf{h}(\mathbf{x}^*) + \mu^{*T}\nabla\mathbf{g}(\mathbf{x}^*) = \mathbf{0},$$

where, for example, $\mu^{*T}\nabla\mathbf{g}(\mathbf{x}^*)$ denotes matrix multiplication of the p-row vector μ^{*T} by the $p \times n$ matrix $\nabla\mathbf{g}(\mathbf{x}^*)$.

2. All the μ_k are non negative:

$$\mu_k \geq 0, \text{ for } 1 \leq k \leq p. \tag{31.1.8}$$

3. Complementary slackness holds, meaning that

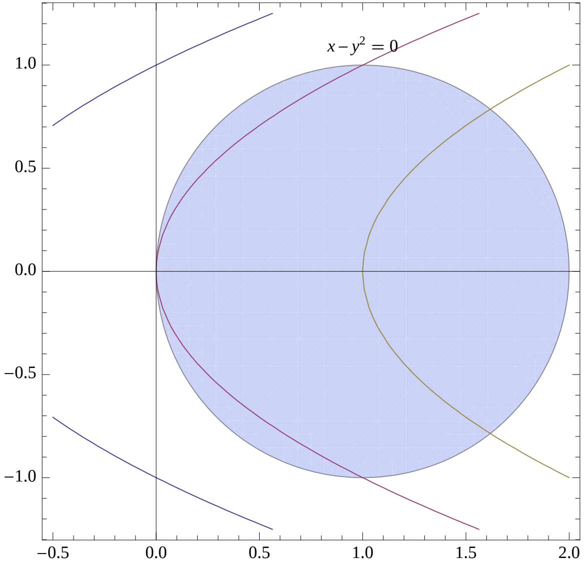
$$\mu^T\mathbf{g}(\mathbf{x}^*) = \sum_{k=1}^{p}\mu_k g_k(\mathbf{x}^*) = 0. \tag{31.1.9}$$

We could also write down a similar statement for a local maximizer are easy to write down, just by replacing the function f by $-f$, and rewriting the conditions above above. We see that the one change necessary is to replace 31.1.8 by $\mu_k \leq 0, \text{ for } 1 \leq k \leq p$.

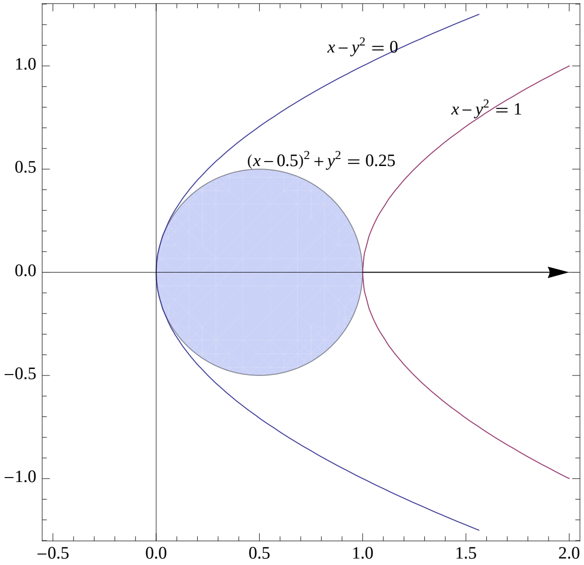
Equation 31.1.9 is called *complementary slackness* , just as in the linear case: see the Equilbrium Theorem 25.6.1. Because $\mu_k \geq 0$ by (31.1.8), and $g_k \leq 0$ on the feasible set by (31.1.2), we have $\mu_k g_k(\mathbf{x}^*) \leq 0$ for each k. Thus for the sum in (31.1.9) to be 0, each term must be 0, so either the multiplier μ_k is 0 or the constraint g_k is active at $\mathbf{x}^*$.

The goal of this lecture is to find first-order necessary conditions for $\mathbf{x}^*$ to be a local minimizer for the function f. Let's look at some simple cases first. If $\nabla f(\mathbf{x}^*) = \mathbf{0}$, then $\mathbf{x}^*$ satisfies the necessary condition for the unconstrained problem, so there is nothing more to do. So we will focus on the points $\mathbf{x}^*$ where the gradient of f is not zero. Next, if the inequality constraints are all inactive at $\mathbf{x}^*$, we can ignore them, and solve the problem using ordinary Lagrangian multipliers. We will focus on the inequality constraints.

31.1.10 Example. Let $f(x, y) = x - y^2$ and assume there is only one constraint $g(x, y) = (x-1)^2+y^2-1 \leq 0$. So the feasible set is just the closed disk of radius 1 centered at $(1, 0)$. Let the point of interest $\mathbf{x}^*$ be the origin. Then $\nabla f(\mathbf{0}) = (1, 0)$. The origin is on the boundary of the constraint set, and $\nabla g(\mathbf{0}) = (-2, 0)$, so that (as we will see later) $\mathbf{x}^*$ is regular for the constraint. Then the necessary condition of this lecture says that we can find a number $\mu \geq 0$ such that $\nabla f(\mathbf{0}) + \mu\nabla g(\mathbf{0}) = \mathbf{0}$. Here $\mu = 1/2$ works. On the boundary of the constraint, namely the circle of radius 1 given by $(x - 1)^2 + y^2 = 1$, we can solve for y^2 and then put that value into the function $f = x - y^2$. Then f takes on the value $x^2 - x$, so for small x, the circle is below the level curve $x - y^2 = 0$, showing that $\mathbf{0}$ is not a local minimum. We can see this on the graph below, where the three level curves for values -1, 0, and 1 are shown, and the feasible set is shaded. We will come back to this example later in this lecture: Example 31.4.1.



On the other hand, if we make constraint region a smaller disk, the origin becomes the local minimum. In fact, if you experiment with the disk $(x-r)^2 + y^2 - r^2 \leq 0$, you will see that the switch occurs when $r = 1/2$. Here is the graph:



This example illustrates the subtlety of optimization with inequalities.

31.2 Constraint Qualification

First, we extend the definition of regularity given in 28.3.3.

31.2.1 Definition. The point $\mathbf{x}^*$ is *regular* for the constraints 31.1.2 if the $(m+p_a)$ rows of the matrices $\nabla \mathbf{h}(\mathbf{x}^*)$ and $\nabla \mathbf{g}_K(\mathbf{x}^*)$ are linearly independent in $\mathbb{R}^n$. Thus we are considering the equality constraints and the active inequality constraints at $\mathbf{x}^*$.

If we replace the minimization problem 31.1.1 by the equality constrained problem with the same objective function, the same equality constraints and the active inequality constraints $\nabla \mathbf{g}_K \leq 0$ replaced by the equality constraints $\nabla \mathbf{g}_K = 0$, then a $\mathbf{x}^*$ that is regular for Problem 31.1.1 is still regular for the equality constraint problem. To apply the methods we developed for equality constraints, we need to assume $m + p_a < n$, which limits the number of inequality constraints.

31.2.2 Example. In Example 28.2.3, the cuspidal cubic, replace the equality constraint by the inequality constraint $g(x_1, x_2) \leq 0$. Then the point $(0, 0)$, for which the constraint is active, is not regular for the constraint according to this definition.

As in the Lagrange Theorem 28.3.9, the Kuhn-Tucker Theorem 31.3.1 says:

> If $\mathbf{x}^*$ is a solution to the minimization problem 31.1.1, then a sufficient condition for there to be a solution to the KKT conditions 31.1.6 is that $\mathbf{x}^*$ be regular.

In many situations, regularity 31.2.1 is too restrictive: we need a weaker condition that we now define.

31.2.3 Definition. The constraints (31.1.2) satisfy the qualification CQ at a point $\mathbf{x}^*$ in the feasible set if

- The m rows of the matrices $\nabla \mathbf{h}(\mathbf{x}^*)$ are linearly independent.
- The system of equations in the variable $\mathbf{z} \in \mathbb{R}^n$:

  $$\nabla \mathbf{h}(\mathbf{x}^*)\mathbf{z} = \mathbf{0}, \quad \nabla \mathbf{g}_K(\mathbf{x}^*)\mathbf{z} \prec \mathbf{0},$$

  has a non-zero solution. Here, as in Definition 31.1.5, $\mathbf{g}_K$ denotes the constraints that are active at $\mathbf{x}^*$.

31.2.4 Example. Here is an example where regularity fails, but CQ works. Take a minimization problem in the plane where the constraints are $x \leq 0$ and $y \leq 0$. The origin, where both constraints are active, is not regular for the constraints, simply because we require that $m < n$. On the other hand, CQ is satisfied. The gradients of the two inequality constraints evaluated at $(0, 0)$ are $(1, 0)$ and $(0, 1)$, so we need to find a (z_1, z_2) with $z_1 < 0$, and $z_2 < 0$. This can obviously be done.

Thus, all standard and canonical linear optimization problems fail regularity, but satisfy CQ.

The following proposition shows that CQ is less restrictive than regularity.

31.2.5 Proposition. *If the constraints* (31.1.2) *are regular at* $\mathbf{x}^*$*, they satisfy* CQ.

Proof. See 31.2.1 for the notation. The easy proof is left to you. □

31.2.6 Example. Now for an example where CQ fails. Again, we are in the plane. We have two constraints $x^2 - y \le 0$, and $y - 2x^2 \le 0$. The origin is feasible, but notice that the gradients of the two constraints are $(0, -1)$ and $(0, 1)$. We obviously cannot find a $\mathbf{z}$ with $z_2 < 0$ and $z_2 > 0$. Note the similarity with Peano's example 13.5.3: there is no line segment in the feasible set starting at the origin.

Now we turn to a third constraint qualification with an even weaker hypothesis. First a condition on vectors $\mathbf{z}$ at $\mathbf{x}^*$.

31.2.7 Definition. The non-zero vector $\mathbf{z} \in \mathbb{R}^n$ satisfies *tangential constraints* at the feasible point $\mathbf{x}^*$ for the constraints (31.1.2) if

$$\nabla \mathbf{h}(\mathbf{x}^*)\mathbf{z} = \mathbf{0}, \text{ and } \quad \nabla \mathbf{g}_K(\mathbf{x}^*)\mathbf{z} \preceq \mathbf{0}$$

where $\mathbf{g}_K$ denotes the effective inequality constraints at $\mathbf{x}^*$. We partition the active constraints K into two subsets:

- K_1 such that $\nabla \mathbf{g}_{K_1}(\mathbf{x}^*)\mathbf{z} = \mathbf{0}$;
- K_2 such that $\nabla \mathbf{g}_{K_2}(\mathbf{x}^*)\mathbf{z} \prec \mathbf{0}$.

Note that the set of active constraints K depends on $\mathbf{x}^*$ and then K_1 and K_2 depend also on $\mathbf{z}$. Let p_1 be the number of constraints in K_1, and p_2 be the number of constraints in K_2, so $p_1 + p_2 = p_a$. Assuming that the equality constraints and the inequality constraints in K_1 are linearly independent, linear algebra tells us that for a non-zero $\mathbf{z}$ to satisfy tangential constraints at $\mathbf{x}^*$, we must have $m + p_1 < n$, since the associated matrix must have a non-trivial nullspace.

31.2.8 Definition. The constraints (31.1.2) satisfy *tangential constraints* at the feasible point $\mathbf{x}^*$ if there is a non-zero vector $\mathbf{z}$ satisfying tangential constraints at $\mathbf{x}^*$ with associated active constraints K_1 and K_2, such that the m rows of the matrix $\nabla \mathbf{h}(\mathbf{x}^*)$ and the p_1 rows of the matrix $\nabla \mathbf{g}_{K_1}(\mathbf{x}^*)$ are linearly independent. In particular, $m + p_1 < n$.

Even if $m + p_1$ is less than n, Example 31.2.6 shows that it may be impossible to find a vector satisfying tangential constraints. Still, we have the easy

31.2.9 Proposition. *If the constraints satisfy CQ at $\mathbf{x}^*$, then they satisfy tangential constraints at $\mathbf{x}^*$.*

Proof. CQ is just the case where K_1 is empty, so we are done. □

We now apply Corollaries 17.9.2 and 17.9.4 to get:

31.2.10 Proposition. *If $\mathbf{z}$ satisfies the tangential constraints at $\mathbf{x}^*$, there is a curve $\mathbf{x}(t)$ defined on a small enough open interval $I \subset \mathbb{R}$ containing 0 such that, if X denotes the set where all equality constraints mentioned are satisfied, then*

1. $\mathbf{x}(0) = \mathbf{x}^*$.
2. $\dot{\mathbf{x}}(0) = \mathbf{z}$*, where $\dot{\mathbf{x}}$ denotes the derivative with respect to t.*
3. $\mathbf{x}(t)$ *is* $\mathcal{C}^1$.
4. $\mathbf{x}(t) \in X$*, for all $t \in I$. In other words,* $\mathbf{h}(\mathbf{x}(t)) = \mathbf{0}$ *and* $\mathbf{g}_{K_1}(\mathbf{x}(t)) = \mathbf{0}$.
5. $\mathbf{g}_{K_2}(\mathbf{x}(t)) \prec \mathbf{0}$*, so that when $t \geq 0$ in I, $\mathbf{x}(t)$ is feasible.*

Proof. The Implicit Function Theorem applies to the set given by the equality constraints and the constraints $\mathbf{g}_{K_1}(\mathbf{x}) = \mathbf{0}$ in a neighborhood of $\mathbf{x}^*$. Then Corollary 17.9.4 gives the result. □

31.3 The Karush-Kuhn-Tucker Multiplier Theorem

We state and prove the main theorem of this lecture for Tangential Constraints 31.2.7. Thus the theorem is proved for regularity constraints and CQ, since they are stronger by Propositions 31.2.5 and 31.2.9. We first give the proof for regularity constraints, since it is conceptually simpler, and shows what we need for the general case.

31.3.1 Theorem. *Let $\mathbf{x}^*$ be a local minimizer for the standard inequality constrained minimization problem 31.1.1. Then if $\mathbf{x}^*$ is regular for the constraints, there are unique numbers $\lambda_1^*, \dots, \lambda_m^*$ and $\mu_1^*, \dots, \mu_p^*$ such that the KKT conditions 31.1.6 hold.*

Before proving the theorem, let's consider what it says when there are no equality constraints:

1) if none on the inequality constraints are active, then we deal with the problem as an unconstrained problem, so all the μ_k are 0.

2) the KKT equation 31.1.7 says $-\nabla f(\mathbf{x}^*) = \mu^T \nabla \mathbf{g}(\mathbf{x}^*)$. Since the μ_k are non-negative, $-\nabla f(\mathbf{x}^*)$ is in the finite cone C generated by the $\nabla g_k(\mathbf{x}^*)$. Assume

that $\nabla f(\mathbf{x}^*) \neq \mathbf{0}$. Consider the hyperplane H given by $\nabla f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) = 0$, and its two halfspaces H_+, where $\nabla f(\mathbf{x}^*)$ lies, and H_-. Then C lies in H_-. Thus H is a supporting hyperplane for the cone C.

Proof. Consider the equality minimization problem where we keep the same m equality constraints, and we replace the p_a active inequality constraints by the corresponding equality constraints. By hypothesis, the point $\mathbf{x}^*$ is regular for these equality constraints, so we can apply the Lagrange Multiplier Theorem 28.3.9, which tells us that we can find uniquely determined Lagrange multipliers λ_i^* and μ_k^*, for the active constraints at $\mathbf{x}^*$. For the inactive constraints, take μ to be 0. Note that complementary slackness holds. To finish the proof we need to show that $\mu_k^* \geq 0$ for each active constraint g_k. This follows from a chain rule computation, as we now see. Assume by contradiction that there is a k such that $\mu_k^* < 0$

Since the vectors $\nabla h_i(\mathbf{x}^*)$, $1 \leq i \leq m$, and $\nabla g_k(\mathbf{x}^*)$, $1 \leq k \leq p_a$, are linearly independent, for any given k_0, $1 \leq k_0 \leq p_a$, there is a vector $\mathbf{v}$ starting at $\mathbf{x}^*$ whose dot product with all these vectors except $\nabla g_{k_0}(\mathbf{x}^*)$ is 0. In addition choose $\mathbf{v}$ so that its dot product with $\nabla g_{k_0}(\mathbf{x}^*)$ is negative. Then $\mathbf{v}$ points into the feasible set. By Corollary 17.9.2 of the implicit function theorem, there is a curve $\mathbf{x}(t)$ parametrized by the real variable t, such that $\mathbf{x}(0) = \mathbf{x}^*$, and the derivative $\frac{d\mathbf{x}}{dt}(0) = \mathbf{v}$. The composite function $g_{k_0}(\mathbf{x}(t))$ must have a non-positive derivative at $t = 0$, since for feasible $\mathbf{x}$ we have $g(\mathbf{x}) \leq 0$. Thus the composite must not be increasing at $t = 0$. Compute the derivative of the composite via the chain rule, to get

$$\nabla g_{k_0}(\mathbf{x}^*)\mathbf{v} \leq 0. \tag{31.3.2}$$

On the other hand, the composite $f(\mathbf{x}(t))$ must have a non-negative derivative at $\mathbf{x}^*$, since $\mathbf{x}^*$ is a local minimum, so by the chain rule again:

$$\nabla f(\mathbf{x}^*)\mathbf{v} \geq 0. \tag{31.3.3}$$

Now dot (31.1.7) with $\mathbf{v}$ to get

$$\nabla f(\mathbf{x}^*)\mathbf{v} + \mu_{k_0}^* \nabla \mathbf{g}_{k_0}(\mathbf{x}^*)\mathbf{v} = 0.$$

Equations (31.3.2) and (31.3.3) imply $\mu_{k_0}^* \geq 0$. If $\nabla f(\mathbf{x}^*)\mathbf{v} > 0$, then $\mu_{k_0}^* > 0$, which is allowable for complementary slackness since the constraint is active. □

To show the non-negativity of the μ, we needed a large enough supply of vectors $\mathbf{v}$ at $\mathbf{x}^*$ pointing into the feasible set.

31.3.4 Theorem. *Let $\mathbf{x}^*$ be a local minimum for the standard inequality constrained minimization problem 31.1.1. Then if the constraints satisfy CQ at $\mathbf{x}^*$, there are numbers $\lambda_1, \dots, \lambda_m$ and $\mu_1, \dots, \mu_p$ such that the KKT conditions 31.1.6 hold.*

31.3.5 Example. Let us examine the asymmetric linear optimization problem 25.1.5 in this context. So the objective function is $f(\mathbf{x}) = \mathbf{c}\cdot\mathbf{x}$, the m equality constraints are $A\mathbf{x}-\mathbf{b} = 0$, for a $m\times n$ matrix A and a m-vector $\mathbf{b}$. We also have n inequality constraints that we now write $-\mathbf{x} \le \mathbf{0}$, to be consistent with the current notation.

Then the Lagrangian for the problem is

$$\mathcal{L}(\mathbf{x}, \lambda_1, \dots, \lambda_m, \mu_1, \dots \mu_n) = \mathbf{c}\cdot\mathbf{x} - \lambda^T A\mathbf{x} - \mu^T\mathbf{x} \tag{31.3.6}$$

Note that the minus sign in front of μ^T is forced on us, since the sign of μ is important, but that in front of λ^T is just to make the notation agree with past notation.

The KKT conditions say

1. There exist numbers $\lambda_1, \dots, \lambda_m$ and $\mu_1, \dots, \mu_m$ such that

$$\mathbf{c} - \lambda^T A - \mu^T = 0,$$

2. where all the μ_k are non negative:

$$\mu_k \ge 0 \text{ for } 1 \le k \le n,$$

3. and complementary slackness holds:

$$\mu\cdot\mathbf{x}^* = 0.$$

The first equation says that $\lambda^T A \preceq \mathbf{c}$ by positivity of the μ_k, so that the λ play the role of the dual variables $\mathbf{y}$, and the constraint is that the dual variable satisfy the constraint for the feasible set of the dual. The third equation says that for each j, either $\mu_j = 0$ or $x^*_j = 0$. If $\mu_j = 0$, then the j-th line of the first equation says that

$$c_j - \sum_{i=1}^{m} \lambda_i a_{ij} = 0 \quad \text{ or } (\lambda^T A)_j = c_j$$

so we recover complementary slackness in the sense of Lecture 25. Finally we should check the constraint qualification condition. We have m equality constraints, and n inequality constraints. We work at a point $\mathbf{x}^*$ in the feasible set. Then the system satisfies CQ (see Definition 31.2.3) at $\mathbf{x}^*$ because the condition reduces to finding a solution $\mathbf{z}$ of $A\mathbf{z} = \mathbf{0}$ and $z_j > 0$ for the constraints that are active at $\mathbf{x}^*$.

31.4 Examples

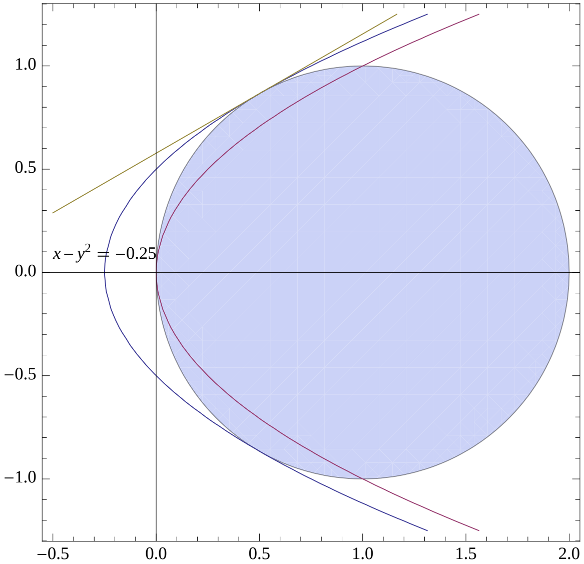
It is important to understand how to use Theorem 31.3.1 to find the minima. So let's work some examples.

31.4.1 Example. We solve Example 31.1.10. Since the Lagrangian is $\mathcal{L} = x - y^2 + \mu(x^2 - 2x + y^2)$, the KKT equations are

$$\begin{aligned} 1 + 2\mu(x-1) &= 0, \\ 2y(\mu - 1) &= 0. \end{aligned}$$

and complementary slackness must be satisfied: $\mu((x-1)^2 + y^2 - 1) = 0$. We want to solve this system for $\mu \geq 0$.

The second KKT equation gives two possibilities: either $\mu = 1$ or $y = 0$. The second case leads to the solution we already found. Our graph shows that it is not a local minimum, so our necessary condition was not sufficient. The case $\mu = 1$, then gives $x = 1/2$ and $y^2 = 3/4$. The value of f at the two points satisfying this is $-1/4$, and these are local minimizers and as the following graph shows, the global minimizers, too. The level curve of f for $-1/4$ and 0 are drawn, as is the line given both by the gradient of f and that of g at the point $(1/2, \sqrt{3}/2)$.



Now assume we want to find the maximizer. Then we study the same KKT equations, but this time for $\mu \leq 0$. Then we must have $y = 0$. An examination of the graph and the level curves shows that $x = 2$ and $\mu = -1/2$.

31.4.2 Example. Minimize $2x^2 + 2xy + y^2 - 10x - 10y$ subject to

$$x^2 + y^2 - 5 \leq 0 \text{ and } 3x + y - 6 \leq 0$$

The set determined by the first constraint is the disk of radius $\sqrt{5}$ centered at the origin, and the set determined by the second constraint is one of the half-spaces bounded by a line which intersects the circle bounding the ball in two points P_1 and P_2. These are the only points where both constraints are active.

We know, since the constraint set is compact, and the object function continuous, that there is a minimum, so the issue is just to find it.

Let's check all possible combinations of active constraints. Later we will see that one could avoid a lot of this work. Still, let us enumerate all the configuration of constraints:

1. No active constraints. We forget the constraints - assuming that neither one is active - and just consider the unconstrained minimization problem, as the theorem directs us to do. So we just set the partials of f to 0 and solve: we get

$$\begin{aligned} 4x + 2y &= 10, \\ 2x + 2y &= 10. \end{aligned}$$

   so $x = 0$ and $y = 5$. This point is not in the feasible set, so we discard this solution.

2. Only the first constraint active: we only need to introduce the multiplier μ_1 and we get the system:

$$\begin{aligned} 4x + 2y + 2\mu_1 x &= 10, \\ 2x + 2y + 2\mu_1 y &= 10, \\ x^2 + y^2 &= 5. \end{aligned}$$

   It is elementary but painful to solve this system: fortunately by inspection we notice the solution $x = 1$, $y = 2$ and $\mu_1 = 1$. μ_1 has the right sign, so the last thing we have to do is check that this solution is feasible. We plug it into the second constraint $3x + y - 6 \leq 0$ and get $3 + 2 - 6 \leq 0$, so we get a critical point and therefore, perhaps, a minimum, at $P_3 = (1, 2)$. Note that $f(1, 2) = 2 + 4 + 4 - 10 - 20 = -20$.

3. Only the second constraint active. Thus we only need to introduce the multiplier μ_2 and we get the system:

$$\begin{aligned} 4x + 2y + 3\mu_2 &= 10, \\ 2x + 2y + \mu_2 &= 10, \\ 3x + y &= 6. \end{aligned}$$

Subtracting the second equation from the first, we get $x = -\mu_2$. Then the third equation gives $y = 6 + 3\mu_2$, so substituting these values into the second equation, we get $\mu_2 = -\frac{2}{5}$. So μ_2 is negative, and we can discard this solution.

4. Both constraints active: we find the coordinates of P_1 and P_2, to see if we have local minima there. A gradient computation shows that the answer is no.

Most of this analysis is unnecessary, as we will see in Lecture 23: the objective function is convex and the feasible set is convex, so that as soon as you find the first potential solution, you can stop looking.

31.4.3 Example. Given a constant vector $\mathbf{a} = (a_1, a_2) \in \mathbb{R}^2$, minimize $f(\mathbf{x}) = x_1^2 + x_2^2 - \langle \mathbf{a}, \mathbf{x} \rangle$, subject to $x_1^2 + x_2^2 \leq 1$.

Since the feasible set is compact (it is the unit disk centered at the origin), we know there will be a finite minimum. We write the lone constraint as $g(\mathbf{x}) = x_1^2 + x_2^2 - 1 \leq 0$, to have the problem in standard form. The constraint will be regular at $\mathbf{x}^*$ unless $\nabla g(\mathbf{x}^*)$ is the 0 vector, which only happens if $\mathbf{x}^*$ is the origin, where f takes the value 0. If $\mathbf{x}^*$ is not the origin, we write the Lagrangian as $f(\mathbf{x}) + \mu g(\mathbf{x})$, so that the KKT conditions are

$$\begin{aligned} 2x_1 - a_1 + 2\mu x_1 &= 0 \\ 2x_2 - a_2 + 2\mu x_2 &= 0 \\ \mu &\geq 0 \\ \mu(x_1^2 + x_2^2 - 1) &= 0 \end{aligned}$$

The first two equations imply that $x_i = \frac{a_i}{2(1+\mu)}$. Using the last (complementary slackness) equation, we need to check two cases.

On one hand, if $\mu = 0$, we have $x_i = a_i/2$. The point $(a_1/2, a_2/2)$ is feasible, if and only if $a_1^2 + a_2^2 \leq 4$, in which case $f(a_1/2, a_2/2) = -(a_1^2 + a_2^2)/4$.

On the other hand, if $x_1^2 + x_2^2 = 1$, we can combine this equation with the first two equations, and solve. We get $\mu = \sqrt{a_1^2 + a_2^2}/2 - 1$. Since μ is non-negative, this forces $a_1^2 + a_2^2 \geq 4$. Then we have $x_i = \frac{a_i}{\sqrt{a_1^2+a_2^2}}$ and we are on the circle of radius 1 where the constraint is active, as required. The value of f as this point is $1 - \sqrt{a_1^2 + a_2^2}$.

In both cases the value that we have found is less than the value of f at $\mathbf{0}$, so we have the unique minimizer. Finally note that the case where $\mu = 0$ and that where $x_1^2 + x_2^2 = 1$ only overlap when $a_1^2 + a_2^2 = 4$, and that the answer we have found is the same.

We can now let $\mathbf{a}$ vary, and think of the unique minimizer $\mathbf{x}$ as a function of $\mathbf{a}$. We get the answer

$$\mathbf{x}(\mathbf{a}) = \begin{cases} (a_1/2, a_2/2) & \text{if } a_1^2 + a_2^2 \leq 4; \\ \big(\frac{a_1}{\sqrt{a_1^2+a_2^2}}, \frac{a_2}{\sqrt{a_1^2+a_2^2}}\big), & \text{if } a_1^2 + a_2^2 > 4. \end{cases}$$

Notice that $\mathbf{x}(\mathbf{a})$ is a continuous function.

Finally, we computed the 'value function' $v(a_1, a_2)$, which associates to each $\mathbf{a}$ the minimum value of f for that choice of $\mathbf{a}$. We obtained:

$$\mathbf{v}(a_1, a_2) = \begin{cases} -(a_1^2 + a_2^2)/4, & \text{if } a_1^2 + a_2^2 \leq 4; \\ 1 - \sqrt{a_1^2 + a_2^2}, & \text{if } a_1^2 + a_2^2 > 4. \end{cases} \tag{31.4.4}$$

31.4.5 Example. A small modification of Example 31.4.3 provides the computation of the conjugate of $f(\mathbf{x}) = x_1^2 + x_2^2$ restricted to the closed unit disk D, promised in Example 20.3.8. The conjugate of this f is, by definition:

$$f^*(\mathbf{x}) = \max_{\mathbf{y} \in D} \big(\langle \mathbf{y}, \mathbf{x} \rangle - y_1^2 - y_2^2\big)$$

So we are maximizing the function $h(\mathbf{y}) = \langle \mathbf{y}, \mathbf{x} \rangle - y_1^2 - y_2^2$, where $\mathbf{x}$ is a parameter, on the unit disk centered at the origin. This is the same as minimizing $-h = y_1^2 + y_2^2 - \langle \mathbf{y}, \mathbf{x} \rangle$, which we just did in Example 31.4.3: just replace the names a by x, and x by y.

Translate (31.4.4) into this notation, and multiply by -1. We get the conjugate function

$$\mathbf{f}^*(x_1, x_2) = \begin{cases} (x_1^2 + x_2^2)/4, & \text{if } x_1^2 + x_2^2 \leq 4; \\ \sqrt{x_1^2 + x_2^2} - 1, & \text{if } x_1^2 + x_2^2 > 4. \end{cases}$$

From Example 22.3.11 we know that this is a convex function, the conjugate of the function $x_1^2 + x_2^2$, restricted to $x_1^2 + x_2^2 \leq 1$. Indeed, it is easy to see this in polar coordinates: write $x_1 = r\cos\theta$, $x_2 = r\sin\theta$. Then the conjugate function only depends of r:

$$\mathbf{f}^*(r, \theta) = \begin{cases} r^2/4, & \text{if } r \leq 2; \\ r - 1, & \text{if } r > 2. \end{cases}$$

so unsurprisingly we get a result similar to Example 20.3.7.

Lecture 32

The Second Order KKT conditions

We continue the study of the KKT conditions by looking at the degree 2 necessary and sufficient conditions. For the time being we only look at the case where the constraints satisfy regularity, the most restrictive of the constraint qualification conditions. We work out some more examples. In particular in §32.3 we show how the arithmetic-geometry mean inequality can be proved by solving a KKT problem.

32.1 The Second Order Necessary Condition

We continue working with the standard problem 31.1.1, and form the Lagrangian 31.1.4. We now suppose all the functions are $\mathcal{C}^2$.

32.1.1 Theorem. *Assume that* $\mathbf{x}^*$ *is regular for the constraints that are active at* $\mathbf{x}^*$. *If the feasible point* $\mathbf{x}^*$ *is a local minimum for our problem, then by the first-order theorem 31.3.1, there exist* λ, μ *satisfying the Lagrange equations* (31.1.7) *, where the* μ *are non-negative and complementary slackness holds. Then the Hessian* L *of the Lagrangian evaluated at* $\mathbf{x}^*$ *and at the* λ *and* μ*:*

$$L = F + \lambda^T H + \mu^T G \tag{32.1.2}$$

is positive semidefinite on the tangent subspace of the active constraints at $\mathbf{x}^*$.

This follows immediately from Theorem 29.3.2 of Lecture 29. Indeed, that theorem tells us that (32.1.2) is a necessary condition for the minimization problem where all the active inequalities are replaced by equalities. *A fortiori*, it is necessary for the problem we are considering, since the feasible set is larger.

Let us check what happens in Example 31.4.2. The Hessian of f is constant and equal to

$$\begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}$$

This is positive definite on the whole space, which explains why the level sets are ellipses, as the graph shows. Thus it is obviously positive definite when restricted to the tangent space of the circle at $(1, 2)$, namely the line $x + 2y = 0$, so we have a minimum.

32.1.3 Example. We continue Example 31.4.1, putting a parameter $r > 0$ for the radius into the constraint equation, so we can vary the equation of the circle: $g = x^2 - 2rx + y^2$. The gradient of $f(x, y) = x - y^2$ is $\nabla f = (1, -2y)$ and the gradient of g is $(2x-2r, 2y)$. Since the Lagrangian is $\mathcal{L} = x-y^2+\mu(x^2-2rx+y^2)$, the KKT equations are

$$\begin{aligned} 1 + 2\mu(x - r) &= 0, \\ 2y(\mu - 1) &= 0. \end{aligned}$$

and complementary slackness must be satisfied: $\mu(x^2 - 2rx + y^2) = 0$. We want to solve this system for $\mu \geq 0$.

The second KKT equation gives two possibilities: either $\mu = 1$ or $y = 0$.

1. Assume $y = 0$. Then $\mu x(x - 2r)$, so three possibilities: $\mu = 0$ or $x = 0$ or $x = 2r$. We treat them each in turn.

   (a) $\mu = 0$. The first KKT equation makes this impossible.

   (b) $x = 0$. Then $1 - 2\mu r = 0$, so $\mu = 1/(2r) > 0$, so this is a possible solution.

   (c) $x = 2r$. This leads to a negative value of μ: this give the maximum, as we already noted.

   So this case leads to one solution $x = 0$, $y = 0$, $\mu = 1/(2r)$.

2. The case $\mu = 1$, then gives $x = r - 1/2$ and $y^2 = r^2 - 1/4$. For there to be a solution, we must have $r \geq 1/2$.

So when $r < 1/2$, we have found only one solution $x = 0$, $y = 0$, $\mu = 1/(2r)$. When $r \geq 1/2$, there are two solutions: the one already found and the pair of solutions $x = r - 1/2$, $y = \pm\sqrt{r^2 - 1/4}$. $\mu = 1$.

The Hessian of f and g are

$$F = \begin{bmatrix} 0 & 0 \\ 0 & -2 \end{bmatrix} \text{ and } G = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

The Hessian of the Lagrangian at the first solution $x = 0$, $y = 0$, $\mu = 1/(2r)$ is therefore

$$L_1 = \begin{bmatrix} 1/r & 0 \\ 0 & 1/r - 2 \end{bmatrix}$$

The tangent line to the constraint locus is the y-axis, so the restriction of the Hessian is just the number $1/r - 2$. This is positive semidefinite when this number is non-negative, so $r \le 1/2$. It is positive definite when $r < 1/2$.

The Hessian of the Lagrangian at the second solution

$$x = r - 1/2, \quad y = \pm\sqrt{r^2 - 1/4}, \quad \mu = 1,$$

which only exists when $r \ge 1/2$, is

$$L_2 = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

which is positive semidefinite. Furthermore it is positive definite when restricted to any line except a vertical line. So in our case, the restriction will always be positive definite, and the sufficient condition of the next section will be satisfied.

Finally, we can compute the value function $v(r)$ in terms of r, meaning, as usual, that for each r we compute the minimum value of the objective function:

$$v(r) = \begin{cases} 0, & \text{if } r \le 1/2; \\ -r^2 + r - 1/4, & \text{if } r > 1/2; \end{cases}$$

Let's see if this makes sense. As r increases, the constraint set gets larger, so the minimum value must decrease (weakly). Since the maximum value of the quadratic $-r^2 + r - 1/4$ is attained at $r = 1/2$, this is indeed the case. Finally, notice that the value function is concave - in particular it is continuous. As we will see in Lecture 23, this follows from Theorem 23.6.3, since the feasible set in convex and the objective function f is concave.

32.2 The Second Order Sufficient Condition

Here the theorem is almost parallel toTheorem 29.4.1 for equality constrained optimization, but not quite:

32.2.1 Theorem. *Now suppose all the functions are $\mathcal{C}^2$. Assume that $\mathbf{x}^*$ is regular for the constraints that are active at $\mathbf{x}^*$, and that there are λ, μ satisfying the first order theorem at $\mathbf{x}^*$, so that the μ are non-negative and complementary slackness holds. Then if the Hessians evaluated at $\mathbf{x}^*$:*

$$L = F + \lambda^T H + \mu^T G$$

is positive definite on the tangent subspace of the active constraints at $\mathbf{x}^*$ *for which the multiplier* μ *is strictly positive, then the function has a strict minimum at* $\mathbf{x}^*$.

This theorem does *not* follow immediately from the corresponding Theorem 29.4.1 of Lecture 29. The proof of this theorem is not yet given in these notes.

This raises an interesting point. There can be active constraints at a point $\mathbf{x}^*$ for which the corresponding multipliers are 0. Such a constraint is called *degenerate*.

A simple modification of Example 31.4.2 gives a degenerate constraint:

32.2.2 Example. Minimize $2x^2 + 2xy + y^2 - 10x - 10y$ subject to

$$x^2 + y^2 - 5 \le 0 \text{ and } 2x + y - 4 \le 0$$

The objective function and the first constraint have not changed. The equation of the constraint line has been changed so that it goes through the minimizing point $(1, 2)$. Clearly the critical point found does not change. You should compute the Lagrange multipliers at $(1, 2)$ to show that the multiplier corresponding to the linear constraint is 0, even though it is active.

32.3 Application: The Arithmetic-Geometry Mean Inequality

In §22.6, we proved the arithmetic-geometric mean inequality, and then deduced many famous inequalities from it. Here we prove it using the optimization techniques we just learned.

Here is what the inequality says, in slightly different notation that suits our purpose now. Start with a collection of n positive real numbers x_i, $1 \le i \le n$. The *arithmetic mean* of the $\mathbf{x}$ is

$$A(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and their *geometric mean* is

$$G(\mathbf{x}) = \Bigl(\prod_{i=1}^{n} x_i\Bigr)^{\frac{1}{n}}$$

The *arithmetic-geometry mean inequality* says that for all such $\mathbf{x}$:

$$G(\mathbf{x}) \le A(\mathbf{x}) \tag{32.3.1}$$

Here we give another proof of this result by solving a KKT problem.

32.3.2 Theorem. *Consider the minimization problem*

Minimize $f(\mathbf{y}) = \sum_{i=1}^n y_i$ subject to the equality constraint $\prod_{i=1}^n y_i = 1$, and inequality constraints $y_i \geq 0$, for $1 \leq i \leq n$.

This problem has a unique minimizer $y_i = 1$, for all i, so that the minimum value $f(\mathbf{y}) = n$.

Proof. None of the positivity constraints can be active, since that would cause the equality constraint $\prod_{i=1}^n y_i = 1$ to fail. Thus by complementary slackness we can assume that all the μ are zero, so the Lagrangian is

$$\sum_{i=1}^n y_i + \lambda(\prod_{i=1}^n y_i - 1)$$

and so the i-th Lagrange equation, multiplied by y_i is

$$y_i + \lambda(\prod_{i=1}^n y_i) = y_i + \lambda = 0.$$

To get the last equation we substituted in the equality constraint. So $y_i + \lambda = 0$. Putting this into the positivity constraint, we get $(-\lambda)^n = 1$. The only value that satisfies the positivity constraints is then $\lambda = -1$ and all $y_i = 1$. □

We use this optimization result to establish the arithmetic-geometry mean inequality (32.3.1). Let $p = \prod_{i=1}^n x_i$, where the x_i are all positive. Let $q = p^{\frac{1}{n}}$ and

$$y_i = \frac{x_i}{q}. \tag{32.3.3}$$

Then the y_i are positive and $\prod_{i=1}^n y_i = 1$, so they satisfy the constraints of Theorem 32.3.2. The unique minimizer is $\mathbf{y}^*$, with $y_i^* = 1$, for all i. Since $f(\mathbf{y}^*) = n$, for any $\mathbf{y}$ satisfying the constraint, we have $n \leq \sum_{i=1}^n y_i$. Use (32.3.3) to eliminate the y_i in favor of the x_i:

$$n \leq \sum_{i=1}^n \frac{x_i}{q}.$$

Multiply by the denominator q and divide by n:

$$q \leq \frac{\sum_{i=1}^n x_i}{n}.$$

This is the same as (32.3.1) so the arithmetic-geometric mean inequality is proved.

Part IX

Advanced Topics

Lecture 33

Convex Optimization without Differentiability

In this lecture we consider a convex optimization problem where the objective function f and the inequality constraints g_i are convex but not necessarily differentiable. The equality constraints h_k are affine, as before.

33.1 Introduction

Here is an important example of duality in optimization.

Let $k(x, y)\colon\ : A \times B \to \overline{\mathbb{R}}$ be a function from any sets A and B to the real numbers extended by $-\infty$ and ∞. So $x \in A$ and $y \in B$.

So we have the functions $f_b(x) = k(x, b)$, for each $b \in B$, and $g_a(y) = k(a, y)$, for each $a \in A$, as above. We now define

$$f(x) = \sup_{b \in B} f_b(x).$$

In other words, for each value x, we consider all the values $f_b(x)$, $\forall b \in B$. If this collection is bounded, we let $f(x)$ be the least upper bound (see Definition 14.2.3); if the collection is unbounded, we let $f(x)$ be ∞. Also see Remark 14.2.4.

We also define

$$g(y) = \inf_{a \in A} g_a(y).$$

For each value y, we consider all the values $g_a(y)$, $\forall a \in A$. If this collection is bounded, we let $g(y)$ be the greatest lower bound (see Definition 14.2.3); if the collection is unbounded, we let $g(y)$ be $-\infty$.

Then consider the two optimization problems: minimize $f(x)$ over A, and maximize $g(y)$ over B. We say these optimization problems are dual.

The key remark is that

$$g(y) \leq k(x,y) \leq f(x), \forall x \in A, y \in B. \tag{33.1.1}$$

This follows from elementary logic: let's just treat the right hand side $k(x,y) \leq f(x)$. The value x is held fixed. Write out the definition of $f(x)$ to get: $k(x,y) \leq \sup_{b\in B} k(x,b)$. Since y is one of the possible values for b, this is clear.

Then, in complete generality, we get a result known as the Weak Minimax Theorem:

$$\sup_{y\in B} \inf_{x\in A} k(x,y) := \sup_{y\in B} g(y) \leq \inf_{x\in A} f(x) := \inf_{x\in A} \sup_{y\in B} k(x,y) \tag{33.1.2}$$

Now assume that equality holds in (33.1.2): the common value is called the *saddle value* k^*. Furthermore, assume that there is a point (x^*, y^*) where the saddle value is achieved: $k(x^*, y^*) = k^*$. Such a point is called a *saddle point*. As we will see later in these lectures, even if there is a saddle value, there need not be a saddle point.

If (x^*, y^*) is a saddle point, then

$$k(x^*, y) \leq k(x^*, y^*) \leq k(x, y^*), \quad \forall x \in A, y \in B. \tag{33.1.3}$$

Conversely, if this equation is satisfied, then (x^*, y^*) is a saddle point.

33.1.4 Theorem. *A point (x^*, y^*) is a saddle point for $k(x,y)$ if and only if x^* is a minimizer for $f(x)$ over A and y^* is a maximizer for $g(y)$ over B and the saddle value k^* exists.*

Since (33.1.3) can be written $f(x^*) = k(x^*, y^*) = g(y^*)$, this is clear.

Thus the key issue is to find saddle points. They do not exist in full generality, but in some important contexts they do. For example, see Theorem 26.5.6 in the context of linear optimization and the more general Theorem 33.4.2 in the context of convex optimization.

33.1.5 Example. The prototypical example is the function $k(x,y) = x^2 - y^2$, for $x \in \mathbb{R}$ and $y \in \mathbb{R}$. Then $f_b(x) = x^2 - b^2$, so that the supremum over b is $f(x) = x^2$. In the same way, $g_a(y) = a^2 - y^2$, so that the infimum over a is $g(x) = -y^2$. Thus the saddle value is 0, and there is a unique saddle point $(0,0)$.

Saddle points will arise in the following way. We will start with, say, a minimization problem concerning a function $f(x)\colon A \to \mathbb{R}$. Then we will construct a space B and a function $k(x, y)\colon A \times B \to \mathbb{R}$ so that the original function $f(x)$ is $\sup_{b \in B} k(x, b)$. Then if we can find a saddle value and a saddle point for $k(x, y)$, we not only get a solution for our original minimization problem, but we also get a solution for the dual problem of maximizing the dual function $g(y)$.

This works best if the function $f(x)$ is convex, as we will see in Lecture 33. There we discuss the existence of saddle points in the context of the Lagrangian of f: compare Definition 33.2.2 to (33.1.3).

33.1.6 Example. The following modification of Example 33.1.5 shows that (33.1.1) can be satisfied, without there being a saddle point. Let $k(x, y) = -x^2 + y^2$. Then $f_b(x) = -x^2 + b^2$ so that the supremum over b is the function identically equal to ∞. Similarly $g_a(y) = -a^2 + y^2$, so that the infimum over a is $g(x) = -\infty$.

This example also shows why we need to consider functions to the extended real numbers $\overline{\mathbb{R}}$, rather than ordinary real-valued functions.

33.1.7 Exercise. Graph the function $k(x, y)$ of Example 33.1.5 on the unit disk in $\mathbb{R}^2$. What are the functions $f_b(x) = x^2 - b^2$ on this graph? What are the functions $g_a(y)$? Now graph the level curves $k(x, y) = c$ for various constants $c \in \mathbb{R}$. Next confirm Theorem 33.1.4.

Finally do the same thing for Example 33.1.6.

33.2 Saddle Points

Before defining saddle points, consider the following problem:

$$\text{Minimize } f(\mathbf{x}) \text{ subject to } \mathbf{h}(\mathbf{x}) = \mathbf{0} \text{ and } \mathbf{g}(\mathbf{x}) \preceq \mathbf{0}. \tag{33.2.1}$$

where we make no convexity or even continuity hypotheses of the objective function $f(\mathbf{x})$ and the constraints $\mathbf{h}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$.

Write the usual Lagrangian

$$\mathcal{L}(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i h_k(\mathbf{x}) + \sum_{k=1}^{p} \mu_k g_k(\mathbf{x}$$

for this problem, where $\mu \succeq \mathbf{0}$.

33.2.2 Definition. A *saddle point* for $\mathcal{L}(\mathbf{x}, \mu)$ is a point $(\mathbf{x}^*, \lambda^*, \mu^*)$ in $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p_{\succeq \mathbf{0}}$ such that for all $(\mathbf{x}, \lambda, \mu)$ in the same space,

$$\mathcal{L}(\mathbf{x}^*, \lambda, \mu) \le \mathcal{L}(\mathbf{x}^*, \lambda^*, \mu^*) \le \mathcal{L}(\mathbf{x}, \lambda^*, \mu^*) \tag{33.2.3}$$

33.2.4 Theorem. *If* $(\mathbf{x}^*, \lambda^*, \mu^*)$ *is a saddle point for* $\mathcal{L}(\mathbf{x}, \lambda, \mu)$*, then* $\mathbf{x}^*$ *is a solution of Problem 33.2.1, and* $f(\mathbf{x}^*) = \mathcal{L}(\mathbf{x}^*, \lambda^*, \mu^*)$*. Conversely, if* $(\mathbf{x}^*, \lambda^*, \mu^*)$ *is a triple in* $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p_{\succeq \mathbf{0}}$ *such that the right-hand inequality of* (33.2.3) *holds,* $\mathbf{x}^*$ *is feasible for the constraints of Problem* (33.2.1)*, and complementary slackness holds, then* $(\mathbf{x}^*, \lambda^*, \mu^*)$ *is a saddle point.*

Proof. By subtraction, the left side inequality in (33.2.3) implies

$$\langle \lambda - \lambda^*, \mathbf{h}(\mathbf{x}^*) \rangle + \langle \mu - \mu^*, \mathbf{g}(\mathbf{x}^*) \rangle \leq \mathbf{0}. \tag{33.2.5}$$

Assume $\mathbf{h}(\mathbf{x}^*) \neq \mathbf{0}$, so that a coordinate, say the i-th coordinate h_i of $\mathbf{h}$ is nonzero at $\mathbf{x}^*$. Then by choosing λ_i either very large or very small, depending on the sign of $h_i(\mathbf{x}^*)$, we can contradict the saddle point inequality (33.2.5). In a similar way, if $g_k(\mathbf{x}^*) > 0$, by making μ_k very large we can contradict (33.2.5). Thus $h_i(\mathbf{x}^*) = 0$ for all i and $g_k(\mathbf{x}^*) \leq 0$ for all k, which shows that $\mathbf{x}^*$ is feasible for Problem (33.2.1).

Now choose $\mu = \mathbf{0}$ in (33.2.5). We get $\langle \mu^*, \mathbf{g}(\mathbf{x}^*) \rangle \geq 0$. Since we have just established that $\langle \mu^*, \mathbf{g}(\mathbf{x}^*) \rangle \leq 0$, we have $\langle \mu^*, \mathbf{g}(\mathbf{x}^*) \rangle = 0$, complementary slackness. This shows that $f(\mathbf{x}^*) = \mathcal{L}(\mathbf{x}^*, \lambda^*, \mu^*)$, the last claim of the theorem.

We now use the right-hand inequality and complementary slackness to get

$$f(\mathbf{x}^*) \leq f(\mathbf{x}) + \langle \lambda^*, \mathbf{h}(\mathbf{x}) \rangle + \langle \mu^*, \mathbf{g}(\mathbf{x}) \rangle.$$

Now assume that $\mathbf{x}$ is feasible. Then $\mathbf{h}(\mathbf{x}) = 0$, so

$$f(\mathbf{x}^*) \leq f(\mathbf{x}) + \langle \mu^*, \mathbf{g}(\mathbf{x}) \rangle,$$

and since $\mathbf{g}(\mathbf{x}) \preceq \mathbf{0}$ and $\mu \succeq \mathbf{0}$, we get $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all feasible $\mathbf{x}$, as required. □

33.2.6 Remark. The general definition of a saddle point is given in §33.1. We already encountered a saddle-point result when we looked at matrix games in §26.5. We started with a $m \times n$ matrix A, and we looked at the values $\mathbf{q}^T A\mathbf{p}$ where $\mathbf{p}$ and $\mathbf{q}$ are probability vectors of the length n and m respectively. . We called a pair of probability vectors $\mathbf{p}^0$ and $\mathbf{q}^0$ optimal, if, for all $\mathbf{p}$ and $\mathbf{q}$,

$$(\mathbf{q}^0)^T A\mathbf{p} \leq (\mathbf{q}^0)^T A\mathbf{p}^0 \leq \mathbf{q}^T A\mathbf{p}^0$$

and we proved that an optimal pair exists for each matrix A: Theorem 26.5.6. Thus $\mathbf{q}^0$ is a probability vector that minimizes $\mathbf{q}^T A\mathbf{p}^0$ and $\mathbf{p}^0$ a probability vector that maximizes $(\mathbf{q}^0)^T A\mathbf{p}$.

33.3 Implications of the Existence of Saddle Points

33.3.1 Theorem. *If $f(\mathbf{x})$ is $\mathcal{C}^1$ at $\mathbf{x}^*$, and if $(\mathbf{x}^*, \lambda^*, \mu^*)$ is a saddle point for the Lagrangian $\mathcal{L}(\mathbf{x}, \lambda, \mu)$ of f, then f satisfies the KKY conditions at $(\mathbf{x}^*, \lambda^*, \mu^*)$.*

Proof. The right-hand inequality of (33.2.3) says that $\mathbf{x}^*$ is an unconstrained minimizer of the function $\mathcal{L}(\mathbf{x}, \lambda^*, \mu^*)$, so its gradient vanishes there. Since we already have positivity of μ^* by hypothesis, and since we have established complementary slackness even without the differentiability hypothesis, we have all the KKT conditions. □

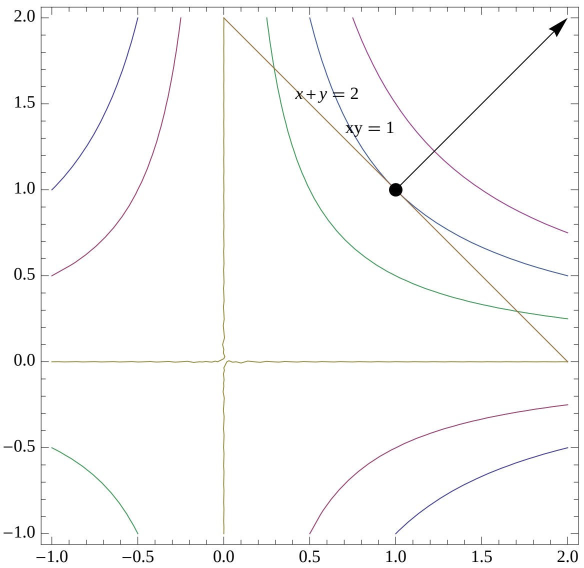
33.3.2 Theorem. *If $f(\mathbf{x})$ is $\mathcal{C}^2$ at $\mathbf{x}^*$, and if $(\mathbf{x}^*, \lambda^*, \mu^*)$ is a saddle point for the Lagrangian $\mathcal{L}(\mathbf{x}, \lambda, \mu)$ of f, then the Hessian of $\mathcal{L}(\mathbf{x}, \lambda^*, \mu^*)$ is positive semidefinite.*

Proof. Again, the right-hand inequality of (33.2.3) says that $\mathbf{x}^*$ is an unconstrained minimizer of the function $\mathcal{L}(\mathbf{x}, \lambda^*, \mu^*)$, so the Hessian of the Lagrangian is positive semidefinite by Theorem 13.1.2. □

33.3.3 Example. We minimize $f(x, y) = xy$ with the single constraint $x + y = 2$. The Lagrange equations are $y + \lambda = 0$ and $x + \lambda = 0$, so we get the triple $(x^*, y^*, \lambda^*) = (1, 1, -1)$ as the unique solution to the first order equations. The Hessian of the Lagrangian at this point is

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

which is not positive definite: the eigenvalues are ± 1 and the eigenvectors $(1, 1)$ and $(1, -1)$. Thus the constraint restricts us to the line spanned by the $+1$ eigenvector, so the Hessian is positive definite there, and we have a strict minimum. There are many other ways we could have seen this of course, most easily by graphing the level curves , the constraint line, and the gradient of the objective function at the minimizer.



So this is as simple and well-behaved minimization problem as can be imagined, and yet the point $(x^*, y^*, \lambda^*) = (1, 1, -1)$ is not a saddle point, because it fails the right-hand inequality in the saddle point definition:

$$1 \leq xy - (x + y - 2)$$

for all x and y. Indeed, just let $y = -x$, and let x get very large.

If f is a convex function, this kind of example cannot occur, because the Hessian of a convex function is positive semidefinite. Thus the notion of saddle point is only useful when one is dealing with convex optimization, of which linear optimization is a special case.

33.4 The Basic Theorem

We keep the same set-up as in Lecture 23, in particular the standard problem is the same, except that the differentiability requirements are weakened: all we know about f and the g_k is that they are convex, and therefore (by Theorem 21.4.3) continuous. Here it is:

33.4.1 Definition. Minimize $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, subject to

$$A\mathbf{x} = \mathbf{b} \text{ and } g_k(\mathbf{x}) \leq 0\,,\, 1 \leq k \leq p\,,$$

where $f(\mathbf{x})$ and the $g_k(\mathbf{x})$ are convex functions, and A is a $m \times n$ matrix of maximal rank m.

As usual we write the Lagrangian for this problem:

$$\mathcal{L}(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i (A\mathbf{x} - \mathbf{b}) + \sum_{k=1}^{p} \mu_k g_k(\mathbf{x})$$

We need the Slater condition (see Definition 23.7.1). Note the strict inequality in the inequality constraints of the Slater condition, so that we are postulating the existence of a special kind of feasible vector.

33.4.2 Theorem. *Assume that Problem 33.4.1 satisfies the Slater condition. Then a necessary condition for a feasible point $\mathbf{x}^*$ to be a minimizer for Problem 33.4.1 is that there exist a $\mu^* \succeq 0 \in \mathbb{R}^p$ such that $\langle \mu^*, \mathbf{g}(\mathbf{x}^*) \rangle = 0$ and a $\lambda^* \in \mathbb{R}^m$ such that $(\mathbf{x}^*, \lambda^*, \mu^*)$ is a saddle point for the Lagrangian of f.*

Proof. The proof of Theorem 23.7.2 shows that there are λ^* and $\mu^* \geq 0$ such that complementary slackness holds: $\langle \mu^*, \mathbf{g}(\mathbf{x}^*) \rangle = 0$, and such that $\mathcal{L}(\mathbf{x}, \lambda^*, \mu^*)$ is minimized at $\mathbf{x}^*$. Then we just apply Theorem 33.2.4: the converse statement. □

Lecture 34

Nonlinear Duality

For convex minimization problem (so f and the g_i are convex, and the h_k are affine), there is a theory of duality that is a beautiful generalization of the duality we have already studied in the linear case. These notes define the new notion and show how it generalizes linear duality. We also show what happens in the quadratic case.

34.1 The Dual Function

We start with the same minimization problem as in Lecture 31: Minimize $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, subject to

$$h_i(\mathbf{x}) = 0 \;, 1 \le i \le m \text{ and } g_k(\mathbf{x}) \le 0 \;, 1 \le k \le p \tag{34.1.1}$$

Let F be the feasible set for this problem, and let f^* be the minimum value of $f(\mathbf{x})$ on F. We allow the value $-\infty$ if f is unbounded below on F. To this problem we associate the Lagrangian:

$$\begin{aligned} \mathcal{L}(\mathbf{x}, \lambda, \mu) &= f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i h_i + \sum_{k=1}^{p} \mu_k g_k \qquad (34.1.2) \\ &= f(\mathbf{x}) + \lambda^T \mathbf{h}(\mathbf{x}) + \mu^T \mathbf{g}(\mathbf{x}) \end{aligned}$$

defined for all λ, and for all $\mu \succeq \mathbf{0}$.

34.1.3 Definition. The *dual function* $\phi(\lambda, \mu)$ to $f(\mathbf{x})$, with its constraints $\mathbf{g}(\mathbf{x})$ and $\mathbf{h}(\mathbf{x})$, is the function

$$\phi(\lambda, \mu) = \inf_{\mathbf{x} \in F} \mathcal{L}(\mathbf{x}, \lambda, \mu)$$

We write '$\inf$' rather that '$\min$' because the value might be $-\infty$ at some points.

34.1.4 Remark. Because $\phi(\lambda, \mu)$ is the $\inf$ of a collection of linear and therefore concave functions, by the results of §22.3, ϕ is a concave function. Indeed, in Example 22.3.11, we showed that the $\sup$ of an arbitrary collection of convex functions is convex, so here we just multiply by -1 to get the result.

34.1.5 Theorem. *For all λ, and for all $\mu \geq 0$,*

$$\phi(\lambda, \mu) \leq f^*, \tag{34.1.6}$$

where f^* is the minimum value of f on F. If there are Lagrange multipliers λ^*, and $\mu^* \succeq \mathbf{0}$, associated to the minimizer $\mathbf{x}^*$, so that complementary slackness is satisfied:

$$\langle \mu^*, \mathbf{g}(\mathbf{x}^*) \rangle = 0$$

then

$$\phi(\lambda^*, \mu^*) = f^*.$$

This occurs in the convex case when the Slater condition is satisfied: see Theorem 23.7.2.

Proof. At a feasible point $\mathbf{x}$, for any λ, and for any $\mu \succeq \mathbf{0}$, we have

$$\lambda^T \mathbf{h}(\mathbf{x}) + \mu^T \mathbf{g}(\mathbf{x}) = \mu^T \mathbf{g}(\mathbf{x}) \leq 0,$$

since $\mu \succeq 0$ and $\mathbf{g}(\mathbf{x}) \preceq \mathbf{0}$. So

$$\mathcal{L}(\mathbf{x}, \lambda, \mu) \leq f(\mathbf{x}) \tag{34.1.7}$$

Now we take the infimum with respect to $\mathbf{x}$ of both sides, getting $\phi(\lambda, \mu) \leq f^*$. This establishes the first claim. If complementary slackness is achieved at μ^* and $\mathbf{x}^*$, then (34.1.7) becomes an equality and the second part of the theorem is proved. □

This shows that once the dual function $\phi(\lambda, \mu)$ is defined, we should maximize it in order to approximate as closely as possible the minimum value of the original (primal) function f. Thus dually, the problem of minimizing f becomes that of maximizing $\phi(\lambda, \mu)$.

34.2 Example: The Linear Case

34.2.1 Example. We first compute the dual function for the problem:

$$\text{Minimize } \mathbf{c}^T\mathbf{x} \text{ subject to } A\mathbf{x} = \mathbf{b} \text{ and } \mathbf{x} \geq \mathbf{0}$$

The Lagrangian is

$$\mathcal{L}(\mathbf{x}, \lambda, \mu) = \mathbf{c}^T\mathbf{x} + \lambda^T(\mathbf{b} - A\mathbf{x}) - \mu^T\mathbf{x},$$

which we can rewrite, grouping the terms involving $\mathbf{x}$:

$$\mathcal{L}(\mathbf{x}, \lambda, \mu) = \lambda^T\mathbf{b} + (\mathbf{c} - A^T\lambda - \mu)^T\mathbf{x},$$

so the dual function can be written:

$$\phi(\lambda, \mu) = \lambda^T\mathbf{b} + \inf_{\mathbf{x}}(\mathbf{c} - A^T\lambda - \mu)^T\mathbf{x}.$$

As soon as a coefficient of $\mathbf{x}$ is non-zero, a suitable value of $\mathbf{x}$ can drive the dual function to $-\infty$, so we get the result:

34.2.2 Theorem. *With the constraints $\mathbf{c} - A^T\lambda \succeq \mathbf{0}$, $\mu = \mathbf{c} - A^T\lambda$, the dual function $\phi(\lambda, \mu) = \lambda^T\mathbf{b}$. When these constraints are not satisfied, $\phi(\lambda, \mu) = -\infty$.*

Thus the dual problem to the standard linear optimization problem, asymmetric form (25.1.5), is maximize $\lambda^T b$ with the constraints $\mathbf{c} - A^T\lambda \geq \mathbf{0}$. Note that the dual variables here are the λ and the μ. The μ do not play an important role, so we are left with the λ. These variables were called $\mathbf{y}$ in Lecture 25.

As noted in the previous section, to understand the primal function we want to maximize the dual, so we want to

$$\text{Maximize } \lambda^T\mathbf{b} \text{ subject to } \lambda^T A \leq \mathbf{c} \tag{34.2.3}$$

Compare to Example 25.4.1.

34.2.4 Example. Next we compute the dual function for the symmetric linear optimization problem:

$$\text{Minimize } \mathbf{c}^T\mathbf{x} \text{ subject to } A\mathbf{x} \geq \mathbf{b} \text{ and } \mathbf{x} \geq \mathbf{0}.$$

There are no equality constraints, and two sets of inequality constraints. We use the dual variables μ and ν for them, so they must both be non-negative. The Lagrangian is

$$\mathcal{L}(\mathbf{x}, \mu, \nu) = \mathbf{c}^T\mathbf{x} + \nu^T(\mathbf{b} - A\mathbf{x}) - \mu^T\mathbf{x}$$

which we can rewrite, grouping the terms involving $\mathbf{x}$:

$$\mathcal{L}(\mathbf{x},\mu,\nu) = \nu^T\mathbf{b} + (\mathbf{c} - A^T\nu - \mu)^T\mathbf{x}$$

so the dual function

$$\phi(\mu,\nu) = \nu^T\mathbf{b} + \inf_{\mathbf{x}}(\mathbf{c} - A^T\nu - \mu)^T\mathbf{x}$$

As before, as soon as a coefficient of $\mathbf{x}$ is different from 0, a suitable value of $\mathbf{x}$ can drive the dual function to $-\infty$, so we get the result:

34.2.5 Theorem. *With the constraints* $\mathbf{c} - A^T\nu \geq \mathbf{0}$*,* $\mu = \mathbf{c} - A^T\nu$*, the dual function* $\phi(\mu,\nu) = \nu^T\mathbf{b}$*. When these constraints are not satisfied,* $\phi(\mu,\nu) = -\infty$*.*

In this case the dual variable ν must be nonnegative. Again, we have recovered the same dual function found in Lecture 25:

$$\text{Maximize } \nu^T\mathbf{b} \text{ subject to } \nu^T A \leq \mathbf{c} \text{ and } \nu \geq 0.$$

See 25.3.11.

34.3 Example: The Quadratic Case

34.3.1 Example. We now consider the quadratic problem

$$\text{Minimize } \frac{1}{2}\mathbf{x}^T Q\mathbf{x} + \mathbf{c}^T\mathbf{x} \text{ subject to } A\mathbf{x} \leq \mathbf{b}$$

where we assume that Q is symmetric and positive semidefinite.

The Lagrangian is

$$\mathcal{L}(\mathbf{x},\mu) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mu^T(A\mathbf{x} - \mathbf{b})$$

so the dual function is

$$\phi(\mu) = -\mu^T\mathbf{b} + \inf_{\mathbf{x}}\big(\frac{1}{2}\mathbf{x}^T Q\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mu^T A\mathbf{x}\big)$$

for $\mu \geq 0$.

We minimize $\frac{1}{2}\mathbf{x}^T Q\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mu^T A\mathbf{x}$ for a fixed value of μ. This is an unconstrainted convex minimization problem. A necessary and sufficient condition for it to have a minimum at $\mathbf{x}^*$ is that the gradient with respect to $\mathbf{x}$ vanish there:

$$\mathbf{x}^{*T}Q + \mathbf{c}^T + \mu^T A = 0 \tag{34.3.2}$$

Because we are not assuming that Q is invertible, we cannot solve for $\mathbf{x}^*$. Instead we multiply by $\mathbf{x}^*$ on the right, getting

$$\mathbf{x}^{*T}Q\mathbf{x}^* + \mathbf{c}^T\mathbf{x}^* + \mu^T A\mathbf{x}^* = 0$$

Substituting this into the dual function we get

$$\phi(\mu) = -\mu^T\mathbf{b} - \frac{1}{2}\mathbf{x}^{*T}Q\mathbf{x}^* \tag{34.3.3}$$

where $\mathbf{x}^*$ depends on μ through (34.3.2).

If Q is positive definite, then we can solve for $\mathbf{x}$ in (34.3.2), getting

$$\mathbf{x}^* = -Q^{-1}\big(A^T\mu + \mathbf{c}\big) \tag{34.3.4}$$

We substitute this into (34.3.3), writing $R = Q^{-1}$, to R is also symmetric, to get:

$$\begin{aligned}\phi(\mu) &= -\mu^T\mathbf{b} - \frac{1}{2}\big(\mu^T A + \mathbf{c}^T\big)R^T Q R\big(A^T\mu + \mathbf{c}\big) && (34.3.5)\\ &= -\mu^T\mathbf{b} - \frac{1}{2}\big(\mu^T A + \mathbf{c}^T\big)R\big(A^T\mu + \mathbf{c}\big) && (34.3.6)\\ &= -\mu^T\mathbf{b} - \frac{1}{2}\big(\mu^T AQ^{-1}A^T\mu\big) - \frac{1}{2}\big(\mathbf{c}^TQ^{-1}\mathbf{c}\big) - \mathbf{c}^TQ^{-1}A^T\mu && (34.3.7)\end{aligned}$$

This is a quadratic polynomial in μ, with quadratic part $-AQ^{-1}A^T$ and linear part $-(\mathbf{c}^TQ^{-1}A^T + \mathbf{b})\mu$. If A has rank $m \le n$, then, as we saw in Proposition 30.3.2, $-AQ^{-1}A^T$ is a $m \times m$ negative definite matrix.

If $\mathbf{c}$ is the zero vector, this simplifies to

$$\phi(\mu) = -\mu^T\mathbf{b} - \frac{1}{2}\big(\mu^T AQ^{-1}A^T\mu\big)$$

The key question is: what kind of matrix is $-AQ^{-1}A^T$ in general?

34.3.8 Proposition. *Assume the nullspace of A^T has dimension k. Then the quadratic form $-AQ^{-1}A^T$ is negative semidefinite, with a 0-part of dimension k, and a negative part of dimension $m - k$.*

When we are dealing with inequalities, we could have $m > n$, in which case $-AQ^{-1}A^T$ will have rank at most n.

Here is an example.

34.3.9 Example. Minimize

$$f(\mathbf{x}) = 3x_1^2 + 2x_2^2 - 2x_1x_2 - 3x_1 - 4x_2$$

subject to

$$2x_1 + 3x_2 \leq 6\,,\; -x_1 + 2x_2 \leq 2\,, \mathbf{x} \geq \mathbf{0}\,.$$

so

$$Q = 2\begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix}, \mathbf{c} = \begin{bmatrix} -3 \\ -4 \end{bmatrix}, A = \begin{bmatrix} 2 & 3 \\ -1 & 2 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 6 \\ 2 \\ 0 \\ 0 \end{bmatrix}$$

Notice that Q is positive definite, and that the last two constraints, positivity constraints for the variables, have to be multiplied by -1 so the inequalities go in the right direction.

We want to write $\mathbf{x}^*$ as a function of μ. We need

$$Q^{-1} = \frac{1}{10}\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$$

so, using this in (34.3.4)

$$\mathbf{x}^* = -\frac{1}{10}\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 2\mu_1 - \mu_2 - \mu_3 - 3 \\ 3\mu_1 + 2\mu_2 - \mu_4 - 4 \end{bmatrix} = -\frac{1}{10}\begin{bmatrix} 7\mu_1 - 2\mu_3 - \mu_4 - 10 \\ 11\mu_1 + 5\mu_2 - \mu_3 - 3\mu_4 - 15 \end{bmatrix}$$

Now we plug this into (34.3.3):it is clear we get a quadratic polynomial in the μ. The purely quadratic part is given by $AQ^{-1}A^T$ which we now compute:

$$AQ^{-1}A^T = \frac{1}{10}\begin{bmatrix} 2 & 3 \\ -1 & 2 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 2 & -1 & -1 & 0 \\ 3 & 2 & 0 & -1 \end{bmatrix} = \frac{1}{10}\begin{bmatrix} 47 & 15 & -7 & -11 \\ 15 & 10 & 0 & -5 \\ -7 & 0 & -12 & 1 \\ -11 & -5 & 1 & 3 \end{bmatrix}$$

How to solve the primal problem? As always with KKT, one means to guess which constraints are active. It is worth determining the minimum for the unconstrained problem given by $f(\mathbf{x})$. A little gradient computation shows that it occurs at $(1, \frac{3}{2})$. This is not in the feasible set of the constrainted problem, so it is not our answer.

It is worth graphing the feasible set. You get a polygonal region whose vertices, going counterclockwise, are $(0,0)$,$(3,0)$, $(\frac{6}{7}, \frac{10}{7})$,$(0,1)$.

Since the level sets of f are ellipses centered at $(1, \frac{3}{2})$, it is clear that the only active constraints are going to be μ_1 and perhaps μ_1. So we try with just μ_1, which we call μ.

We solve the two gradient equations with the first constraint active equation: $2x_1 + 3x_2 = 6$. This is a system of three linear equations, with a unique solution with a positive $\mu = \frac{5}{47}$ and feasible (x_1, x_2), so we are done.

34.4 The Duality Theorem for Convex Minimization

This simply reiterates some results mentioned in the first section. Keeping the same notation, we now let ϕ^* stand for $\max_{\lambda,\mu} \phi(\lambda, \mu)$ just as f^* stands for $\min_{\mathbf{x}} f(\mathbf{x})$.

34.4.1 Definition. *Weak duality* is the assertion that

$$\phi^* \leq f^*$$

We saw in the first section that it always holds.

34.4.2 Definition. *Strong duality* is the assertion that

$$\phi^* = f^*$$

in which case the pair (λ^*, μ^*) that achieves the maximum for ϕ are the Lagrange multipliers associated to the minimizer for the primal problem.

34.4.3 Theorem. *Strong duality holds for the convex minimization problem when the Slater condition is satisfied.*

This was already mentioned in Theorem 34.1.5. For more general optimization problems it may fail, in which case the difference $f^* - \phi^*$ is called the *duality gap*.

Part X

Appendices

Appendix A

Symbols and Notational Conventions

A.1 Logic

We will use the universal quantifier $\forall$ *for all* and the existential qualifier $\exists$ *there exists*. See §2.1.1 for details.

If S is a set, and $R \subset S$ a subset of S, then $S \setminus R$ denotes the elements of S that are not in R.

We will occasionally use the 'exclusive' *or*: if P and R are two statements, in ordinary mathematics P or R means that at least one of the two statements is true. P 'exclusive' or R means that exactly one of the two statements is true. See Corollary 7.2.4 for an example.

A.2 Number Systems

$\mathbb{N}$ denotes the natural numbers, namely the positive integers.

$\mathbb{Z}$ denotes the integers.

$\mathbb{Q}$ denotes the rational numbers.

$\mathbb{R}$ denotes the real numbers.

$\overline{\mathbb{R}}$ is $\mathbb{R}$ extended by $-\infty$ and ∞.

$[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$.

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

A.3 Real Vector Spaces

The n-th cartesian product of the real numbers $\mathbb{R}$ is written $\mathbb{R}^n$. Lower-case bold letters such as $\mathbf{x}$ and $\mathbf{a}$ denote vectors in $\mathbb{R}^n$, each with coordinates represented by non-bold letters $(x_1, \ldots, x_n)$ and $(a_1, \ldots, a_n)$, respectively. We typically use $\mathbf{x}$ (and $\mathbf{y}$, $\mathbf{z}$, etc.) for variables and $\mathbf{a}$ (and $\mathbf{b}$, $\mathbf{c}$, etc.) for constants.

Vectors are also called points, depending on the context. When the direction is being emphasized, it is called a vector.

With the exception of gradients, vectors are always column matrices.

In the body of the text, an expression such as $[a_1, a_2, \ldots, a_n]$ denotes a column vector while $(a_1, a_2, \ldots, a_n)$ denotes a row vector.

The length of a vector $\mathbf{v}$ is written $\|\mathbf{v}\|$, and the inner product of $\mathbf{v}$ and $\mathbf{w}$ is $\langle \mathbf{v}, \mathbf{w} \rangle$, or, more rarely, $\mathbf{v} \cdot \mathbf{w}$.

The partial order in $\mathbb{R}^n$ leads to the following notation:

$\mathbf{x} \prec \mathbf{y}$ means that $x_i < y_i$ for all $1 \leq i \leq n$

$\mathbf{x} \preceq \mathbf{y}$ means that $x_i \leq y_i$ for all $1 \leq i \leq n$

$\mathbf{x} \preccurlyeq \mathbf{y}$ means that $x_i \leq y_i$ for all $1 \leq i \leq n$ and $x_j < y_j$ for some j

and therefore

$$\begin{aligned} \mathbb{R}^n_{\succ} &= \{\mathbf{x} \mid \mathbf{x} \succ \mathbf{0}\} \\ \mathbb{R}^n_{\succeq} &= \{\mathbf{x} \mid \mathbf{x} \succeq \mathbf{0}\} \\ \mathbb{R}^n_{\succcurlyeq} &= \{\mathbf{x} \mid \mathbf{x} \succcurlyeq \mathbf{0}\} \end{aligned}$$

The open ball of radius r centered at the point $\mathbf{p} \in \mathbb{R}^n$ is written

$$N_r(\mathbf{p}) = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{p}\| < r\}$$

and the closed ball

$$\overline{N}_r(\mathbf{p}) = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{p}\| \leq r\}$$

The tangent space of $\mathbb{R}^n$ at a point $\mathbf{p}$ is written $T_{\mathbb{R}^n,\mathbf{p}}$ or simply $T_{\mathbf{p}}$. See §17.2. If M is a submanifold of $\mathbb{R}^n$, its tangent space at $\mathbf{p}$ is written $T_{M,\mathbf{p}}$. See §17.3.

A.4 Analysis

A function is $\mathcal{C}^1$ if it is continuously differentiable, and $\mathcal{C}^n$ if it is is n-times differentiable, and its n-th derivative is continuous. This, both for functions of a single variable and of several variables.

We use the Lagrange notation $f'(x)$, $f''(x)$, …, $f^{(n)}$ for the successive derivatives of a function of one variable. On the other hand we use the Leibniz notation $\frac{\partial f}{\partial x}$ for the partial derivative of f with respect to x. On occasion we will use the Cauchy notation $D_x f$ for the same partial to simplify the notation.

For a differentiable real valued function $f\colon \mathbb{R}^n \to \mathbb{R}$, we write $\nabla f(\mathbf{x})$ for its gradient vector evaluated at $\mathbf{x}$.

$$\nabla f(\mathbf{x}) = \big(\frac{\partial f}{\partial x_1}(x_1, \dots, x_n), \dots, \frac{\partial f}{\partial x_n}(x_1, \dots, x_n)\big)$$

and if f is twice differentiable, $F(\mathbf{x})$ for its $n \times n$ Hessian matrix at $\mathbf{x}$:

$$F(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}) \\ \dots & \dots & \dots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}) & \dots & \frac{\partial^2 f}{\partial x_n^2}(\mathbf{x}) \end{bmatrix} \tag{A.4.1}$$

A.5 Linear Algebra

Matrices are written with square brackets as in (A.4.1). Matrices are denoted by capital roman letters such as A, and have as entries the corresponding lower case letter. So $A = [a_{ij}]$. A is an $m \times n$ matrix if it has m rows and n columns, so $1 \le i \le m$ and $1 \le j \le n$. We write the columns of A as $\mathbf{a}_j$ and the rows as $\mathbf{a}^i$.
A^T is the transpose of the matrix A.
$D(d_1, d_2, \dots, d_n)$ is the $n \times n$ diagonal matrix

$$\begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & 0 \\ 0 & 0 & 0 & \dots & d_n \end{bmatrix}$$

I_n or just I stands for the $n \times n$ identity matrix $D(1, 1, \dots, 1)$.
If A is an $m \times n$ matrix, T_A is the linear transformation from $\mathbb{R}^m$ to $\mathbb{R}^n$ given by $T_A(\mathbf{x}) = A\mathbf{x}$, the matrix product of A by the n-column vector $\mathbf{x}$. The nullspace of T_A is written $\mathcal{N}(A)$, and its range $\mathcal{R}(A)$.

A.6 Affine Geometry

The affine hyperplane $\{\mathbf{x} \mid \langle \mathbf{a}, \mathbf{x} \rangle = c\}$ in $\mathbb{R}^n$ is written $H_{\mathbf{a},c}$. It separates $\mathbb{R}^n$ into two closed halfspaces

$$\begin{aligned} H^+_{\mathbf{a},c} &= \{\mathbf{x} \mid \langle \mathbf{a}, \mathbf{x} \rangle \geq c\} \\ H^-_{\mathbf{a},c} &= \{\mathbf{x} \mid \langle \mathbf{a}, \mathbf{x} \rangle \leq c\} \end{aligned}$$

A.7 Convexity

CoS is the convex hull of S. See Definition 18.1.17. KS is the set of convex combinations of S. See Definition 18.1.25.

$P(A, \mathbf{b})$ is the polyhedron $\{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} \leq \mathbf{b}\}$, where A is an $m \times n$ matrix and $\mathbf{b}$ an m-vector. See Definition 18.3.16.

$\hat{P}(A, \mathbf{b})$ satisfies the additional positivity condition $\mathbf{x} \geq \mathbf{0}$. See Definition 19.2.8.

C_A denotes the cone defined in Definition 19.3.1. C^* is the dual cone defined in Theorem 20.4.8.

A.8 Optimization

The objective function is always f, and typically $f : \mathbb{R}^n \to \mathbb{R}$.
To the extent possible, the equality constraints are written $\mathbf{h}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^m$, and the i-th equality constraint refers to the function $h_i(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ whose scalar output is the i-th coordinate of $\mathbf{h}(\mathbf{x})$.
Similarly the inequality constraints are written $\mathbf{g}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^p$, and the j-th inequality constraint refers to the function $g_j(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ whose scalar output is the j-th coordinate of $\mathbf{g}(\mathbf{x})$.
The feasible set is usually written D. It is usually the intersection of the domain of f with the set $\mathbf{h}(\mathbf{x}) = 0$ and the set $\mathbf{g}(\mathbf{x}) \leq 0$.

References

[1] Gabriel Archinard and Bernard Guerrien, *Analyse Mathématique Pour Economistes*: *Cours et exercices corrigés*, quatrième édition, Economica, 49, rue Héricart, 75015 Paris, 1992. ↑365

[2] K. J. Arrow and A. C. Enthoven, *Quasiconcave Programming*, Econometrica **29** (1961), 779-800. ↑365, 368

[3] Angelo Barone-Netto, Gianluca Gorni, and Gaetano Zampieri, *Local Extrema of Analytic Functions*, NoDEA **3** (1996), 287-303. ↑174

[4] Alexander Barvinok, *A Course in Convexity*, Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2002. ↑xiii, 235, 236, 267, 296

[5] Mokhtar S. Bazaraa, Hanif D. Sherali, and C. M. Shetty, *Nonlinear Programming*: *Theory and Algorithms*, Wiley Interscience, Hoboken, NJ, 2006. ↑4, 235, 360

[6] Richard Bellman, *Dynamic Programming*, Princeton University Press, Princeton, 1957. reprint of the sixth (1972) edition by Dover in 2003. ↑xv

[7] Leonard D. Berkovitz, *Convexity and Optimization in* $\mathbf{R}^n$, Wiley, New York, 2002. ↑xiii, 3, 235, 236, 291, 330, 332, 375, 418

[8] R. G. Bland, *New Finite Pivoting Rules for the Simplex Method*, Math. Oper. Res. **2** (1977), no. 2, 205-206. ↑432

[9] Jonathan M. Borwein and Adrien S. Lewis, *Convex Analysis and Nonlinear Optimization*: *Theory and Examples*, Second Edition, Springer, New York, 2006. ↑267

[10] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge U.P, Cambridge, 2004. ↑xiii, 3, 138, 179, 235, 236, 360, 362, 375

[11] Robert E. Bradley and C. Edward Sandifer, *Cauchy's cours d'analyse*: *An Annotated Translation*, Springer, New York, 2009. ↑342

[12] David M. Bressoud, *A Radical Approach to Real Analysis*, Second Edition, MAA, 2007. ↑31, 39, 40, 42, 47, 200

[13] Otto Bretscher, *Linear Algebra*: *With Applications*, Prentice Hall, New York, 1996. ↑71, 73, 91

[14] N. L. Carothers, *Real Analysis*, Cambridge U.P, Cambridge, 2000. ↑345

[15] Augustin-Louis Cauchy, *Cours d'Analyse de l'École Royale Polytechnique*, Debure frères, Paris, 1821. ↑342

[16] ———, *Résumé des Leçons Données à l'École Royale Polytechnique*, Debure frères, Paris, 1823. ↑34, 152

[17] A. Charnes, *Optimality and Degeneracy in Linear Programming*, Econometrica **20** (1952), 160-170. ↑432

[18] Edwin K. P. Chong and Stanislaw H. Zak, *An Introduction to Optimization*, Second Edition, Wiley-Interscience, New York, 2001. ↑4

[19] Gerard Debreu, *Definite and Semidefinite Quadratic Forms*, Econometrica **20** (1952), no. 2, 295-300. ↑464

[20] Zdeněk Dostál, *Optimal Quadratic Programming Algorithms*: *With Applications to Variational Inequalities*, Springer Optimization and its Applications, Springer, New York, 2009. ↑3

[21] Bob A. Dumas and John E. McCarthy, *Transition to Higher Mathematics*, Walter Rudin Student Series in Advanced Mathematics, McGraw Hill, Boston, 2007. ↑xiii

[22] Monique Florenzano and Cuong Le Van, *Finite Dimensional Convexity and Optimization*, Springer, Berlin, 2001. In cooperation with Pascal Gourdel. ↑235, 236, 256, 332, 375

[23] Joel N. Franklin, *Methods of Mathematical Economics*, Classics in Applied Mathematics, SIAM, Philadelphia, 2002. reprint with errata of 1980 Springer -Verlag original. ↑xiii, 3, 291, 375, 418

[24] David Gale, *The Theory of Linear Economic Models*, McGraw-Hill, New York, 1960. ↑8, 375

[25] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1959. translation from the Russian original by K. A. Hirsch. ↑77, 79, 91, 108, 117, 139, 412

[26] Angelo Genocchi, *Calcolo Differenziale*: *Principii di Calcolo Integrale*, Fratelli Bocca, Roma, Torino, Firenze, 1884. Pubblicato con Aggiunte dal Dr. Giuseppe Peano. ↑156, 172, 174, 175

[27] Édouard Goursat, *A Course in Mathematical Analysis*, translated by Earle Raymond Hedrick, Ginn, Boston, 1904. ↑175, 341

[28] E. Hainer and G. Wanner, *Analysis by Its History*, Undergraduate Texts in Mathematics, Springer, New York, 2008. ↑31, 32, 156, 168, 187

[29] J. A. J Hall and K. I. M McKinnon, *The simplest examples where the simplex method cycles and conditions where EXPAND fails to prevent cycling*, Math. Program. **100** (2004), 133-150. ↑437

[30] Harris Hancock, *Theory of Maxima and Minima*, Dover, New York, 1917. ↑175, 447

[31] F. L. Hitchcock, *The Distribution of a Product from Several Sources to Numerous Localities*, J. Math. Physics **20** (1941), 224-230. ↑11

[32] Roger A. Horn and Charles R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1990. ↑135, 137

[33] Paul J. Kelly and Max L. Weiss, *Geometry and Convexity*, John Wiley, New York, 1979. republished by Dover in 2009. ↑236, 241, 248, 256

[34] T. C. Koopmans, *Optimum Utilization of the Transportation System*, Econometrica **17** (1949), 136-145. ↑11

[35] Steven G. Krantz and Harold R. Parks, *The Implicit Function Theorem*, Birkhäuser, Boston, Basel, Berlin, 2002. ↑227

[36] H. W. Kuhn, *On a Pair of Dual Nonlinear Programs*, Nonlinear Programming (J. Abadie, ed.), North-Holland, Amsterdam, 1967, pp. 37–54. ↑341, 342

[37] Joseph L. Lagrange, *Recherches sur la mthode de maximis et minimis*, Miscellanea Taurinensia **1** (1759), 18-42. Oeuvres, I, 3-20, 1867. ↑117

[38] Serge Lang, *Introduction to Linear Algebra*, Second Edition, Springer, New York, 1997. ↑73, 77

[39] Peter D. Lax, *Linear Algebra*, Wiley-Interscience, New York, 1997. ↑8, 73, 91, 267, 409, 411, 413, 418, 446

[40] Steven R. Lay, *Convex Sets and Their Applications*, John Wiley, New York, 1982. reprint by Dover in 2007. ↑235, 236, 244, 249, 296

[41] Mark Levi, *The Mathematical Mechanic*: *Using Physical Reasoning to Solve Problems*, Princeton U.P., Princeton, 2009. ↑342

[42] David G. Luenberger and Yinyu Ye, *Linear and Nonlinear Programming*, Third Edition, Springer Science, New York, 2008. ↑375, 409

[43] Mikuláš Luptáčik, *Mathematical Optimization and Economic Analysis*, Springer, New York, 2010. ↑4

[44] Olvi L. Mangasarian, *Nonlinear Programming*, Classics in Applied Mathematics, SIAM, Philadelphia, 1994. reprint of 1969 McGraw Hill original. ↑4

[45] Garth P. McCormick, *Nonlinear Programming*: *theory, algorithms, and applications*, Wiley, New York, 1983. ↑4

[46] J. R. Norris, *Markov Chains*, Cambridge U. P., Cambridge, 1997. ↑412

[47] Barrett O'Neill, *Elementary Differential Geometry*, Academic Press, New York, 1966. ↑218

[48] W. F. Osgood, *Lehrbuch der Funktionentheorie*, Fünfte Auflage, Teubner, Leibzig, Berlin, 1928. ↑152

[49] K. Otani, *A Characterization of Quasiconvex Functions*, Journal of Economic Theory **31** (1983), 194-196. ↑367

[50] Pablo Pedregal, *Introduction to Optimization*, Springer, New York, 2004. ↑445

[51] Victor V. Prasolov, *Polynomials*, Algorithms and Computation in Mathematics, Springer, New York, 2001. ↑138

[52] A. Wayne Roberts and Dale E. Varberg, *Convex Functions*, Academic Press, New York, 1973. ↑xiii, 3, 236, 241, 256, 291, 332

[53] R. Tyrrell Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, 1970. ↑235, 236, 297, 318, 332

[54] ———, *Conjugate Duality and Optimization*, Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, 1974. ↑297

[55] Walter Rudin, *Principles of Mathematical Analysis*, Third Edition, McGraw-Hill, New York, 1976. ↑xiii, xv, 41, 42, 43, 47, 154, 156, 187, 197, 208, 223, 227

[56] Herbert Scarf, *The Computation of Economic Equilibria*, Cowles Foundation Monographs, Yale University Press, New Haven, 1973. with the collaboration of Terje Hansen. ↑173

[57] Denis Serre, *Matrices*: *Theory and Applications*, Graduate Texts in Mathematics, Springer, New York, 2002. Errata online at www.umpa.ens-lyon.fr/ serre/DPF/errata.pdf. ↑xiii, 91

[58] J.-A. Serret, *Cours de Calcul Différentiel et Intégral*, Gauthier-Villars, Paris, 1868. ↑172, 341

[59] Carl P. Simon and Lawrence Blume, *Mathematics for Economists*, Norton, New York, 1994. ↑76

[60] Lawrence E. Spense, Arnold J. Insel, and Stephen H. Friedberg, *Elementary Linear Algebra*: *A Matrix Approach*, Prentice Hall, Upper Saddle River, NJ., 2000. ↑xii, 62, 71, 73, 77, 87, 88, 91, 177

[61] J. Michael Steele, *The Cauchy-Schwarz Master Class*, Cambridge, New York, 2004. ↑58

[62] L. A. Steen, *Highlights in the History of Spectral Theory*, American Math. Monthly **80** (1973), no. 4, 359-381. ↑117, 129

[63] James Stewart, *Calculus, Early Transcendentals*, Fifth Edition, Brooks/Cole, Belmont, CA, 2002. ↑xii, xiv, 3, 4, 5, 27, 35, 37, 54, 55, 141, 145, 154, 157, 158, 179, 206, 208, 217, 220, 316, 439

[64] G. J. Stigler, *The Cost of Subsistence*, J. Farm Economics **27** (1945), 303-314. ↑8

[65] David Stirzaker, *Elementary Probability*, Second Edition, Cambridge U. P., Cambridge, 2003. ↑412

[66] Josef Stoer and Christoph Witzgall, *Convexity and Optimization in Finite Dimensions I*, Die Grundlehren der mathematischen Wissenschlaften in Einzeldarstellungen, Springer-Verlag, New York, 1970. ↑236, 332

[67] Gilbert Strang, *Linear Algebra and its Applications*, Third Edition, Harcourt Brace Jovanovich, San Diego, CA, 1988. ↑xiii, 77, 89, 91, 179, 409, 418

[68] ______, *Introduction to Linear Algebra*, Second Edition, Wellesley-Cambridge, Wellesley, MA, 1998. ↑xii, 55, 62, 73, 77, 87, 89, 177

[69] ______, *Introduction to Applied Mathematics*, Second Edition, Wellesley-Cambridge, Wellesley, MA, 1986. ↑71

[70] Robert S. Strichartz, *The Way of Analysis*, Revised Edition, Jones and Bartlett, Sudbury, MA, 2000. ↑xiii, xiv, 31, 42, 43, 47, 157, 163, 187, 197, 204

[71] Rangarajan K. Sundaram, *A First Course in Optimization Theory*, Cambridge U.P., Cambridge, 1996. ↑4

[72] J. J. Sylvester, *A Demonstration of the Theorem that Every Homogeneous Quadratic Polynomial is Reducible by Real Orthogonal Substitutions to the Form of a Sum of Positive and Negative Squares*, Philosophical Magazine **23** (1852), 47-51. ↑113

[73] Herman Weyl, *The Elementary Theory of Convex Polyhedra*, Contributions to the Theory of Games, 1950. ↑296, 304

[74] Vladimir A. Zorich, *Mathematical Analysis I*, translated by Roger Cooke, Universitext, Springer, Berlin, Heidelberg, New York, 2004. ↑32, 43, 149, 156

Index