

# OPTIMIZATION

M.I.Ostrovskii

June 9, 2000

## 1 Partial derivatives and extrema

Let  $\Omega$  be a subset of  $\mathbf{R}^n$  ( $n$ -dimensional space of  $n$ -tuples of real numbers) and let  $f$  be a real-valued function on  $\Omega$ . Our course is devoted to the following problem

**PROBLEM 1.1** *Minimize (or maximize)  $f(x)$  subject to  $x \in \Omega$ .*

The set  $\Omega$  will be called a *feasible set*.

Some notation. Let  $x \in \mathbf{R}^n$ , that is

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Then  $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$ .

We consider several kinds of minimizers and maximizers.

**DEFINITION 1.1** Let  $f : \Omega \rightarrow \mathbf{R}$  be a real-valued function on a subset  $\Omega$  of  $\mathbf{R}^n$ . A point  $x^*$  is called a *local minimizer (maximizer)* of  $f$  over  $\Omega$ , if there exists  $\varepsilon > 0$  such that  $f(x) \geq f(x^*)$  ( $f(x) \leq f(x^*)$ ) for all  $x \in \Omega$  satisfying  $\|x - x^*\| < \varepsilon$ .

**DEFINITION 1.2** Let  $f : \Omega \rightarrow \mathbf{R}$  be a real-valued function on a subset  $\Omega$  of  $\mathbf{R}^n$ . A point  $x^*$  is called a *global minimizer (maximizer)* of  $f$  over  $\Omega$ , if  $f(x) \geq f(x^*)$  ( $f(x) \leq f(x^*)$ ) for all  $x \in \Omega$ .

**DEFINITION 1.3** If the inequality in any of these definitions is strict for  $x \neq x^*$ , the corresponding minimizer (maximizer) is called *strict*.

**DEFINITION 1.4** A point that is either a minimizer or a maximizer is called an *extremum*.

We restrict our attention to twice continuously differentiable functions. The standard approach to the problem in one-dimensional case is based on the following result.

**THEOREM 1.1** *Let  $f$  be twice continuously differentiable function on some open interval  $I$  ( $I$  may be infinite). Let  $x$  and  $x^*$  be two different points from  $I$ . Then there exists a point  $z$  between  $x$  and  $x^*$  such that*

$$f(x) = f(x^*) + f'(x^*)(x - x^*) + \frac{f''(z)}{2}(x - x^*)^2.$$

**PROOF.** See almost any Calculus textbook (e.g. [4] (p. 140) or [6] (Appendix B, p. A41)). Usually this theorem is called ‘Taylor’s theorem’ or ‘Taylor’s formula’. ■

Using this theorem we can derive the following result.

**THEOREM 1.2** *Let  $f$  be a twice continuously differentiable function on an open interval  $I$ .*

(1) *If  $x^*$  is a local minimizer of  $f$  over  $I$ , then  $f'(x^*) = 0$  and  $f''(x^*) \geq 0$ .*

(2) *If  $f'(x^*) = 0$  and  $f''(x^*) > 0$ , then  $x^*$  is a strict local minimizer of  $f$  over  $I$ .*

(3) *If  $f'(x^*) = 0$  and  $f''(x) \geq 0$  for every  $x \in I$ , then  $x^*$  is a global minimizer of  $f$  over  $I$ .*

**PROOF.** (1) Suppose  $x^*$  is a local minimizer of  $f$  over  $I$ . Let  $\varepsilon > 0$  be such that the condition of the Definition 1.1 is satisfied. Then

$$\frac{f(x) - f(x^*)}{x - x^*} \geq 0, \quad \text{if } x^* < x < x^* + \varepsilon,$$

and

$$\frac{f(x) - f(x^*)}{x - x^*} \leq 0, \text{ if } x^* - \varepsilon < x < x^*.$$

Since  $f$  is differentiable at  $x^*$ , then

$$f'(x^*) = \lim_{x \rightarrow x^*+} \frac{f(x) - f(x^*)}{x - x^*} \geq 0$$

and

$$f'(x^*) = \lim_{x \rightarrow x^*-} \frac{f(x) - f(x^*)}{x - x^*} \leq 0.$$

Hence  $f'(x^*) = 0$ .

To prove  $f''(x^*) \geq 0$  we assume the contrary, that is,  $f''(x^*) < 0$ . Since  $f''$  is continuous, then there exists  $\delta > 0$  such that  $f''(z) < 0$  if  $|z - x^*| < \delta$ . Let  $x \neq x^*$  be any point satisfying  $|x - x^*| < \delta$ . By Theorem 1.1

$$f(x) - f(x^*) = \frac{f''(z)}{2}(x - x^*)^2 \tag{*}$$

for some  $z$  between  $x$  and  $x^*$ . Such  $z$  satisfies  $|z - x^*| < \delta$ . Hence the right-hand side in (\*) is negative and  $x^*$  is not a local minimizer. This contradiction proves the second statement in (1).

(2) Since  $f''$  is continuous, then there exists  $\delta > 0$  such that  $f''(z) > 0$  for every  $z$  satisfying  $|z - x^*| < \delta$ . Let  $x \neq x^*$  be such that  $|x - x^*| < \delta$ . We use the same argument as above, but in this case the right-hand side of (\*) is  $> 0$ . Hence  $x^*$  is a local minimizer.

(3) Let  $x$  be an arbitrary point in  $I$ . There exist a point  $z$  between  $x$  and  $x^*$  satisfying (\*). Since  $f''(z) \geq 0$ , then  $f(x) - f(x^*) \geq 0$ . Hence  $x^*$  is a global minimizer. ■

**DEFINITION 1.5** Let  $f$  be a differentiable function on an open interval  $I$ . A point  $x \in I$  is called a *critical point* of  $f$  if  $f'(x) = 0$ .

**Remarks. 1** There are some more variations on the theme of Theorem 1.2. For example, if  $f'(x^*) = 0$ ,  $f''(x^*) > 0$  and  $f''(x) \geq 0$  for every  $x \in I$ , then  $x^*$  is a strict global minimizer of  $f$  over  $I$ .

2. One can state and prove similar result for maximizers.

3. One of the approaches to 2 is: to show that the problem of maximization of  $f$  over  $\Omega$  is equivalent to the problem of minimization of  $-f$  over  $\Omega$ .

4. Using the theorem and its analogue for maximizers we can develop the following

### STRATEGY

for finding local extrema of a differentiable function  $f$  of one variable.

1. We find its critical points.

2. We determine the sign of the second derivative at them.

- If  $f''$  at a critical point is  $> 0$ , then the critical point is a local minimizer;

- If  $f''$  at a critical point is  $< 0$ , then the critical point is a local maximizer.

- If  $f''$  at a critical point is 0, then an additional investigation is needed in order to determine the character of the critical point.

One of the standard approaches is to check whether  $f'$  changes the sign at the critical point.

- If from  $+$  to  $-$ , then local maximizer;

- If from  $-$  to  $+$ , then local minimizer;

- If does not change sign, then neither one.

The problem of finding global maximizers and minimizers is more difficult. In some situations the following definition is useful.

**DEFINITION 1.6** A real-valued function defined on some set  $\Omega$  is called *bounded from below* if there exists a real number  $m$  such that  $f(x) \geq m$  for every  $x \in \Omega$ . The function  $f$  is called *bounded from above* if there exists a real number  $M$  such that  $f(x) < M$  for every  $x \in \Omega$ .

The following result immediately follows from the definitions.

**PROPOSITION 1.1** *Let  $f$  be a real-valued function on  $\Omega$ . If  $f$  has a global minimizer on  $\Omega$ , then  $f$  is bounded from below. If  $f$  has a global maximizer on  $\Omega$  then  $f$  is bounded from above.*

**Example.** Find the local and global maximizers and minimizers of the following functions:

(a)  $f(x) = 2x^3 - 9x^2 + 12x + 1$ .

(b)  $f(x) = x + \sin x$ .

(c)  $f(x) = x^4$ .

Our next purpose is to extend Theorem 1.2 to  $n$ -dimensional case.

We need to recall some definitions from Linear Algebra and Multivariable Calculus.

**DEFINITION 1.7** Let  $L : \mathbf{R}^n \rightarrow \mathbf{R}^m$  be a function. The function  $L$  is called *linear* if

(1)  $L(x + y) = L(x) + L(y)$  for every  $x, y \in \mathbf{R}^n$ .

(2)  $L(ax) = aL(x)$  for every  $a \in \mathbf{R}$  and every  $x \in \mathbf{R}^n$ .

**DEFINITION 1.8** A function  $f : \mathbf{R}^n \rightarrow \mathbf{R}^m$  is said to be *differentiable* at  $x_0 \in \mathbf{R}^n$  if there exist a linear function  $L : \mathbf{R}^n \rightarrow \mathbf{R}^m$  satisfying

$$\lim_{x \rightarrow x_0} \frac{\|f(x) - f(x_0) - L(x - x_0)\|}{\|x - x_0\|} = 0.$$

The linear function  $L$  determined by  $f$  and  $x_0$  is called the *derivative* of  $f$  at  $x_0$ .

Recall (from Linear Alg.). There exists a one-to-one correspondence between the set of all linear functions  $L : \mathbf{R}^n \rightarrow \mathbf{R}^m$  and the set of all  $m \times n$  matrices with real entries. In more detail: for each such  $L$  there exists an  $m \times n$ -matrix

$$\begin{bmatrix} l_{1,1} & \dots & l_{1,n} \\ \vdots & \ddots & \vdots \\ l_{m,1} & \dots & l_{m,n} \end{bmatrix},$$

such that

$$L \left( \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \right) = \begin{bmatrix} l_{1,1} & \dots & l_{1,n} \\ \vdots & \ddots & \vdots \\ l_{m,1} & \dots & l_{m,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} l_{1,1}x_1 + \dots + l_{1,n}x_n \\ l_{2,1}x_1 + \dots + l_{2,n}x_n \\ \vdots \\ l_{m,1}x_1 + \dots + l_{m,n}x_n \end{bmatrix}.$$

It turns out that the matrix of the derivative can be easily described in terms of partial derivatives.

**THEOREM 1.3 (Without proof).** *Let  $f : \mathbf{R}^n \rightarrow \mathbf{R}^m$ . The function  $f$  can be written as*

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix}.$$

*If the partial derivatives of  $f_1, \dots, f_m$  are continuous at  $x$ , then  $f$  is differentiable at  $x$  and the matrix of the derivative of  $f$  at  $x$  is given by:*

$$Df(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \dots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \dots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix}.$$

The matrix of the derivative is usually called the *Jacobian matrix*.

If  $f$  is differentiable at every  $x$  in  $\mathbf{R}^n$ , then  $Df(x)$  may be considered as a function from  $\mathbf{R}^n$  to the set of all  $m \times n$  matrices. Using a natural notion of norm on the set of all  $m \times n$  matrices we can define the derivative of the function  $Df(x)$  in the same way as before. This derivative will be a linear function from  $\mathbf{R}^n$  into the set of  $m \times n$  matrices. It is a rather complicated object. We shall not study it in this generality.

We shall consider a special case of real-valued functions  $f : \mathbf{R}^n \rightarrow \mathbf{R}$  satisfying the condition:  $f$  has continuous second partial derivatives. In this case

$$Df(x) = \left[ \frac{\partial f}{\partial x_1}(x) \quad \frac{\partial f}{\partial x_2}(x) \quad \dots \quad \frac{\partial f}{\partial x_n}(x) \right].$$

**DEFINITION 1.9** Let  $A$  be an  $n \times m$ -matrix,

$$A = \begin{bmatrix} a_{1,1} & \dots & a_{1,m} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \dots & a_{n,m} \end{bmatrix}.$$

The *transpose* of  $A$  is the  $m \times n$  matrix  $B$  satisfying  $b_{i,j} = a_{j,i}$ . The transpose of  $A$  is denoted by  $A^T$ .

**DEFINITION 1.10** The transpose of  $Df(x)$  is called the *gradient* of  $f$  at  $x$  and is denoted by  $\nabla f(x)$ . That is

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}.$$

So  $\nabla f : \mathbf{R}^n \rightarrow \mathbf{R}^n$ . The matrix of the derivative of the gradient is called the *Hessian* of  $f$  at  $x$  and (with some abuse of notation) is denoted by  $D^2 f(x)$ . By the description of the derivative mentioned above we have

$$D^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{bmatrix}.$$

**THEOREM 1.4** *If  $f$  has continuous second partial derivatives, then*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial^2 f}{\partial x_j \partial x_i}(x)$$

for every  $i, j \in \{1, \dots, n\}$ .

**PROOF.** See e.g [7] (p. 174). ■

**DEFINITION 1.11** Let  $A$  be an  $n \times n$ -matrix, The matrix  $A$  is called *symmetric* if  $A^T = A$ . (Equivalent definition: if  $a_{i,j} = a_{j,i}$  for every  $i, j \in \{1, \dots, n\}$ .)

Theorem 1.4 can be restated in the following way: the Hessian of a twice continuously differentiable function is a symmetric matrix.

**Example.** Find the Jacobian matrix, the gradient, and the Hessian of the function

$$f(x_1, x_2, x_3) = x_1^2 \sin(x_2 + x_3).$$

**Remark.** We use the notation  $f(x_1, x_2, \dots, x_n)$  instead of the more formal

$$f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}\right).$$

We are going to state an  $n$ -dimensional analogue of the Taylor's Theorem. For this we need an analogue of the notion of the open interval.

**DEFINITION 1.12** A subset  $\Omega \subset \mathbf{R}^n$  is called *open* if for every point  $x \in \Omega$  there exists  $\varepsilon > 0$  such that  $y \in \Omega$  provided  $\|x - y\| < \varepsilon$ .

**Example.** The set  $A = \{x \in \mathbf{R}^n : \|x\| < 1\}$  is open.

The set  $B = \{x \in \mathbf{R}^n : \|x\| \leq 1\}$  is not open.

To prove the first statement we need the following important inequality, called “the triangle inequality”. (Its proof is usually discussed in courses of Linear Algebra.)

$$\|x + y\| \leq \|x\| + \|y\|, \text{ for every } x \text{ and } y \text{ in } \mathbf{R}^n.$$

Now, let  $x \in A$  (two-dimensional picture). We choose  $\varepsilon = 1 - \|x\|$ . Let  $\|x - y\| < \varepsilon$ . By the triangle inequality we get  $\|y\| \leq \|x\| + \|y - x\| < \|x\| + (1 - \|x\|) = 1$  (here we use  $\|y - x\| = \|x - y\|$ ). Hence  $y$  is also in  $A$ . ■

To prove the second statement we need to show that for some  $x \in B$  and for every  $\varepsilon > 0$  there exists  $y \in \mathbf{R}^n$  such that  $\|x - y\| < \varepsilon$  but  $y$  is not in  $B$ .

If we sketch the picture in two-dimensional case and observe that the condition above means that there are points that are ‘very close’ to  $x$  but

not in  $B$ . For this reason it is natural to choose  $x$  satisfying  $\|x\| = 1$ . Now let  $\varepsilon > 0$  be arbitrary. One of the possible choices of  $y$  is:

$$y = \left(1 + \frac{\varepsilon}{2}\right)x.$$

With this choice of  $y$  we have  $\|x - y\| = \frac{\varepsilon}{2} < \varepsilon$  and  $\|y\| = \frac{\varepsilon}{2} + 1 > 1$ , so  $y \notin B$ . ■

**DEFINITION 1.13** The *line segment between* points  $x$  and  $y$  in  $\mathbf{R}^n$  is the set of all points  $z$  of the form  $z = \alpha x + (1 - \alpha)y$ ,  $0 \leq \alpha \leq 1$ .

**THEOREM 1.5 (Without Proof).** *Let  $\Omega$  be an open subset of  $\mathbf{R}^n$ . Suppose that  $x, x^* \in \Omega$  and that the line segment between  $x$  and  $x^*$  is contained in  $\Omega$ . Suppose also that  $f : \Omega \rightarrow \mathbf{R}$  has continuous second partial derivatives. Then there exists  $z$  on the line segment between  $x$  and  $x^*$  such that*

$$f(x) = f(x^*) + Df(x^*)(x - x^*) + \frac{1}{2}(x - x^*)^T D^2 f(z)(x - x^*),$$

where the products are products of matrices and by  $A^T$  we denote the transpose of  $A$ .

**Remark.** The standard approach to the proof of this theorem is the following. We consider a function  $h(t)$  defined by  $h(t) = f(tx + (1 - t)x^*)$ . This function is twice continuously differentiable and is defined on some open interval containing  $[0, 1]$ . Observe, also that  $h(0) = f(x^*)$  and  $h(1) = f(x)$ . Applying the (one-dimensional) Taylor's Theorem to  $h$  we get the result.

**THEOREM 1.6** *If  $x^*$  is a local minimizer (or a local maximizer) of  $f$  over an open set  $\Omega$ , then  $Df(x^*) = 0$ , that is  $\frac{\partial f}{\partial x_i}(x^*) = 0$  for every  $i = 1, \dots, n$ .*

PROOF. We shall prove the result for local minimizers only (for maximizers the proof is the same). Let

$$x^* = \begin{bmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_n^* \end{bmatrix}$$

For each positive integer  $i$  satisfying  $1 \leq i \leq n$  we introduce a function  $h_i$  of one variable  $x_i$  as

$$h_i(x_i) = f(x_1^*, \dots, x_{i-1}^*, x_i, x_{i+1}^*, \dots, x_n^*).$$

Let  $\varepsilon > 0$  be such that  $\|x - x^*\| < \varepsilon$  implies  $x \in \Omega$  and  $f(x) \geq f(x^*)$ . (Existence of such  $\varepsilon$  follows from the definitions of an open set and of a local minimizer.) With such choice of  $\varepsilon$  the point  $x_i^*$  is a (global) minimizer of  $h_i$  over the open interval  $(x_i^* - \varepsilon, x_i^* + \varepsilon)$  (two-dimensional picture). By the one-dimensional result we have  $\frac{dh_i}{dx_i}(x_i^*) = 0$ . It remains to observe that

$$\frac{dh_i}{dx_i}(x_i^*) = \frac{\partial f}{\partial x_i}(x^*).$$

■

To derive some consequences of these theorems we need some more definitions.

Let  $A$  be an  $n \times n$  matrix with real entries. The real-valued function  $x \mapsto x^T A x$  on  $\mathbf{R}^n$  is called the *quadratic form associated with  $A$* .

**DEFINITION 1.14** (a) A quadratic form  $x^T A x$  is called *positive definite* if  $x^T A x > 0$  for every  $x \in \mathbf{R}^n$ ,  $x \neq 0$ .

(b) A quadratic form  $x^T A x$  is called *positive semidefinite* if  $x^T A x \geq 0$  for every  $x \in \mathbf{R}^n$ .

(c) A quadratic form  $x^T A x$  is called *negative definite* if  $x^T A x < 0$  for every  $x \in \mathbf{R}^n$ ,  $x \neq 0$ .

(d) A quadratic form  $x^T A x$  is called *negative semidefinite* if  $x^T A x \leq 0$  for every  $x \in \mathbf{R}^n$ .

(e) A quadratic form  $x^T Ax$  is called *indefinite* if  $x^T Ax > 0$  for some  $x \in \mathbf{R}^n$  and  $x^T Ax < 0$  for some other  $x \in \mathbf{R}^n$ .

With this terminology we can state a multidimensional analogue of the one-dimensional result proved earlier.

**THEOREM 1.7** *Let  $f : \mathbf{R}^n \rightarrow \mathbf{R}$  be twice continuously differentiable.*

1. *If  $x^*$  is a local minimizer of  $f$  over  $\mathbf{R}^n$ , then  $Df(x^*) = 0$  and  $x^T D^2 f(x^*)x$  is positive semidefinite.*

2. *If  $Df(x^*) = 0$  and  $x^T D^2 f(x^*)x$  is positive definite, then  $x^*$  is a strict local minimizer of  $f$  over  $\mathbf{R}^n$ .*

3. *If  $Df(x^*) = 0$  and  $x^T D^2 f(z)x$  is positive semidefinite for every  $z \in \mathbf{R}^n$ , then  $x^*$  is a global minimizer of  $f$  over  $\mathbf{R}^n$ .*

PROOF. 1. By Theorem 1.6 it is enough to show that if  $Df(x^*) = 0$  and  $x^T D^2 f(x^*)x$  is not positive semidefinite, then  $x^*$  is not a local minimizer. Suppose that  $x^T D^2 f(x^*)x$  is not positive semidefinite. It means that there exists  $u \in \mathbf{R}^n$  such that  $u^T D^2 f(x^*)u < 0$ . Since  $f$  is twice continuously differentiable, then the mapping  $z \mapsto u^T D^2 f(z)u$  is a continuous function from  $\mathbf{R}^n$  to  $\mathbf{R}$ . Hence there exists  $\varepsilon > 0$  such that if  $\|z - x^*\| < \varepsilon$ , then  $u^T D^2 f(z)u < 0$ . Let  $\delta > 0$  be any number satisfying  $\|\delta u\| < \varepsilon$ . Let  $x = x^* + \delta u$ . Using the Taylor's Theorem we get

$$f(x^* + \delta u) = f(x^*) + \frac{1}{2}(\delta u)^T D^2 f(z)(\delta u),$$

where  $z$  is on the line segment between  $x^* + \delta u$  and  $x^*$ . That is  $z = \alpha(x^* + \delta u) + (1 - \alpha)x^* = x^* + \alpha\delta u$ , where  $0 \leq \alpha \leq 1$ . Therefore

$$\|z - x^*\| = \|\alpha\delta u\| = \alpha\|\delta u\| < \alpha\varepsilon \leq \varepsilon.$$

Hence

$$\frac{1}{2}(\delta u)^T D^2 f(z)(\delta u) = \frac{1}{2}\delta^2(u^T D^2 f(z)u) < 0$$

and  $f(x) < f(x^*)$ , so  $x^*$  is not a local minimizer.

2. Is similar, but somewhat more complicated. Without proof.

**3.** Let  $x \in \mathbf{R}^n$  be arbitrary. By Taylor's Theorem we have

$$f(x) = f(x^*) + \frac{1}{2}(x - x^*)^T D^2 f(z)(x - x^*)$$

for some  $z$  between  $x$  and  $x^*$ . Since the quadratic form associated with  $D^2 f(z)$  is positive semidefinite, then  $(x - x^*)^T D^2 f(z)(x - x^*) \geq 0$ . Hence  $f(x) \geq f(x^*)$ . ■

**Remarks. 1.** There exists a natural analogue of this result for maximizers.

**2.** Parts **1** and **2** of Theorem 1.7 remain to be true for any twice continuously differentiable real-valued function on an open subset of  $\mathbf{R}^n$ .

Theorem 1.7 and the first remark after it suggest the following approach for finding local maximizers and minimizers of  $f : \mathbf{R}^n \rightarrow \mathbf{R}$ .

**DEFINITION 1.15** A point  $x^* \in \mathbf{R}^n$  is called a *critical point* of  $f$  if  $Df(x^*) = 0$ .

## STRATEGY

**Step 1.** We find all critical points of  $f$ .

**Step 2.** For each critical point  $x^*$  we verify whether the quadratic form  $x^T D^2 f(x^*)x$  is positive definite or negative definite. (It is the so-called Second Order Sufficient Condition (SOSC)). The points that satisfy SOSC are either strict local minimizers (positive definite) or strict local maximizers (negative definite).

**Step 3.** For each critical point  $x^*$  that does not satisfy SOSC we verify whether the quadratic form  $x^T D^2 f(x^*)x$  is indefinite. The points at which it is the case are neither local minimizers nor local maximizers.

In the remaining cases the problem is more complicated and we shall not discuss it.

At the moment we cannot really use this strategy because we do not know how to establish the required properties of quadratic forms.

Now we shall study quadratic forms. We need the following result on properties of transposes.

**THEOREM 1.8** (Lin. Alg.) *Let matrices  $A$  and  $B$  be such that the product  $AB$  is defined. Then*

$$(a) (AB)^T = B^T A^T;$$

$$(b) (A^T)^T = A;$$

*Let  $A$  be a square matrix. Then*

$$(c) \frac{A+A^T}{2} \text{ is a symmetric matrix.}$$

*(d) The quadratic forms associated with  $A$  and  $\frac{A+A^T}{2}$  are the same.*

**PROOF.** We shall prove the statement (d) only. Observe that all  $1 \times 1$  matrices are symmetric. Hence  $(x^T Ax)^T = x^T Ax$ . Using (a) and (b) we get  $x^T A^T x = x^T Ax$ . Therefore

$$x^T \left( \frac{A + A^T}{2} \right) x = \frac{x^T Ax + x^T A^T x}{2} = \frac{2x^T Ax}{2} = x^T Ax.$$

Another proof: to write both quadratic forms in terms of entries of  $A$  and  $x$ . ■

Now we prove a criterion that allows us to determine whether the quadratic form  $x^T Qx$  is positive definite for a symmetric matrix  $Q$ .

**DEFINITION 1.16** Let  $Q$  be an  $n \times n$  matrix. The *leading principal minors* of  $Q$  are  $\det Q$  and the determinants of the matrices obtained from  $Q$  by removing the last  $k$  columns and the last  $k$  rows ( $k = 1, \dots, n - 1$ ).

**Example.** Let

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

Leading principal minors of  $Q$  are

$$\det \begin{bmatrix} 1 & 2 & 5 \\ 3 & 4 & 6 \\ 7 & 8 & 9 \end{bmatrix}; \det \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}; \det[1].$$

**DEFINITION 1.17** A symmetric matrix  $A$  is called *positive definite* (*positive semidefinite*, *negative definite*, *negative semidefinite*, *indefinite*) if the associated quadratic form is positive definite (positive semidefinite, negative definite, negative semidefinite, indefinite).

**THEOREM 1.9 (Sylvester's Criterion.)** *A symmetric matrix  $Q$  is positive definite if and only if all leading principal minors of  $Q$  are positive.*

**PROOF.** We shall use the following notation:

$$\Delta_1 = q_{1,1}, \Delta_2 = \det \begin{bmatrix} q_{1,1} & q_{1,2} \\ q_{2,1} & q_{2,2} \end{bmatrix}, \dots, \Delta_n = \det \begin{bmatrix} q_{1,1} & \dots & q_{1,n} \\ \vdots & \ddots & \vdots \\ q_{n,1} & \dots & q_{n,n} \end{bmatrix}.$$

**Step 1.** If  $\Delta_k = 0$  for some  $1 \leq k \leq n$ , then  $Q$  is not positive definite.

By a well-known result from Linear Algebra the condition  $\Delta_k = 0$  implies that there exists a non-zero vector

$$\begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} \in \mathbf{R}^k$$

such that

$$\begin{bmatrix} q_{1,1} & \dots & q_{1,k} \\ \vdots & \ddots & \vdots \\ q_{k,1} & \dots & q_{k,k} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}. \quad (*)$$

Let

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbf{R}^n.$$

Let us show that  $x^T Q x = 0$ . Consider

$$Qx = \begin{bmatrix} q_{1,1} & \cdots & q_{1,k} & \cdots & q_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ q_{k,1} & \cdots & q_{k,k} & \cdots & q_{k,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ q_{n,1} & \cdots & q_{n,k} & \cdots & q_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

From the definition of the product of matrices and (\*) it follows that this product is of the form

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \alpha_{k+1} \\ \vdots \\ \alpha_n \end{bmatrix},$$

that is, the upper  $k$  entries are zeros.

Hence

$$x^T Q x = [x_1 \ \cdots \ x_k \ 0 \ \cdots \ 0] \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \alpha_{k+1} \\ \vdots \\ \alpha_n \end{bmatrix} = 0.$$

**Step 2.** Assume that the numbers  $\Delta_1, \Delta_2, \dots, \Delta_n$  are nonzero. Then there exists a basis in  $\mathbf{R}^n$  such that

$$x^T Q x = \frac{1}{\Delta_1} \tilde{x}_1^2 + \frac{\Delta_1}{\Delta_2} \tilde{x}_2^2 + \dots + \frac{\Delta_{n-1}}{\Delta_n} \tilde{x}_n^2,$$

where  $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n$  are the coordinates of  $x$  in the new basis.

Let  $\{f_1, \dots, f_n\}$  be some basis in  $\mathbf{R}^n$ ;

$$f_i = \begin{bmatrix} f_{1,i} \\ f_{2,i} \\ \vdots \\ f_{n,i} \end{bmatrix} \in \mathbf{R}^n, \quad i = 1, 2, \dots, n.$$

Let

$$\tilde{x} = \begin{bmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_n \end{bmatrix},$$

where  $\tilde{x}_1, \dots, \tilde{x}_n$  are the coordinates of  $x$  in the basis  $f_1, \dots, f_n$ . Then, by the definition of these notions we have

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} f_{1,1} & \dots & f_{1,n} \\ \vdots & \ddots & \vdots \\ f_{n,1} & \dots & f_{n,n} \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_n \end{bmatrix}.$$

We write this equality in the form  $x = F\tilde{x}$ . Hence  $x^T = \tilde{x}^T F^T$  and  $x^T Q x = \tilde{x}^T F^T Q F \tilde{x} = \tilde{x}^T \tilde{Q} \tilde{x}$ , where  $\tilde{Q} = F^T Q F$ .

We shall choose the basis  $f_1, \dots, f_n$  in such a way that

**I.** The matrix  $F$  is *upper triangular*, that is  $f_{i,j} = 0$  if  $j < i$  (each entry below the diagonal is equal to zero).

**II.**  $QF$  is *lower triangular* and each entry on the diagonal is equal to 1.

**III.**  $f_{1,1} = \frac{1}{\Delta_1}, f_{2,2} = \frac{\Delta_1}{\Delta_2}, \dots, f_{n,n} = \frac{\Delta_{n-1}}{\Delta_n}$ .

To find such  $F$  we solve the following systems of equations

$$\begin{bmatrix} q_{1,1} & \dots & q_{1,k} \\ \vdots & \ddots & \vdots \\ q_{k,1} & \dots & q_{k,k} \end{bmatrix} \begin{bmatrix} f_{1,k} \\ \vdots \\ f_{k,k} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

This system has a unique solution because (as we have assumed)  $\Delta_k \neq 0$ .

If  $k = 1$ , we get a trivial system  $q_{1,1}f_{1,1} = 1$ . Its solution is  $f_{1,1} = \frac{1}{q_{1,1}} = \frac{1}{\Delta_1}$ . For  $k > 1$  we get (using Cramer's rule):

$$f_{k,k} = \frac{\det \begin{bmatrix} q_{1,1} & \cdots & q_{1,k-1} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ q_{k-1,1} & \cdots & q_{k-1,k-1} & 0 \\ q_{k,1} & \cdots & q_{k,k-1} & 1 \end{bmatrix}}{\det \begin{bmatrix} q_{1,1} & \cdots & q_{1,k} \\ \vdots & \ddots & \vdots \\ q_{k,1} & \cdots & q_{k,k} \end{bmatrix}}.$$

Expanding the determinant in the numerator with respect to the last column we find that it is equal to  $\Delta_{k-1}$ . Hence

$$f_{k,k} = \frac{\Delta_{k-1}}{\Delta_k}.$$

Straightforward verification shows that  $F$  satisfies **I** and **II**.

What can we say about  $\tilde{Q} = F^T Q F$ ?

Since  $\tilde{Q} = F^T(QF)$ , using the properties of  $F$  and  $QF$  we get:

**A.**  $\tilde{Q}$  has the entries  $f_{1,1}, \dots, f_{n,n}$  on the diagonal.

**B.**  $\tilde{Q}$  has zero entries above the diagonal.

Also  $\tilde{Q}^T = (F^T Q F)^T = F^T Q^T (F^T)^T = F^T Q F$  (we use the facts that  $Q$  is symmetric and  $(F^T)^T = F$ ). Hence

**C.**  $\tilde{Q}$  is a symmetric matrix.

**Conclusion.** The matrix  $\tilde{Q} = F^T Q F$  is of the form

$$\begin{bmatrix} f_{1,1} & 0 & \cdots & 0 \\ 0 & f_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & f_{n,n} \end{bmatrix},$$

where

$$f_{1,1} = \frac{1}{\Delta_1}, f_{2,2} = \frac{\Delta_1}{\Delta_2}, \dots, f_{n,n} = \frac{\Delta_{n-1}}{\Delta_n}.$$

We have

$$x^T Q x = \tilde{x}^T (F^T Q F) \tilde{x} = [\tilde{x}_1 \ \dots \ \tilde{x}_n] \begin{bmatrix} \frac{1}{\Delta_1} & 0 & \dots & 0 \\ 0 & \frac{\Delta_1}{\Delta_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{\Delta_{n-1}}{\Delta_n} \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_n \end{bmatrix} =$$

$$\frac{1}{\Delta_1} \tilde{x}_1^2 + \frac{\Delta_1}{\Delta_2} \tilde{x}_2^2 + \dots + \frac{\Delta_{n-1}}{\Delta_n} \tilde{x}_n^2.$$

**Step 3.** We use this representation to finish the proof of the theorem.

We need to show

**a.** If all leading principal minors of  $Q$  are positive, then  $Q$  is positive definite.

**b.** If one of the leading principal minors of  $Q$  is non-positive then  $Q$  is not positive definite.

To prove **a** we use the representation proved in Step 2. In fact, if  $x \neq 0$ , then at least one of the numbers  $\tilde{x}_1, \dots, \tilde{x}_n$  is nonzero. Let  $\tilde{x}_k \neq 0$ . Then

$$x^T Q x = \frac{1}{\Delta_1} \tilde{x}_1^2 + \frac{\Delta_1}{\Delta_2} \tilde{x}_2^2 + \dots + \frac{\Delta_{n-1}}{\Delta_n} \tilde{x}_n^2 \geq \frac{\Delta_{k-1}}{\Delta_k} \tilde{x}_k^2 > 0.$$

To prove **b** we consider two possibilities:

- one of the leading principal minors is 0;
- all of the leading principal minors are nonzero and some of them are negative.

The first possibility was already considered in Step 1 where we proved that in this case  $Q$  is not positive definite.

Consider the second possibility. Let  $k$  be the smallest positive integer satisfying  $\Delta_k < 0$ .

Let  $x = f_k$  (the vector of the new basis with the corresponding index. Then the coordinates of  $x$  with respect to the basis  $f_1, \dots, f_n$  are  $0, \dots, 0, 1, 0, \dots, 0$ , where 1 is the  $k$ th coordinate.

Therefore  $x^T Qx = \frac{\Delta_{k-1}}{\Delta_k}$ , if  $k > 1$ , and  $x^T Qx = \frac{1}{\Delta_1}$ , if  $k = 1$ .

In any of these cases  $x^T Qx < 0$  (in the first case we use the assumption that  $\Delta_{k-1} > 0$ ). ■

**Remark.** In order to determine whether  $x^T Qx$  is positive definite for non-symmetric  $Q$  we apply Sylvester's criterion to  $\frac{Q+Q^T}{2}$ . (It is OK because the quadratic forms associated with  $Q$  and  $\frac{Q+Q^T}{2}$  are the same, see Theorem 1.8 (d).)

**Warning.** For non-symmetric  $Q$  it may happen that all leading principal minors of  $Q$  are positive, but the form  $x^T Qx$  is not positive-definite.

**Example.** Let

$$Q = \begin{bmatrix} 2 & 0 & 1 \\ 8 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix}.$$

It can be verified that all leading principal minors of  $Q$  are positive but the quadratic form  $x^T Qx$  is not positive definite.

**Remark.** Observe that  $x^T Qx$  is negative definite if and only if  $x^T (-Q)x$  is positive definite. Therefore in order to determine whether  $x^T Qx$  is negative definite we apply the Sylvester's criterion to  $-Q$ .

**Example.** Show that the matrix

$$Q = \begin{bmatrix} -3 & 1 & 1 \\ 1 & -3 & 1 \\ 1 & 1 & -3 \end{bmatrix}$$

is negative definite.

The remark above can be used to derived the following criterion.

**THEOREM 1.10** *Let  $Q$  be a symmetric matrix,*

$$Q = \begin{bmatrix} q_{1,1} & \cdots & q_{1,n} \\ \vdots & \ddots & \vdots \\ q_{n,1} & \cdots & q_{n,n} \end{bmatrix}.$$

*Let*

$$\Delta_1 = q_{1,1}, \quad \Delta_2 = \det \begin{bmatrix} q_{1,1} & q_{1,2} \\ q_{2,1} & q_{2,2} \end{bmatrix}, \dots, \Delta_n = \det \begin{bmatrix} q_{1,1} & \cdots & q_{1,n} \\ \vdots & \ddots & \vdots \\ q_{n,1} & \cdots & q_{n,n} \end{bmatrix}.$$

*Then  $Q$  is negative definite if and only if*

$$\Delta_1 < 0, \Delta_2 > 0, \Delta_3 < 0, \dots, (-1)^n \Delta_n > 0$$

*(the leading principal minors of odd orders are negative and the leading principal minors of even orders are positive.)*

**PROOF.** Let  $\Lambda_1, \dots, \Lambda_n$  be the leading principal minors of  $-Q$ . By the remark above  $Q$  is negative definite if and only if  $\Lambda_1 > 0, \dots, \Lambda_n > 0$ . So it is enough to prove that  $\Lambda_k = (-1)^k \Delta_k$ . This identity follows from the following fact: the determinant of the matrix obtained after we change the signs of all entries in a column (or a row) of a matrix  $B$  is equal to  $-\det B$ .

■

To formulate a criterion for positive semidefinite quadratic forms we need the following definition.

**DEFINITION 1.18** Let  $Q$  be an  $n \times n$  matrix. A *principal submatrix* of  $Q$  is a submatrix obtained in the following way: we remove some collection of columns of  $Q$  and remove the collection of rows with the same indices.

**Example.** Let

$$Q = \begin{bmatrix} 2 & 3 & 1 \\ 4 & 7 & 6 \\ 5 & 0 & 8 \end{bmatrix}.$$

Principal submatrices of  $Q$  are

$$\begin{bmatrix} 2 & 3 & 1 \\ 4 & 7 & 6 \\ 5 & 0 & 8 \end{bmatrix}; \begin{bmatrix} 2 & 3 \\ 4 & 7 \end{bmatrix}; \begin{bmatrix} 2 & 1 \\ 5 & 8 \end{bmatrix}; \begin{bmatrix} 7 & 6 \\ 0 & 8 \end{bmatrix}; [2]; [7]; [8].$$

**Remark.** It can be proved that an  $n \times n$  matrix has  $2^n - 1$  principal submatrices.

**DEFINITION 1.19** Determinants of principal submatrices of  $Q$  are called *principal minors* of  $Q$ .

**THEOREM 1.11** *A symmetric matrix  $Q$  is positive semidefinite if and only if all principal minors of  $Q$  are non-negative.*

**WITHOUT PROOF.** The proof of this result requires more results on determinants than are usually included into a Linear Algebra course. ■

**Remark.** To prove Theorem 1.11 one can use the fact that  $Q$  is positive semidefinite if and only if  $Q + \varepsilon I$  is positive definite for every  $\varepsilon > 0$ , where  $I$  is the identity matrix.

**Warning.** It may happen that all leading principal minors of a matrix are non-negative, but the associated quadratic form is not positive semidefinite.

**Example.** Let

$$Q = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

It can be shown that although all leading principal minors of  $Q$  are non-negative, the quadratic form  $x^T Q x$  is not positive semidefinite.

**Remark.** A quadratic form  $x^T Q x$  is negative semidefinite if and only if  $x^T (-Q) x$  is positive semidefinite. Therefore in order to determine whether  $x^T Q x$  is negative semidefinite we apply Theorem 1.11 to  $-Q$ .

This remark can be used to prove the following criterion.

**THEOREM 1.12** *A symmetric matrix  $Q$  is negative semidefinite if all of its principal minors of odd orders are non-positive and all of its principal minors of even orders are non-negative.*

PROOF. Can be proved in the same way as Theorem 1.10. ■

**Remark.** A quadratic form  $x^T Q x$  is indefinite if and only if it is neither positive semidefinite nor negative semidefinite.

Using Theorems 1.11 and 1.12 we get:

**THEOREM 1.13** *A symmetric matrix  $Q$  is indefinite if either one of its principal minors of even order is negative or it has principal minors of odd orders of different signs. In particular,  $Q$  is indefinite if some of its diagonal entries are positive and some others are negative.*

Now we know (at least in principle) how to answer the question: whether the quadratic form  $x^T Q x$  positive definite, positive semidefinite, negative definite, negative semidefinite, or indefinite?

1. If  $Q$  is non-symmetric, we symmetrize it, that is find  $\frac{Q^T + Q}{2}$ .
2. We use the table:

positive definite	leading principal minors are $> 0$
negative definite	leading principal minors of odd orders are $< 0$ & leading principal minors of even orders are $> 0$
positive semidefinite	all principal minors are $\geq 0$
negative semidefinite	principal minors of odd orders are $\leq 0$ & principal minors of even orders are $\geq 0$
indefinite	none of the listed above, see Theorem 1.13

**Remark.** It is worthwhile (at least for matrices of high orders) first to find the signs of principal submatrices of small orders. The first step is to look at the diagonal (this can be done even before we symmetrize the matrix, because symmetrization does not change the diagonal).

**Examples.** Is the quadratic form  $x^T Q x$  positive definite, positive semi-definite, negative definite, negative semidefinite, or indefinite?

(a)

$$Q = \begin{bmatrix} -2 & 11 & 100 & 1000 \\ 1 & -2 & 12 & 14 \\ -4 & 1 & -2 & 5 \\ 4 & 7 & 15 & 4 \end{bmatrix}.$$

*Solution.* Since some of the elements on the diagonal (=principal minors of order 1) are negative and some other elements on the diagonal are positive, the matrix is indefinite.

(b)

$$Q = \begin{bmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ -4 & 1 & -2 \end{bmatrix}.$$

*Solution.* All diagonal elements are negative. Hence the quadratic form can be either negative definite, negative semidefinite or indefinite.

The matrix is not symmetric.

So we consider

$$T = \frac{Q + Q^T}{2} = \begin{bmatrix} -2 & 1 & -2 \\ 1 & -2 & 1 \\ -2 & 1 & -2 \end{bmatrix}$$

Principal minors of order 2 of  $T$  are:

$$\begin{vmatrix} -2 & 1 \\ 1 & -2 \end{vmatrix} = 3 \text{ (twice) and } \begin{vmatrix} -2 & -2 \\ -2 & -2 \end{vmatrix} = 0$$

All of them are non-negative, so to answer the question we have to compute the principal minor of order 3, that is  $\det T$ . We have  $\det T = 0$  (since (e.g.) the first and the third columns coincide).

Hence  $x^T Q x$  is negative semidefinite.

**Exercises. 1.** Let

$$Q = \begin{bmatrix} -3 & 0 & 0 \\ 2 & -3 & 0 \\ 0 & 2 & -3 \end{bmatrix}.$$

Is the quadratic form  $x^T Q x$  positive definite, positive semidefinite, negative definite, negative semidefinite, or indefinite?

2. Whether the quadratic form  $x^T Q x$  is positive semidefinite?

(a)

$$Q = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3 & 1 \end{bmatrix}.$$

(b)

$$Q = \begin{bmatrix} 5 & 1 & 5 \\ 1 & 5 & 1 \\ 5 & 1 & 5 \end{bmatrix}.$$

Now we can use the strategy for finding local minimizers and local maximizers that was described earlier.

**Example.** Find the critical points of the function  $f$ . Determine (if possible) the nature of the critical points using the Hessian.

$$f(x_1, x_2, x_3) = x_1^2 + \frac{1}{4}(x_1 + x_2)^4 + \frac{1}{3}(x_1 + x_2)^3 + e^{x_3^2}.$$

*Solution.*  $Df(x_1, x_2, x_3) = 0$  is equivalent to the system

$$\begin{cases} 2x_1 + (x_1 + x_2)^3 + (x_1 + x_2)^2 = 0 \\ (x_1 + x_2)^3 + (x_1 + x_2)^2 = 0 \\ 2x_3e^{x_3^2} = 0 \end{cases}$$

Subtracting the second equation from the first we get  $2x_1 = 0$ , so  $x_1 = 0$ . Now, we get from the second equation  $x_2^3 + x_2^2 = 0$ . It has two solutions  $x_2 = 0$  and  $x_2 = -1$ . The third equation is equivalent to  $x_3 = 0$ .

Two critical points:  $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$  and  $\begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}$

$$D^2f(x) =$$

$$\begin{bmatrix} 2 + 3(x_1 + x_2)^2 + 2(x_1 + x_2) & 3(x_1 + x_2)^2 + 2(x_1 + x_2) & 0 \\ 3(x_1 + x_2)^2 + 2(x_1 + x_2) & 3(x_1 + x_2)^2 + 2(x_1 + x_2) & 0 \\ 0 & 0 & 2e^{x_3^2} + 4x_3^2e^{x_3^2} \end{bmatrix}$$

$$D^2f \left( \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \right) = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

The corresponding form is positive semidefinite, but not positive definite.

$$D^2f \left( \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} \right) = \begin{bmatrix} 2 + 3 - 2 & 3 - 2 & 0 \\ 3 - 2 & 3 - 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Leading principal minors are 3,  $\begin{vmatrix} 3 & 1 \\ 1 & 1 \end{vmatrix} = 3 - 1 = 2$ , and  $\begin{vmatrix} 3 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{vmatrix} = 4$

Hence  $x^T D^2f \left( \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} \right) x$  is positive definite.

**Answer.**

$\begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}$  is a strict local minimizer;

$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$  the nature of this critical point is not determined by the Hessian.

**Exercise.** Find the critical points of the function  $f$ . Determine (if possible) the nature of the critical points using the Hessian.

(a)  $f(x_1, x_2, x_3) = (x_1 - x_2)^2 + x_1^2 + e^{(x_3-1)^2}$ .

(b)  $f(x_1, x_2, x_3) = x_1^4 + (x_2 + x_3)^2 + e^{x_3^2}$ .

(c)  $f(x_1, x_2, x_3) = x_1^2 + x_2^2 + 2x_1x_2 + 2x_3^2 + 4x_1x_3$ .

(d)  $f(x_1, x_2, x_3) = (x_1 + x_2)^3 - 3(x_1 + x_2) + x_2^2 + e^{x_3^2}$ .

## 2 Problems with equality constraints

We consider the following problem:

**minimize (or maximize)  $f(x)$  over the set  $\{x : h(x) = 0\}$ ,**

where  $f : \mathbf{R}^n \rightarrow \mathbf{R}$  and  $h : \mathbf{R}^n \rightarrow \mathbf{R}^m$  ( $m < n$ ) are continuously differentiable.

Usually the problem is phrased in the following way:

**minimize (or maximize)  $f(x)$  subject to  $h(x) = 0$ ,**

Let

$$h(x) = \begin{bmatrix} h_1(x) \\ \vdots \\ h_m(x) \end{bmatrix}.$$

**DEFINITION 2.1** Let  $X_1, \dots, X_n$  be matrices of the same size. They are called *linearly independent* if the equality  $a_1X_1 + \dots + a_nX_n = 0$  (where  $a_1, \dots, a_n$  are real numbers) implies  $a_1 = \dots = a_n = 0$ . A matrix  $X$  is called a *linear combination* of matrices  $X_1, \dots, X_n$  if  $X = b_1X_1 + \dots + b_nX_n$  for some real numbers  $b_1, \dots, b_n$ .

**THEOREM 2.1 (Lagrange Multiplier Theorem)** *Let  $x^*$  be a local minimizer (or maximizer) of  $f(x)$  subject to  $h(x) = 0$ . Suppose that the derivatives*

$$Dh_1(x^*), \dots, Dh_m(x^*)$$

*are linearly independent. Then  $Df(x^*)$  is a linear combination of*

$$Dh_1(x^*), \dots, Dh_m(x^*).$$

**WITHOUT PROOF.** Can be proved using the Implicit Function Theorem. We do not discuss the proof because as far as I know you have studied the Implicit Function Theorem only for one function of two variables. ■

How can we use this theorem to find minimizers and maximizers?

We do the following: we introduce the so-called *Lagrangian function*:

$$L(x_1, \dots, x_n, \lambda_1, \dots, \lambda_m) := f(x_1, \dots, x_n) + \sum_{i=1}^m \lambda_i h_i(x_1, \dots, x_n).$$

and consider the system:

$$\begin{aligned} \frac{\partial L}{\partial x_r}(x_1, \dots, x_n, \lambda_1, \dots, \lambda_m) &= 0, \quad r = 1, \dots, n; \\ h_i(x_1, \dots, x_n) &= 0, \quad i = 1, \dots, m. \end{aligned}$$

**Observation.** The Lagrange Multiplier Theorem implies that if

$$x^* = \begin{bmatrix} x_1^* \\ \vdots \\ x_n^* \end{bmatrix}$$

is a local minimizer (or maximizer) of  $f$  subject  $h(x) = 0$  and the derivatives  $Dh_1(x^*), \dots, Dh_m(x^*)$  are linearly independent, then there is a solution of the system of the form

$$(x_1^*, \dots, x_n^*, \lambda_1^*, \dots, \lambda_m^*).$$

In fact, the Lagrange Multiplier Theorem implies that there exist

$$b_1, \dots, b_m$$

such that

$$Df(x^*) = \sum_{i=1}^m b_i Dh_i(x^*).$$

This equation can be rewritten as

$$\frac{\partial L}{\partial x_r}(x_1, \dots, x_n, -b_1, \dots, -b_m) = 0, \quad r = 1, \dots, n.$$

That is  $(x_1^*, \dots, x_n^*, -b_1, \dots, -b_m)$  is a solution of the system.

To check whether a solution of the system corresponds to a local minimizer we can use the corresponding versions of the second order conditions (see [2], Section 19.5). But in this case the conditions become rather complicated and we shall not study them.

It turns out that in many cases we can find even global minimizers and maximizers without using second order conditions.

To state and prove the corresponding result we need the following.

**DEFINITION 2.2** A subset  $A \subset \mathbf{R}^n$  is called *closed* if its complement (that is the set of all points in  $\mathbf{R}^n$  that are not in  $A$ ) is open.

**Examples.**  $A = \{x : \|x\| \leq 1\}$ ,  $B = \{x : \|x\| \geq 1\}$  are closed sets.

**PROPOSITION 2.1** Let  $h : \mathbf{R}^n \rightarrow \mathbf{R}^m$  be a continuous function. Then the set  $\{x \in \mathbf{R}^n : h(x) = 0\}$  is closed.

**PROOF.** We need to prove that the set  $A = \{x \in \mathbf{R}^n : h(x) \neq 0\}$  is open. Let  $x \in A$ . Then  $\|h(x)\| > 0$ . Let  $\delta = \|h(x)\| > 0$ . By the definition of a continuous function there exists  $\varepsilon > 0$  such that  $\|y - x\| < \varepsilon$  implies  $\|h(y) - h(x)\| < \delta$ . Therefore if  $\|y - x\| < \varepsilon$  then  $\|h(y)\| \geq \|h(x)\| - \|h(x) - h(y)\| > \|h(x)\| - \delta = 0$ , so  $h(y) \neq 0$  and  $y \in A$ . Hence  $A$  is open and the set  $\{x \in \mathbf{R}^n : h(x) = 0\}$  is closed. ■

**DEFINITION 2.3** A subset  $A \subset \mathbf{R}^n$  is called *bounded* if there exists  $0 \leq C < \infty$  such that  $\|x\| \leq C$  for every  $x$  in  $A$ .

**Examples.** 1.  $A_1 = \{x \in \mathbf{R}^n : \|x\| < 2\}$  is bounded.

2.  $A_2 = \{x \in \mathbf{R}^n : \|x\| \leq 2\}$  is bounded.

3.  $A_3 = \{x \in \mathbf{R}^n : \|x\| \geq 2\}$  is unbounded.

4.  $A_4 = \{x \in \mathbf{R}^n : |x_1| < 1, |x_2| < 1, \dots, |x_n| < 1\}$  is bounded.

**THEOREM 2.2 (Weierstrass Theorem).** *Let  $A$  be a closed bounded subset of  $\mathbf{R}^n$  and let  $f$  be a continuous function  $f : A \rightarrow \mathbf{R}$ . Then  $f$  has global minimizers and global maximizers over  $A$ .*

WITHOUT PROOF. I hope that you will study this result in Advanced Calculus. ■

**COROLLARY 2.1** *Consider the problem:*

**Minimize and maximize  $f(x)$  subject to  $h(x) = 0$**

*( $f$  and  $h$  are as above).*

*Suppose that*

*(1) The derivatives*

$$Dh_1(x), \dots, Dh_m(x)$$

*are linearly independent for every  $x$  satisfying  $h(x) = 0$ .*

*(2) The set  $\{x : h(x) = 0\}$  is bounded.*

*Let  $x^1, \dots, x^k \in \mathbf{R}^n$  (here  $1, \dots, k$  are upper indices) be the  $x$ -components of the solutions of the system*

$$\frac{\partial L}{\partial x_r}(x_1, \dots, x_n, \lambda_1, \dots, \lambda_m) = 0, \quad r = 1, \dots, n;$$

$$h_i(x_1, \dots, x_n) = 0, \quad i = 1, \dots, m,$$

*where*

$$L(x_1, \dots, x_n, \lambda_1, \dots, \lambda_m) := f(x_1, \dots, x_n) + \sum_{i=1}^m \lambda_i h_i(x_1, \dots, x_n).$$

*Let  $j \in \{1, \dots, k\}$  be such that*

$$f(x^j) = \min_{1 \leq i \leq k} f(x^i).$$

*Then  $x^j$  is a global minimizer of  $f$  subject  $h(x) = 0$ .*

*Let  $s \in \{1, \dots, k\}$  be such that*

$$f(x^s) = \max_{1 \leq i \leq k} f(x^i).$$

*Then  $x^s$  is a global maximizer of  $f$  subject  $h(x) = 0$ .*

**PROOF.** The set  $\{x : h(x) = 0\}$  is bounded by (2) and is closed because  $h$  is continuous. Hence the function  $f$  has a global minimizer  $x^*$  subject  $h(x) = 0$ . According to the observation above  $x^*$  is the  $x$ -component of some solution of the system. Hence  $x^*$  is among  $x^1, \dots, x^k$ . Hence both  $f(x^j) \leq f(x^*)$  and  $f(x^j) \geq f(x^*)$ . Hence  $x^j$  is a global minimizer. The same argument work for maximizers. ■

**Example.** Find global minimizers and global maximizers of

$$f(x_1, x_2, x_3) = \frac{2}{3}x_1^3 + x_2 + 3x_3$$

subject to

$$h(x_1, x_2, x_3) = 1 - x_1^2 - x_2^2 - 3x_3^2 = 0.$$

*Solution.* We are going to use Corollary 2.1. Let us verify that its conditions are satisfied.

Condition (1). Observe that one vector is linearly independent if and only if it is nonzero. We have

$$Dh(x) = [-2x_1 \quad -2x_2 \quad -6x_3].$$

Hence  $Dh(x) = 0$  if and only if  $x = 0$ . But the vector  $x = 0$  does not satisfy  $h(x) = 0$ . Hence  $Dh(x) \neq 0$  for every  $x$  satisfying  $h(x) = 0$ .

Condition (2). Observe that  $h(x) = 1 - \|x\|^2 - 2x_3^2$ . Therefore, if  $h(x) = 0$ , then  $\|x\|^2 = 1 - 2x_3^2 \leq 1$ . Hence the set  $\{x : h(x) = 0\}$  is bounded.

The Lagrangian function for our problem is:

$$L(\lambda, x_1, x_2, x_3) = \frac{2}{3}x_1^3 + x_2 + 3x_3 + \lambda(1 - x_1^2 - x_2^2 - 3x_3^2)$$

We get the following system of equations:

$$\begin{aligned} 2x_1^2 + \lambda(-2x_1) &= 0 & x_1 &= \lambda \text{ or } x_1 = 0 \\ 1 + \lambda(-2x_2) &= 0 & x_2 &= \frac{1}{2\lambda} \\ 3 + \lambda(-6x_3) &= 0 & x_3 &= \frac{1}{2\lambda} \\ 1 - x_1^2 - x_2^2 - 3x_3^2 &= 0 \end{aligned}$$

In the case when  $x_1 = \lambda$  we get (from the 4th equation)

$$1 - \lambda^2 - \left(\frac{1}{2\lambda}\right)^2 - 3\left(\frac{1}{2\lambda}\right)^2 = 0 \quad \text{or} \quad 1 = \lambda^2 + \frac{1}{\lambda^2}$$

This equation does not have real roots. One of the possible proofs: If  $|\lambda| > 1$ , then  $\lambda^2 > 1$  if  $|\lambda| < 1$ , then  $\frac{1}{\lambda^2} > 1$  if  $|\lambda| = 1$ , then  $\lambda^2 + \frac{1}{\lambda^2} = 2$ .

In the case when  $x_1 = 0$  we get (from the 4th equation)

$$1 - 0^2 - \frac{1}{4\lambda^2} - \frac{3}{4\lambda^2} = 0$$

$$1 = \frac{1}{\lambda^2} \quad \lambda = \pm 1$$

We get two solutions  $\begin{bmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$  and  $\begin{bmatrix} 0 \\ -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}$ .

We have

$$f\left(\begin{bmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}\right) = 2 \quad f\left(\begin{bmatrix} 0 \\ -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}\right) = -2$$

**Answer:**

$\begin{bmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$  is global maximizer of  $f$  subject  $h(x) = 0$

and

$\begin{bmatrix} 0 \\ -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}$  is global minimizer of  $f$  subject  $h(x) = 0$ .

What if the condition (1) of Corollary 2.1 is not satisfied?

If the condition (1) is not satisfied at finitely many points we add these points to the collection  $x^1, \dots, x^k$  and continue in the same way as in the corollary. If there are infinitely many such points, additional investigation is necessary.

What if the condition (2) of Corollary 2.1 is not satisfied? Although there is no theory in general, there are some important cases in which we can use the same approach.

**Some notation.** Let  $f : \mathbf{R}^n \rightarrow \mathbf{R}$ . We write  $\lim_{\|x\| \rightarrow \infty} f(x) = \infty$  if for every real number  $M$  there exists a real number  $N$  satisfying the condition: if  $\|x\| > N$  then  $f(x) > M$ .

**Example.**  $\lim_{\|x\| \rightarrow \infty} e^{\|x\|} = \infty$ .

In fact, let  $M$  be a real number. If  $M \leq 0$  then the condition is satisfied for any choice of  $N$ , because  $e^{\|x\|} > 0$  for every  $x$ . If  $M > 0$ , we let  $N = \ln M$ . If  $\|x\| > N$  then (by the monotonicity of  $e^x$ ) we have  $e^{\|x\|} > e^N = e^{\ln M} = M$ .

**THEOREM 2.3** *Let  $f : \mathbf{R}^n \rightarrow \mathbf{R}$  and  $h : \mathbf{R}^n \rightarrow \mathbf{R}^m$  be continuously differentiable. Suppose that the derivatives*

$$Dh_1(x), Dh_2(x), \dots, Dh_m(x)$$

*are linearly independent for each  $x$  satisfying  $h(x) = 0$ , the set  $A = \{x : h(x) = 0\}$  is unbounded, and*

$$\lim_{\|x\| \rightarrow \infty} f(x) = \infty.$$

*Then the function  $f$  does not have global maximizers subject  $h(x) = 0$  and its global minimizers subject  $h(x) = 0$  can be found in the same way as in Corollary 2.1.*

**PROOF.** The statement concerning global maximizers can be derived from the definitions in the following way. Assume the contrary. Let  $x^*$  be a global maximizer of  $f$  subject  $h(x) = 0$ . Since  $\lim_{\|x\| \rightarrow \infty} f(x) = \infty$ , then there exists  $N$  such that  $f(x) > f(x^*)$  if  $\|x\| > N$  (we apply the definition with  $M = f(x^*)$ ). Since the set  $\{x : h(x) = 0\}$  is unbounded, we can find  $\tilde{x}$  such that  $\|\tilde{x}\| > N$  and  $h(\tilde{x}) = 0$ . Hence there exists  $\tilde{x}$  satisfying  $h(\tilde{x}) = 0$  and  $f(\tilde{x}) > f(x^*)$ . This contradicts the assumption that  $x^*$  is a global maximizer.

**Minimizers.** Let  $u \in \mathbf{R}^n$  be any point satisfying  $h(u) = 0$ . Since  $\lim_{\|x\| \rightarrow \infty} f(x) = \infty$ , then there exists a real number  $N$  such that  $f(x) > f(u)$  when  $\|x\| > N$ . Consider the set

$$\tilde{A} = \{x : h(x) = 0 \text{ and } \|x\| \leq N\}.$$

The set  $\tilde{A}$  is bounded and closed. The boundedness is obvious. Let us sketch the proof of the fact that  $\tilde{A}$  is closed. We need to show that  $\{x : x \notin \tilde{A}\}$  is open. Let  $x \notin \tilde{A}$ . Then either  $\|x\| > N$  or  $h(x) \neq 0$ . In the first case we let  $\varepsilon = \|x\| - N$ . In the same way as in our first example of an open set we show that, if  $\|y - x\| < \varepsilon$ , then  $y \notin \tilde{A}$ . In the second case we let  $\delta = \|h(x)\|$  and repeat the argument of Proposition 2.1.

By the Weierstrass Theorem there exists at least one global minimizer of  $f$  over  $\tilde{A}$ . We denote one of the global minimizers by  $x^*$ . To finish the proof it is enough to show that  $x^*$  is a global minimizer of  $f$  over  $A$ , that is to show that for every  $x$  satisfying  $h(x) = 0$  we have  $f(x) \geq f(x^*)$ . There are two possibilities for  $x$ : either  $\|x\| \leq N$  or  $\|x\| > N$ . In the first case  $x \in \tilde{A}$  and  $f(x) \geq f(x^*)$  by the definition of a global minimizer over  $\tilde{A}$ . In the second case  $f(x) > f(u)$  (by the choice of  $N$ ). On the other hand, since  $u \in \tilde{A}$  (it also follows from the choice of  $N$ ), then  $f(u) \geq f(x^*)$ . We get  $f(x) > f(u) \geq f(x^*)$ . ■

**Observation.** We may use this theorem in unconstrained case.

Here is the corresponding version of the theorem.

**THEOREM 2.4** *Let  $f : \mathbf{R}^n \rightarrow \mathbf{R}$  be continuously differentiable and  $\lim_{\|x\| \rightarrow \infty} f(x) = \infty$ . Then*

- $f$  does not have global maximizers;
- Let  $x^1, \dots, x^k$  be the critical points of  $f$ . Let  $j$ ,  $1 \leq j \leq k$  be such that  $f(x^j) = \min_{1 \leq l \leq k} f(x^l)$ . Then  $x^j$  is a global minimizer of  $f$  over  $\mathbf{R}^n$ .

**Some more notation.** Let  $f : \mathbf{R}^n \rightarrow \mathbf{R}$ . We write  $\lim_{\|x\| \rightarrow \infty} f(x) = -\infty$  if for every real number  $M$  there exists a real number  $N$  satisfying the condition: if  $\|x\| > N$  then  $f(x) < M$ .

**Remark.** There exist natural analogues of Theorems 2.3 and 2.4 for functions satisfying  $\lim_{\|x\| \rightarrow \infty} f(x) = -\infty$ .

In many cases it is difficult to show that  $\lim_{\|x\| \rightarrow \infty} f(x) = \infty$  using the definition. In order to state a theorem that is very useful in this context we need to recall some definitions and facts from linear algebra.

**DEFINITION 2.4** A real-valued linear function on  $\mathbf{R}^n$  is called a *linear functional*. Linear functionals on  $\mathbf{R}^n$  form a *linear space*, it means that the following two conditions are satisfied: (a) the sum of two linear functionals is a linear functional; (b) a scalar multiple of a linear functional is a linear functional. Linear functionals  $q_1, q_2, \dots, q_k$  on  $\mathbf{R}^n$  are called *linearly independent* if  $a_1 q_1 + \dots + a_k q_k = 0$  implies  $a_1 = \dots = a_k = 0$ .

**Fact.** A function  $q : \mathbf{R}^n \rightarrow \mathbf{R}$  is a linear functional on  $\mathbf{R}^n$  if and only if there exist real numbers  $\lambda_1, \dots, \lambda_n$  such that

$$q \left( \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \right) = \lambda_1 x_1 + \dots + \lambda_n x_n.$$

The numbers  $\lambda_1, \dots, \lambda_n$  are uniquely determined by  $q$ .

**Examples.** The functions  $q_1(x_1, x_2, x_3) = x_1 + x_2 + 3x_3$  and  $q_2(x_1, x_2, x_3) = 10x_3$  are linear functionals. The function  $q_3(x_1, x_2, x_3) = x_1 + x_2^2 + x_3^3$  is not a linear functional.

**Fact.** Linear functionals  $q_1, \dots, q_n$  on  $\mathbf{R}^n$  are linearly independent if and only if

$$\det \begin{bmatrix} \lambda_{1,1} & \dots & \lambda_{1,n} \\ \vdots & \ddots & \vdots \\ \lambda_{n,1} & \dots & \lambda_{n,n} \end{bmatrix} \neq 0,$$

where  $\lambda_{k,1}, \dots, \lambda_{k,n}$  are the real numbers corresponding to the functional  $q_k$ .

**Example.** The functionals  $x_1 + x_2$ ,  $x_2 + x_3$  and  $x_1 + x_3$  are linearly independent functionals of  $\mathbf{R}^3$ .

**THEOREM 2.5** (a) *Let  $q_1, q_2, \dots, q_n$  be linearly independent linear functionals on  $\mathbf{R}^n$  and  $P_i : \mathbf{R} \rightarrow \mathbf{R}$   $i = 1, \dots, n$  be such that*

$$\lim_{z \rightarrow \infty} P_i(z) = \infty \quad \text{and} \quad \lim_{z \rightarrow -\infty} P_i(z) = \infty.$$

*Then*

$$\lim_{\|x\| \rightarrow \infty} \sum_{i=1}^n P_i(q_i(x)) = \infty.$$

(b) *Suppose that  $\lim_{\|x\| \rightarrow \infty} f(x) = \infty$  and that  $g(x)$  is such that  $g(x) \geq C$  for some real number  $C$  and every  $x \in \mathbf{R}^n$ . Then  $\lim_{\|x\| \rightarrow \infty} (f(x) + g(x)) = \infty$ .*

(c) *Suppose that  $\lim_{\|x\| \rightarrow \infty} f(x) = \infty$  and that  $h : \mathbf{R} \rightarrow \mathbf{R}$  satisfies  $\lim_{z \rightarrow \infty} h(z) = \infty$ . Then*

$$\lim_{\|x\| \rightarrow \infty} h(f(x)) = \infty.$$

WITHOUT PROOF. ■

**Remark 1.** It is worthwhile to emphasize that the number of functionals in this theorem coincide with the dimension of the space.

**Remark 2.** In particular, the theorem can be used when  $P_i$  are polynomials with even degrees and positive coefficients near the highest powers.

**Remark 3.** There exists a similar result concerning functions satisfying  $\lim_{\|x\| \rightarrow \infty} f(x) = -\infty$ .

**Example.** Let

$$f(x) = \frac{1}{2}(x_1 + x_2)^4 - (x_1 + x_2)^2 + x_2^2 + x_3^4.$$

Show that

$$\lim_{\|x\| \rightarrow \infty} f(x) = \infty.$$

Use this result to find global minimizers and global maximizers of  $f$  over  $\mathbf{R}^n$ .

*Solution.*  $f(x)$  can be written as

$$f(x) = \sum_{i=1}^3 P_i(q_i(x))$$

where

$$\left. \begin{array}{l} q_1(x) = x_1 + x_2 \\ q_2(x) = x_2 \\ q_3(x) = x_3 \end{array} \right\} \text{ are linearly independent}$$

$$\left. \begin{array}{l} P_1(z) = \frac{1}{2}z^4 - z^2 \\ P_2(z) = z^2 \\ P_3(z) = z^4 \end{array} \right\} \begin{array}{l} \text{leading coefficients are positive} \\ \text{and degrees are even} \end{array}$$

Hence,

$$\lim_{\|x\| \rightarrow \infty} f(x) = \infty.$$

Hence  $f$  does not have global maximizers.

The equation  $Df(x) = 0$  is equivalent to the system

$$\begin{aligned} 2(x_1 + x_2)^3 - 2(x_1 + x_2) &= 0 \\ 2(x_1 + x_2)^3 - 2(x_1 + x_2) + 2x_2 &= 0 \\ 4x_3^3 &= 0 \quad \text{equivalent to } x_3 = 0. \end{aligned}$$

Subtracting the first equation from the second we get  $2x_2 = 0$ , so  $x_2 = 0$ . Using this we get from the first equation  $2x_1^3 - 2x_1 = 0$

Three solutions  $x_1 = 1, x_1 = -1, x_1 = 0$ .

Three critical points  $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix},$  &  $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$

We have

$$f\left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right) = -\frac{1}{2} \quad f\left(\begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}\right) = -\frac{1}{2} \quad f\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}\right) = 0.$$

Hence

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}$$

are global minimizers of  $f$  over  $\mathbf{R}^n$ .

**Exercises. 1.** Find global minimizers and global maximizers of

$$f(x_1, x_2, x_3) = 2(x_1 + x_2)^4 - (x_1 + x_2) + x_2^4 + x_3^4.$$

**2.** Let  $f(x) = x_1^8 - 2x_1x_3 - 4x_1x_2 + x_2^4 + 5x_3^2$ . Show that

$$\lim_{x \rightarrow \infty} f(x) = \infty.$$

### 3 Linear Programming

Let  $f, h_1, \dots, h_k$  be linear functionals on  $\mathbf{R}^n$  and  $b_1, \dots, b_k$  be real numbers. The problem

$$\begin{aligned} & \text{minimize (maximize) } f(x) \\ & \text{subject } h_i(x) \leq b_i, \quad i = 1, \dots, k \end{aligned}$$

is called a *linear programming problem* or *LP problem*.

**Example.** Maximize  $10x_1 + 11x_2 + 12x_3$  subject

$$\begin{aligned} x_1 + x_2 &\leq 1; \\ x_1 + x_3 &\leq 2; \\ 11x_1 + 12x_2 + 13x_3 &\leq 100. \end{aligned}$$

It turns out that each LP problem is equivalent to a problem of the form

$$\begin{aligned} & \text{minimize } c^T x \\ & \text{subject } Ax = b \text{ and } x \geq 0, \end{aligned}$$

where  $c \in \mathbf{R}^n$ ,  $A$  is an  $m \times n$  matrix ( $m < n$ ) of rank  $m$ ,  $b \in \mathbf{R}^m$  and  $x \geq 0$  means that each coordinate of  $x$  is non-negative. ( $A$  has *rank*  $m$  means that  $n \geq m$  and that  $A$  has  $m$  linearly independent columns.)

This form of a linear programming problem is called *standard*.

LP problems arise in many different contexts. (See [3], [2] (Section 15.2) or any textbook on linear programming.) We shall discuss only one example.

#### Transportation Problem

Some commodity is made at  $m$  plants  $P_1, \dots, P_m$ . Let  $\sigma_i$  be the supply (the amount of the commodity that can be made) at the plant  $P_i$ . The commodity is sold at  $n$  markets  $M_1, \dots, M_n$ . Let  $\delta_j$  be the demand (the

expected amount of the commodity that will be sold) at the market  $M_j$ . Let  $a_{i,j}$  be the cost of transporting of one unit of the commodity from the plant  $P_i$  to the market  $M_j$ . Suppose that the total supply ( $\sum_{i=1}^m \sigma_i$ ) is equal to the total demand ( $\sum_{j=1}^n \delta_j$ ).

**Problem.** Determine the amounts that should be transported from each plant to each market in order to minimize total transportation cost.

To see that this problem is an LP problem denote by  $x_{i,j}$  the amount of the commodity that we are going to transport from  $P_i$  to  $M_j$ .

We have

$$\sum_{i=1}^m x_{i,j} = \delta_j; \tag{1}$$

$$\sum_{j=1}^n x_{i,j} = \sigma_i; \tag{2}$$

$$x_{i,j} \geq 0. \tag{3}$$

Total transportation cost is  $\sum_{i=1}^m \sum_{j=1}^n a_{i,j} x_{i,j}$ .

So in this case the space is  $m \times n$  dimensional, and the mathematical statement of the problem is:

Minimize the linear functional (on  $\mathbf{R}^{m \times n}$ )

$$f(x) = \sum_{i=1, j=1}^{m, n} a_{i,j} x_{i,j}$$

subject to constraints (1), (2) and (3).

It is clear that this problem satisfies all the conditions on standard-form LP problems except the condition on the rank. (We shall come back to this condition later.)

Now we shall study standard-form problems.

**Observation.** Since  $A$  is of rank  $m$ , then the equation  $Ax = b$  has at least one solution. (Can be proved using some basic Linear Algebra.)

**DEFINITION 3.1** A solution  $x$  of  $Ax = b$  is called *feasible* if  $x \geq 0$ .

**DEFINITION 3.2** A vector  $x$  that minimize  $c^T x$  over the set of vectors satisfying  $Ax = b$  and  $x \geq 0$  is called an *optimal feasible solution*.

Our purpose is to find optimal feasible solutions (or to show that they do not exist).

**Example.** Minimize  $2x_1 + 4x_2 + x_3 + x_4$  subject to

$$x_1 - x_2 = 3;$$

$$x_3 - x_4 = 5;$$

$$x \geq 0.$$

It is a standard-form LP problem with

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix};$$

$$b = \begin{bmatrix} 3 \\ 5 \end{bmatrix}, \quad c = \begin{bmatrix} 2 \\ 4 \\ 1 \\ 1 \end{bmatrix}.$$

The matrix  $A$  has rank 2 because (for example) the first and the third columns are linearly independent.

The vector

$$x^1 = \begin{bmatrix} 3 \\ 0 \\ 5 \\ 0 \end{bmatrix}$$

is a feasible solution of  $Ax = b$ . The vector

$$x^2 = \begin{bmatrix} 0 \\ -3 \\ 0 \\ -5 \end{bmatrix}$$

is an unfeasible solution of  $Ax = b$ .

**THEOREM 3.1** *Let  $x \in \mathbf{R}^n$  be an optimal feasible solution of a standard-form LP problem with the minimal possible number of nonzero coordinates. Let  $x_{i_1}, \dots, x_{i_k}$  be the non-zero coordinates of  $x$ . Let  $a_1, \dots, a_n$  be columns of  $A$ . Then  $a_{i_1}, \dots, a_{i_k}$  are linearly independent.*

PROOF. Assume the contrary. Then there exist  $\{\lambda_j\}_{j=1}^k$ , not all zero, such that

$$\sum_{j=1}^k \lambda_j a_{i_j} = 0. \quad (*)$$

There are two cases: either at least one of the numbers

$$\lambda_1, \dots, \lambda_k$$

is positive or all of them are non-positive. We consider only the first case because in the second case we can replace  $\lambda_1, \dots, \lambda_k$  by  $-\lambda_1, \dots, -\lambda_k$ . Let  $\varepsilon := \min\{\frac{x_{i_j}}{\lambda_j} : \text{over all } j \text{ satisfying } \lambda_j > 0\}$ . Since  $x_{i_1}, \dots, x_{i_k}$  are the non-zero coordinates of  $x$  and  $x \geq 0$ , then  $\varepsilon > 0$ .

The definition of  $\varepsilon$  implies that

$$(1) \quad x_{i_j} - \varepsilon \lambda_j \geq 0.$$

In fact, for  $j$  satisfying  $\lambda_j \leq 0$  the assertion follows from  $\varepsilon > 0$  and  $x_{i_j} \geq 0$ . For  $j$  satisfying  $\lambda_j > 0$  we have  $\varepsilon \leq x_{i_j}/\lambda_j$  (by the definition of  $\varepsilon$ ), hence  $x_{i_j} - \varepsilon \lambda_j \geq 0$ .

$$(2) \quad \text{At least one of the numbers } \{x_{i_j} - \varepsilon \lambda_j\}_{j=1}^k \text{ is equal to 0.}$$

In fact, it is the case for  $j$  satisfying  $\varepsilon = \frac{x_{i_j}}{\lambda_j}$ .

We introduce the vector  $y \in \mathbf{R}^n$  by

$$y_{i_j} = \lambda_j \text{ for } j = 1, \dots, k.$$

$$y_l = 0 \text{ if } l \notin \{i_1, \dots, i_k\}.$$

and let  $z = x - \varepsilon y$ . Then

$$z_{i_j} = x_{i_j} - \varepsilon \lambda_j \text{ for } j = 1, \dots, k.$$

$$z_l = 0 \text{ if } l \notin \{i_1, \dots, i_k\}.$$

From (1) we get  $z \geq 0$ . From (2) we get: the number of non-zero coordinates of  $z$  is strictly less than the number of nonzero coordinates of  $x$ .

Also, the equation (\*) can be rewritten as  $Ay = 0$ . Hence  $Az = A(x - \varepsilon y) = Ax - \varepsilon Ay = Ax = b$ . Hence  $z$  is a feasible solution with less nonzero components than  $x$ .

To get a contradiction it remains to show that  $z$  is an optimal feasible solution, that is  $c^T z = c^T x$ ,

We have  $z = x - \varepsilon y$  and  $c^T z = c^T x - \varepsilon c^T y$ . So it is enough to show that  $c^T y = 0$ .

Let  $\delta = \min\{\frac{x_{i_j}}{|y_{i_j}|} : y_{i_j} \neq 0\}$ . It is clear that  $\delta > 0$ . We have also  $(x + \delta y) \geq 0$  and  $(x - \delta y) \geq 0$ . In fact, for  $k$  satisfying  $y_k = 0$  it follows from  $x_k \geq 0$  and for  $k$  satisfying  $y_k \neq 0$  we have (by the definition of  $\delta$ )  $x_k \geq \delta |y_k|$ . Hence  $x_k - \delta y_k \geq 0$  and  $x_k + \delta y_k \geq 0$ .

Since  $Ay = 0$ , then  $A(x + \delta y) = A(x - \delta y) = 0$ . Therefore the vectors  $(x + \delta y)$  and  $(x - \delta y)$  are feasible.

Since the vector  $x$  is optimal, it implies  $c^T x \leq c^T(x + \delta y)$  and  $c^T x \leq c^T(x - \delta y)$ . It implies  $0 \leq \delta c^T y$  and  $0 \leq -\delta c^T y$ . Hence  $\delta c^T y = 0$ . Since  $\delta$  is non-zero, it implies  $c^T y = 0$ .

We get:  $z$  is an optimal feasible solution and the number of non-zero coordinates of  $z$  is less than  $k$ . It contradicts the choice of  $x$ . This contradiction shows that the vectors  $a_{i_1}, \dots, a_{i_k}$  are linearly independent. ■

Theorem 3.1 can be used to construct an algorithm for finding optimal feasible solutions.

**THEOREM 3.2** (Lin. Alg.) *Let  $A$  be a matrix of rank  $m$  and  $a_{i_1}, \dots, a_{i_k}$  be linearly independent columns of  $A$ . There exists an invertible  $m \times m$  submatrix  $B$  of  $A$  such that  $a_{i_1}, \dots, a_{i_k}$  are among columns of  $B$ .*

Theorems 3.1 and 3.2 can be used to derive the following

**COROLLARY 3.1** If the problem has optimal feasible solutions, then some of them can be found in the following way:

Choose all invertible  $(m \times m)$  submatrices  $B$  in  $A$ . Solve all the systems  $Bz = b$  that we get. Remove all solutions that have negative coordinates. For each of the remaining solutions consider the corresponding vector in  $\mathbf{R}^n$ . (The correspondence is described below.) Compute the values of  $c^T x$  for all obtained vectors. The vectors  $x$  for which the corresponding value is minimal are optimal feasible solutions.

What do we mean by the corresponding vector in  $\mathbf{R}^n$ ? Let  $a_1, \dots, a_n$  be the columns of  $A$ , and let  $a_{i_1}, \dots, a_{i_m}$  be the columns of  $B$ . Let

$$z = \begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix}$$

be a solution of  $Bz = b$ . Then the vector  $x \in \mathbf{R}^n$  corresponding to  $z$  is defined by  $x_{i_j} = z_j$  for  $j = 1, \dots, m$  and  $x_l = 0$  if  $l \notin \{i_1, \dots, i_m\}$ .

It is convenient to use the following definition.

**DEFINITION 3.3** A solution of  $Ax = b$  corresponding to an invertible  $m \times m$  submatrix  $B$  of  $A$  is called a *basic* solution.

Now we can describe the algorithm in the following way:

**ALGORITHM.** We find all basic feasible solutions of  $Ax = b$  and evaluate the function  $c^T x$  at them. If the problem has optimal feasible solutions, then any of the basic feasible solutions minimizing  $c^T x$  is one of them.

Why is it true?

**PROOF.** If the problem has optimal feasible solutions, we can choose one of them, say  $x$ , with the minimal possible number of non-zero coordinates. By Theorem 3.1 the columns  $a_{i_1}, \dots, a_{i_k}$  corresponding to non-zero coordinates of  $x$  are linearly independent. By Theorem 3.2 there exists an invertible

$m \times m$  submatrix  $B$  of  $A$  containing the columns  $a_{i_1}, \dots, a_{i_k}$ . But then  $x$  is the basic solution corresponding to  $B$ . (To see this we need to recall the fact that  $Bz = b$  has a unique solution, the definitions of matrix multiplication and of a solution corresponding to  $B$ .) ■

This algorithm has the following drawbacks:

(1) If optimal solutions do not exist, the algorithm gives a vector that (of course) is not an optimal solution.

(2) Working with this algorithm we have to consider all  $m \times m$  submatrices of  $A$  (and it may happen that all of them are invertible). From Combinatorics we know that an  $m \times n$  matrix has  $\frac{n!}{m!(n-m)!}$   $m \times m$  submatrices. In applications, for example, to transportation problems the numbers  $m$  and  $n$  are large. In such situations it is unpractical and sometimes even impossible to use this method.

Efficient methods for solving linear programming problems were developed by L. Kantorovich, T. Koopmans and G. Dantzig. The most popular is the simplex method designed by G. Dantzig.

But if  $m$  and  $n$  are (very) small numbers the algorithm is efficient for solving standard-form linear programming problems provided we know that optimal feasible solutions exist.

One of the cases when optimal feasible solutions exist is described by the following theorem.

**THEOREM 3.3** *Suppose that all entries of  $A$  and  $b$  are nonnegative and each column of  $A$  has at least one strictly positive entry. If the system  $Ax = b$  has feasible solutions, then it has optimal feasible solutions.*

**PROOF.** Let  $\Omega$  be the set of all feasible solutions. That is  $\Omega = \{x : Ax = b \text{ and } x \geq 0\}$ . Since the system  $Ax = b$  has feasible solutions, the set  $\Omega$  is non-empty.

**Claim 1.**  $\Omega$  is closed.

PROOF. The complement of  $\Omega$  consists of points  $x$  such that either  $Ax \neq b$  or at least one of the coordinates of  $x$  is  $< 0$ .

For  $x$  satisfying the first condition we do the following. We let  $\delta = \|Ax - b\| > 0$ . It is well known that  $Ax - b$  is a continuous function of  $x$ . Hence there exists  $\varepsilon > 0$  such that if  $\|x - y\| < \varepsilon$ , then  $\|(Ax - b) - (Ay - b)\| < \delta$  and

$$\|Ay - b\| \geq \|Ax - b\| - \|(Ax - b) - (Ay - b)\| > \delta - \delta = 0.$$

So  $\|Ay - b\| > 0$ , therefore  $Ay \neq b$  and  $y$  is in the complement of  $\Omega$ .

For  $x$  satisfying the second condition we choose  $k \in \{1, \dots, n\}$  such that  $x_k < 0$ . We let  $\varepsilon = -x_k > 0$ . If  $\|x - y\| < \varepsilon$ , then

$$y_k = x_k + (y_k - x_k) \leq x_k + \|x - y\| < x_k + \varepsilon = -\varepsilon + \varepsilon = 0.$$

So  $y_k < 0$  and  $y$  is also in the complement of  $\Omega$ . Hence the complement of  $\Omega$  is open, and  $\Omega$  is closed. ■

**Claim 2.**  $\Omega$  is bounded.

PROOF. Let  $x \in \Omega$ . Then  $x \geq 0$  and  $Ax = b$ . The last equation can be written in the form.

$$\sum_{i=1}^n a_{k,i} x_i = b_k \quad (k = 1, \dots, m). \quad (*)$$

By the condition of the theorem for each

$$j \in \{1, \dots, n\}$$

there exists  $k(j)$  such that  $a_{k(j),j}$  is strictly positive, that is  $a_{k(j),j} > 0$ . We rewrite the corresponding equality as

$$a_{k(j),j} x_j + \sum_{i=1, i \neq j}^n a_{k(j),i} x_i = b_{k(j)}.$$

Since all the numbers  $x_i$  and  $a_{k,i}$  are nonnegative, we get

$$a_{k(j),j} x_j \leq b_{k(j)}.$$

Since  $a_{k(j),j} > 0$ , we get

$$x_j \leq \frac{b_{k(j)}}{a_{k(j),j}}.$$

Therefore

$$\|x\|^2 = \sum_{j=1}^n x_j^2 \leq \sum_{j=1}^n \left( \frac{b_{k(j)}}{a_{k(j),j}} \right)^2.$$

The number in the right-hand side does not depend on  $x$ . Hence the set  $\Omega$  is bounded. ■

It is well known that  $c^T x$  is continuous (as a function of  $x$ ).

By the Weierstrass theorem it follows that there exist minimizers (and maximizers) of  $c^T x$  over  $\Omega$  (=subject  $Ax = b$  and  $x \geq 0$ ). ■

**COROLLARY 3.2** *The algorithm can be used to find optimal feasible solutions of problems satisfying the conditions of Theorem 3.3.*

**PROOF.** Actually it remains to understand what happens if the problem does not have feasible solutions. In such a case we use the following version of Theorem 3.1 (the same proof works):

**THEOREM 3.4** *Let  $x \in \mathbf{R}^n$  be a feasible solution of a standard-form LP problem with the minimal possible number of nonzero coordinates. Let  $x_{i_1}, \dots, x_{i_k}$  be the non-zero coordinates of  $x$ . Let  $a_1, \dots, a_n$  be columns of  $A$ . Then  $a_{i_1}, \dots, a_{i_k}$  are linearly independent.*

In other words: if the problem has feasible solutions, it has basic feasible solutions. We may state it differently: if all basic solutions are unfeasible, then the problem does not have feasible solutions. So the algorithm solves the problem in this case also. (That is: if all basic solutions are unfeasible, then none of the vectors in  $\mathbf{R}^n$  satisfy the constraints.) ■

**Example.** Use the described algorithm to solve the following problem:

Minimize  $f(x) = 2x_1 - 3x_2 - 4x_3 - 3x_4$ , subject to  $x \geq 0$  and  $Ax = b$  where  $x \in \mathbf{R}^4$  and

$$A = \begin{bmatrix} 1 & 1 & 2 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

It is clear that  $A$  and  $b$  satisfy the conditions of Theorem 3.3. So we may use the algorithm.

We list all  $2 \times 2$  submatrices of  $A$ :

$$B_1 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, B_2 = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, B_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$B_4 = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}, B_5 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, B_6 = \begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix}.$$

It turns out that in this example all of them are invertible.

The corresponding solutions are:

$$x^1 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, x^2 = \begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, x^3 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

$$x^4 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, x^5 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, x^6 = \begin{bmatrix} 0 \\ 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix},$$

Observe that  $x^2$  is not feasible. For the remaining solutions we have  $f(x^1) = -3$ ,  $f(x^3) = -1$ ,  $f(x^4) = f(x^5) = -3$  and  $f(x^6) = -3.5$ . Hence  $x^6$  is an optimal feasible solution.

Now we shall discuss the following question: how to reduce an LP problem to a standard-form LP problem?

1. Any LP problem can be restated as

$$\begin{aligned} & \text{minimize (maximize)} \quad d^T x \\ & \text{subject} \quad Hx \leq b, \end{aligned}$$

where  $d \in \mathbf{R}^k$ ,  $H$  is an  $m \times k$  matrix and  $b \in \mathbf{R}^m$ .

2. It is clear that the problem

$$\begin{aligned} & \text{maximize } d^T x \\ & \text{subject } Hx \leq b, \end{aligned}$$

is equivalent to

$$\begin{aligned} & \text{minimize } (-d)^T x \\ & \text{subject } Hx \leq b, \end{aligned}$$

3. So it is enough to consider a problem of the form:

$$\begin{aligned} & \text{minimize } d^T x \\ & \text{subject } Hx \leq b, \end{aligned}$$

Let

$$H = \begin{bmatrix} h_{11} & \dots & h_{1k} \\ \vdots & \ddots & \vdots \\ h_{m1} & \dots & h_{mk} \end{bmatrix}.$$

We introduce  $2k + m$  parameters  $\{z_i\}_{i=1}^{2k+m}$  in the following way:

$$\begin{aligned} z_j &= \frac{|x_j| + x_j}{2} \text{ if } j = 1, \dots, k; \\ z_j &= \frac{|x_j| - x_j}{2} \text{ if } j = k + 1, \dots, 2k; \\ z_{2k+i} &= b_i - \sum_{j=1}^k h_{ij} x_j, \quad i = 1, \dots, m. \end{aligned}$$

It is clear that if  $x$  satisfies  $Hx \leq b$ , then  $\{z_i\}_{i=1}^{2k+m}$  satisfy the following conditions

$$\begin{aligned} z_i &\geq 0 \text{ for every } i; \\ x_j &= z_j - z_{k+j} \text{ for } j = 1, \dots, k \end{aligned}$$

and therefore

$$\sum_{j=1}^k h_{i,j} (z_j - z_{k+j}) + z_{2k+i} = b_i; \quad i = 1, \dots, m.$$

The last equation can be rewritten as  $Az = b$  where  $A$  is a  $m \times (2k + m)$ -matrix given by

$$A = \begin{bmatrix} h_{11} & \dots & h_{1k} & -h_{11} & \dots & -h_{1k} & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h_{m1} & \dots & h_{mk} & -h_{m1} & \dots & -h_{mk} & 0 & \dots & 1 \end{bmatrix}.$$

Let  $c \in \mathbf{R}^{2k+m}$  be given by

$$c = \begin{bmatrix} d_1 \\ \vdots \\ d_k \\ -d_1 \\ \vdots \\ -d_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

**Claim.** The problem

$$\begin{aligned} & \text{minimize } c^T z \\ & \text{subject } Az = b \text{ and } z \geq 0 \end{aligned}$$

is equivalent to the initial problem.

**PROOF.** It is clear that with such definition of  $z$  there is a one-to-one correspondence between the feasible sets of the problems. Also we have  $c^T z = d^T x$ . Hence the problems are equivalent. ■

**Remark.** It is also clear that the matrix  $A$  has rank  $m$ . Hence the problem

$$\begin{aligned} & \text{minimize } c^T z \\ & \text{subject } Az = b \text{ and } z \geq 0 \end{aligned}$$

is a standard-form LP problem.

**Example.** Find a standard-form LP problem equivalent to the problem

$$\begin{aligned} & \text{minimize } d^T x \\ & \text{subject } Hx \leq b, \end{aligned}$$

where  $d^T = [1 \ 2 \ 3]$ ,

$$H = \begin{bmatrix} 1 & 1 & 1 \\ -1 & -1 & -1 \\ 2 & 1 & 4 \\ 3 & 1 & 3 \\ -1 & -2 & -5 \end{bmatrix}; \quad b = \begin{bmatrix} 5 \\ -1 \\ 10 \\ 10 \\ -3 \end{bmatrix};$$

**Answer.** It is the problem:

$$\begin{aligned} & \text{minimize } c^T z \\ & \text{subject } Az = b \text{ and } z \geq 0, \end{aligned}$$

where  $c^T = [1 \ 2 \ 3 \ -1 \ -2 \ -3 \ 0 \ 0 \ 0 \ 0 \ 0]$ ,

$$A = \begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 2 & 1 & 4 & -2 & -1 & -4 & 0 & 0 & 1 & 0 & 0 \\ 3 & 1 & 3 & -3 & -1 & -3 & 0 & 0 & 0 & 1 & 0 \\ -1 & -2 & -5 & 1 & 2 & 5 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and  $b$  is the same as in the original problem.

**Remark on problems with equality constraints.** Suppose that we have a problem

$$\begin{aligned} & \text{minimize } c^T x \\ & \text{subject } Ax \leq b \text{ and } Bx = d, \end{aligned}$$

where  $c, x \in \mathbf{R}^n$ ,  $A$  is an  $m \times n$ -matrix,  $B$  is a  $k \times n$ -matrix,  $b \in \mathbf{R}^m$  and  $d \in \mathbf{R}^k$

Such problem can be reduced to an LP problem. It is enough to observe that  $Bx = d$  is equivalent to  $Bx \leq d$  and  $(-B)x \leq -d$ . Therefore the constraint can be written in the form  $Ex \leq f$ , where

$$E = \begin{bmatrix} A \\ B \\ -B \end{bmatrix} \quad \text{and} \quad f = \begin{bmatrix} b \\ d \\ -d \end{bmatrix}$$

are partitioned matrices.

### Exercises.

1. One of the important results about determinants is the following identity due to Vandermonde (see [10] (p. 3)):

Let  $a_1, a_2, \dots, a_n$  be real numbers. Then

$$\det \begin{bmatrix} 1 & 1 & \dots & 1 \\ a_1 & a_2 & \dots & a_n \\ a_1^2 & a_2^2 & \dots & a_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^{n-1} & a_2^{n-1} & \dots & a_n^{n-1} \end{bmatrix} = \prod_{i>j} (a_i - a_j).$$

(The right-hand side is the product of all differences between  $a_i$  and  $a_j$  ( $i > j$ ).)

Use this identity to show that for any  $m$  and any  $n \geq m$  there exists an  $m \times n$  matrix all of whose  $m \times m$  submatrices are invertible.

2. Use Corollary 3.2 to solve the following problem:

Minimize  $f(x) = -x_1 - 2x_2 - 3x_3 - 2x_4$ , subject to  $x \geq 0$  and  $Ax = b$  where  $x \in \mathbf{R}^4$  and

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

## 4 The simplex method

As I mentioned before it is unpractical to use the method based on solving the systems  $Bz = b$  for all invertible submatrices. Our next purpose is to discuss a more efficient method for solving LP problems, the simplex method. This method was designed by G. Dantzig. Although it is not efficient for some artificial LP problems, in practice and on average, the method is very efficient.

To describe the main ideas of the method we need some definitions.

**DEFINITION 4.1** A set  $P$  of points in  $\mathbf{R}^n$  is called a *polyhedron* if  $P = \{x : Ax \leq b\}$  for some  $m \times n$  matrix  $A$  and for some  $b \in \mathbf{R}^m$ .

**Examples.** 1.  $n$ -dimensional cube:

$$C_n = \left\{ \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} : 0 \leq x_1 \leq 1, \dots, 0 \leq x_n \leq 1 \right\}.$$

2. Half-plane:

$$P = \left\{ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} : x_1 \geq 0 \right\}.$$

3. Angular region:

$$B = \left\{ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} : x_1 \leq x_2, x_2 \leq 2x_1 \right\}.$$

**Remark.** All polyhedra are closed (the proof is standard). Polyhedra can be bounded or unbounded. (In particular,  $C_n$  is a bounded polyhedron for every  $n$ . Polyhedra  $P$  and  $B$  are unbounded.)

**DEFINITION 4.2** Let  $P$  be a polyhedron. A point  $x \in P$  is called a *vertex* of  $P$  if there are no two distinct points  $u$  and  $v$  in  $P$  such that  $x = \alpha u + (1 - \alpha)v$  for some  $0 < \alpha < 1$ .

**Examples.** Vertices of the cube  $C_n$ : a point  $x$  in  $\mathbf{R}^n$  is a vertex of  $C_n$  if and only if each of the components of  $x$  is either 0 or 1. Half-plane does not have vertices. The origin is the only vertex of the angular region  $B$ .

I shall prove the statements about  $P$  and  $B$  only.

Vertices of  $P$ . Let  $x$  be any point in  $P$ . Then

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 - 1 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 + 1 \end{bmatrix}.$$

Since

$$\begin{bmatrix} x_1 \\ x_2 - 1 \end{bmatrix} \text{ and } \begin{bmatrix} x_1 \\ x_2 + 1 \end{bmatrix} \text{ are in } P,$$

$x$  is not a vertex of  $P$ .

Vertices of  $B$ . Observe that the inequalities in the definition of  $B$  imply  $x_1 \geq 0$  and  $x_2 \geq 0$  for every  $x \in B$ . Therefore, if  $0 = \alpha u + (1 - \alpha)v$ ,  $0 < \alpha < 1$ , we get  $v_1 = u_1 = 0$  and  $v_2 = u_2 = 0$ . It proves that 0 is a vertex. Let us show that it is the only vertex. Let  $x \neq 0$  be a point in  $B$ . It is clear that  $2x \in B$  and that  $x = \frac{1}{2}0 + \frac{1}{2}(2x)$ . Hence  $x$  is not a vertex of  $B$ .

Why we are interested in vertices of polyhedra?

Consider an LP problem

$$\begin{aligned} & \text{minimize } d^T x \\ & \text{subject } Hx \leq b. \end{aligned}$$

Observe that the set  $\{x : Hx \leq b\}$  is a polyhedron.

**THEOREM 4.1** *If a polyhedron  $P$  has vertices and a linear function  $d^T x$  is bounded from below on  $P$  (that is  $d^T x \geq C$  for some real number  $C$  and for every  $x \in P$ ), then  $d^T x$  attains its minimum on  $P$  at some of the vertices of  $P$ .*

WITHOUT PROOF. We are not going to use this result. We need it to understand the idea of the simplex method. ■

**Remark.** It does not mean that  $d^T x$  attains its minimum **only** at vertices of  $P$ .

**Example.** The theorem can be illustrated by an example of a two-dimensional polyhedron and a linear function of two variables.

**Exercise.** Let

$$A = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in \mathbf{R}^3 : x_1 + x_2 + x_3 \leq 1 \right\}$$

Show that the polyhedron  $A$  does not have vertices.

**DEFINITION 4.3** A line segment  $I$  joining two vertices of a polyhedron  $P$  is called an *edge* of  $P$  if none of the points of  $I$  can be represented as  $\alpha u + (1 - \alpha)v$ , where  $0 < \alpha < 1$ ,  $u, v \in P$ , but  $u, v \notin I$ .

**Examples.** Half-plane and angular regions do not have edges because in order to have edges a polyhedron has to have at least two vertices.

**PROPOSITION 4.1** *The line segment joining two vertices of  $C_3$  is an edge of the cube if and only if the vertices have only one different coordinate.*

**PROOF.** A formal complete proof of this statement is rather boring. We shall give proofs only in two special cases (all other cases are quite similar).

**Case 1.** The line segment  $I$  joining

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

is an edge of the cube. By definition

$$I = \{x \in \mathbf{R}^3 : x = \lambda \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + (1 - \lambda) \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} ; 0 \leq \lambda \leq 1\} = \\ \left\{ \begin{bmatrix} 1 \\ 0 \\ 1 - \lambda \end{bmatrix} : 0 \leq \lambda \leq 1 \right\}.$$

So, suppose that some point  $x \in I$  can be represented as  $\alpha u + (1 - \alpha)v$  for some  $0 < \alpha < 1$  and  $u, v$  from the cube. Let

$$u = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \quad \text{and} \quad v = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}.$$

Let us show that  $u_1 = v_1 = 1$  and  $u_2 = v_2 = 0$ . I give the details only for the first statement, the second is similar.

We have  $1 = \alpha u_1 + (1 - \alpha)v_1$  and  $0 \leq u_1, v_1 \leq 1$ . Assume that one of the numbers  $u_1, v_1$  is strictly less than 1. We consider the case when  $u_1 < 1$ , the case when  $v_1 < 1$  is similar. Then, since  $0 < \alpha$  we get  $\alpha u_1 < \alpha$ . Since  $v_1 \leq 1$  and  $1 - \alpha > 0$  we get  $(1 - \alpha)v_1 \leq 1 - \alpha$ .

We get

$$1 = \alpha u_1 + (1 - \alpha)v_1 < \alpha + (1 - \alpha) = 1.$$

This contradiction implies that  $u_1 = 1$ .

In the same way we prove that  $v_1 = 1$  and  $u_2 = v_2 = 0$ . If we compare this fact with the description of  $I$  and the definition of the cube, we see that it implies that  $u, v \in I$ . Hence  $I$  is an edge of the cube.

**Case 2.** The line segment  $J$  joining

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

is not an edge of the cube.

We have

$$J = \left\{ \alpha \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + (1 - \alpha) \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} : 0 \leq \alpha \leq 1 \right\} = \\ \left\{ \begin{bmatrix} 1 \\ 1 - \alpha \\ 1 - \alpha \end{bmatrix} : 0 \leq \alpha \leq 1 \right\}.$$

In particular,

$$x = \frac{1}{2} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} \in J.$$

On the other hand

$$x = \frac{1}{2} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

and the vectors

$$\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

are not in  $J$  (because the 2nd and the 3rd coordinates of points in  $J$  are equal to each other). ■

**DEFINITION 4.4** Two vertices of a polytope  $P$  are called *adjacent* if the line segment joining them is an edge of  $P$ .

**Example.** Vertices of  $C_n$  are adjacent if and only if they have only one different component. (See Proposition 4.1, where this result is proved in the case  $n = 3$ . The proof in the general case is almost the same.)

### Description of the simplex algorithm

(The case when the constraint set is bounded)

Consider an LP problem: minimize  $c^T x$  subject to  $Ax \leq b$ .

**Step 1.** Find a vertex  $x^0$  of the constraint set.

**Step 2.** If the function  $c^T x$  is increasing along all edges with the endpoint at  $x^0$  then **stop**,  $x^0$  is a minimizer.

**Step 3.** If the function  $c^T x$  is decreasing along one of the edges with the endpoint at  $x^0$ , then we find the corresponding adjacent vertex  $x^1$ , and start the **Step 2** anew with  $x^0$  replaced by  $x^1$ .

**Remark.** This description is somewhat imprecise. The actual simplex algorithm is somewhat more complicated, in particular, because it is supposed to give an answer even if the feasible set is unbounded.

To compare the simplex algorithm with the algorithm discussed earlier we need the following result.

**PROPOSITION 4.2** *The set of vertices of the polyhedron corresponding to a standard-form LP problem coincide with the set of basic feasible solutions.*

**PROOF.** The feasible set of the problem is

$$\Omega = \{x \in \mathbf{R}^n : Ax = b, x \geq 0\},$$

where  $A$  is an  $m \times n$  matrix. Let  $a_1, \dots, a_n$  be the columns of  $A$ .

Let  $x$  be the basic feasible solution corresponding to an invertible matrix

$$B = [a_{i_1}, \dots, a_{i_m}].$$

In particular,  $x_l = 0$  if  $l \notin \{i_1, \dots, i_m\}$ . Suppose that  $x = \alpha u + (1 - \alpha)v$ , where  $u, v \in \Omega$ ,  $0 < \alpha < 1$ .

Let us show that  $u_l = v_l = 0$  if  $l \notin \{i_1, \dots, i_m\}$ .

We have  $u_l \geq 0, v_l \geq 0$ . If one of these numbers is  $> 0$  then  $\alpha u_l + (1 - \alpha)v_l > 0$  (we use the fact that  $0 < \alpha < 1$ ). It contradicts the fact that  $0 = x_l = \alpha u_l + (1 - \alpha)v_l$ . Hence  $u_l = v_l = 0$ . Hence the vectors

$$\begin{bmatrix} u_{i_1} \\ \vdots \\ u_{i_m} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} v_{i_1} \\ \vdots \\ v_{i_m} \end{bmatrix}$$

are solutions of the system

$$Bz = b.$$

But this system has only one solution (since  $B$  is invertible). Hence

$$\begin{bmatrix} u_{i_1} \\ \vdots \\ u_{i_m} \end{bmatrix} = \begin{bmatrix} v_{i_1} \\ \vdots \\ v_{i_m} \end{bmatrix} = \begin{bmatrix} x_{i_1} \\ \vdots \\ x_{i_m} \end{bmatrix}.$$

So  $u = v = x$ . Hence  $x$  is a vertex.

The other direction. Suppose that  $x \in \Omega$  is not a basic solution of  $Ax = b$ . Let  $x_{i_1}, \dots, x_{i_k}$  be the non-zero coordinates of  $x$  and let  $a_1, \dots, a_n$  be the columns of  $A$ .

Then, by the definition of a basic solution the columns  $a_{i_1}, \dots, a_{i_k}$  are linearly dependent.

Let us show that  $x$  is not a vertex.

In fact, since  $a_{i_1}, \dots, a_{i_k}$  are linearly dependent, then there exist

$$\lambda_1, \dots, \lambda_k,$$

not all zero, such that

$$\sum_{j=1}^k \lambda_j a_{i_j} = 0.$$

Let  $\delta = \min\{\frac{x_{i_j}}{|\lambda_j|} : \text{over } j \text{ satisfying } \lambda_j \neq 0\}$ . Then  $\delta > 0$ . Let  $u, v \in \mathbf{R}^n$  be defined by

$$u_{i_j} = x_{i_j} + \delta \lambda_j;$$

$$v_{i_j} = x_{i_j} - \delta \lambda_j;$$

and  $u_l = v_l = 0$  if  $l \notin \{i_1, \dots, i_k\}$ . Since  $\delta > 0$  and  $\lambda_1, \dots, \lambda_k$  are not all zero, then  $u \neq v$ .

By the choice of  $\lambda_j$  and  $\delta$  we get  $u, v$  are in  $\Omega$ . Direct computation shows that  $x = \frac{1}{2}u + \frac{1}{2}v$ . Hence  $x$  is not a vertex. ■

**COROLLARY 4.1** *Every LP problem is equivalent to an LP problem whose feasible set has vertices.*

Another interesting consequence of Proposition 4.2 is the following geometric description of the algorithm (find all basic feasible solutions and minimize the function over them): we find **all** vertices of the feasible set and choose the vertices at which the value of the linear functional is minimal (maximal). And the main drawback of this algorithm is that for many problems that arise in applications the feasible set has too many vertices. The simplex method is much better because usually using it we have to consider relatively few vertices. There exists a precise statement that describes in mathematical terms what do we mean by “usually” and “relatively few” here.

**A more detailed description of the simplex algorithm** for a problem of the form:

$$\begin{aligned} & \text{maximize } c^T x \\ & \text{subject to } Ax \leq b, \end{aligned}$$

where  $A$  is an  $k \times n$  matrix,  $b \in \mathbf{R}^k$ ,  $c \in \mathbf{R}^n$ .

**Remark.** It turns out that the details of the simplex algorithm are easier for standard-form LP problems, but the general description is easier for problems of the form described above.

We have observed that every LP problem is equivalent to an LP problem whose feasible set has vertices. Therefore we may assume that the set  $\Omega = \{x : Ax \leq b\}$  has vertices. Suppose also that we have found one of its vertices (later we shall discuss how to find a vertex). Denote the vertex by  $x^0$ .

Let  $\alpha_1, \dots, \alpha_k$  be rows of  $A$  and let  $b_1, \dots, b_k$  be entries of  $b$ . We need the following result.

**LEMMA 4.1** *If  $x^0$  is a vertex of  $\Omega$ , then there exists an  $n \times n$  invertible*

submatrix  $A^0$  of  $A$  such that  $A^0 x^0 = b^0$ , where

$$A^0 = \begin{bmatrix} \alpha_{i_1} \\ \vdots \\ \alpha_{i_n} \end{bmatrix} \quad \text{and} \quad b^0 = \begin{bmatrix} b_{i_1} \\ \vdots \\ b_{i_n} \end{bmatrix}.$$

PROOF. Assume the contrary. Introduce  $D = \{i : 1 \leq i \leq k, \alpha_i x^0 = b_i\}$ . Our assumption means that  $D$  contains no more than  $n - 1$  linearly independent rows.

From Linear Algebra we know that a homogeneous linear system with  $n$  variables and  $\leq n - 1$  equations has non-zero solutions. It implies that a linear homogeneous system with  $\leq n - 1$  linearly independent rows also has non-zero solutions.

We apply this fact to the system

$$\alpha_i x = 0, \quad i \in D.$$

Let  $y$  be a non-zero solution of this system.

Observe that for  $i \notin D$  we have  $\alpha_i x^0 < b_i$ . Hence there exists  $\varepsilon > 0$  such that for  $i \notin D$  we have  $\alpha_i(x^0 + \varepsilon y) < b_i$  and  $\alpha_i(x^0 - \varepsilon y) < b_i$ .

For  $i \in D$  we have  $\alpha_i(x^0 + \varepsilon y) = \alpha_i(x^0 - \varepsilon y) = b_i$ . Hence  $(x^0 + \varepsilon y) \in \Omega$  and  $(x^0 - \varepsilon y) \in \Omega$ . Also  $x^0 = \frac{(x^0 + \varepsilon y) + (x^0 - \varepsilon y)}{2}$ . Since  $(x^0 + \varepsilon y) \neq (x^0 - \varepsilon y)$ , then  $x^0$  is not a vertex. ■

Next problem is: how to find  $A^0$  given  $x^0$ ?

To find  $A^0$  we do the following. We find all rows  $\alpha_i$  of  $A$  satisfying the condition  $\alpha_i x^0 = b_i$  and then find linearly independent subset in this collection of rows using one of the standard procedures from Linear Algebra.

So we assume that a vertex  $x^0 \in \Omega$  and an invertible  $n \times n$  submatrix  $A^0$  of  $A$  satisfying the condition above are given. Let  $E^0 = \{i_1, \dots, i_n\}$  be the numbers of rows included in  $A^0$ .

**Claim.** There exists  $u^0 \in \mathbf{R}^k$  such that  $u_i^0 = 0$  if  $i \notin E^0$  and  $(u^0)^T A = c^T$ .

In fact, we may define  $u^0 \in \mathbf{R}^k$  in the following way  $u_i^0 = 0$  if  $i \notin E^0$

$$\begin{bmatrix} u_{i_1}^0 & \dots & u_{i_n}^0 \end{bmatrix} = c^T (A^0)^{-1}.$$

**Case 1.**  $u^0 \geq 0$ . Then stop:  $x^0$  is optimal, because

$$c^T x^0 = (\text{by the choice of } u^0) = (u^0)^T A x^0 =$$

$$\begin{aligned} & (\text{we use the fact that } b_i = \alpha_i x^0, \text{ if } i \in E^0 \text{ and } u_i^0 = 0, \text{ if } i \notin E^0) \\ & = (u^0)^T b \geq \end{aligned}$$

(we use the fact that  $u^0 \geq 0$  and  $b \geq Ax$  for  $x \in \Omega$ )

$$\geq \max\{(u^0)^T Ax : x \in \Omega\} =$$

(we use the definition of  $u^0$  again)

$$= \max\{c^T x : x \in \Omega\}.$$

**Case 2.**  $u^0 \not\geq 0$ . Choose the smallest index  $i^*$  for which  $u^0$  has negative component  $u_{i^*}^0$ . Let  $y^0 \in \mathbf{R}^n$  be a solution of the system

$$\alpha_k y = 0, \text{ if } k \in E^0 \text{ and } k \neq i^*;$$

$$\alpha_{i^*} y = -1.$$

This system **has** a (unique) solution because the matrix of this system coincide with  $A^0$ , and  $A^0$  is invertible.

**Important Observation.**

$$c^T y^0 = (u^0)^T A y^0 =$$

(we use the definition of  $y^0$ , if  $i \in E^0$  and  $u_i^0 = 0$ , if  $i \notin E^0$ )

$$= -u_{i^*}^0 > 0.$$

Case 2 split into two cases:

**Case 2a.**  $\alpha_i y^0 \leq 0$  for every  $i \in \{1, \dots, k\}$ . Then stop. In this case  $x^0 + \lambda y^0 \in \Omega$  for all  $\lambda \geq 0$ , and we have

$$c^T(x^0 + \lambda y^0) = c^T x^0 + \lambda(c^T y^0) = c^T x^0 + \lambda(-u_{i^*}^0).$$

Hence

$$\max\{c^T x : x \in \Omega\} = \infty,$$

and the problem does not have maximizers.

**Case 2b.**  $\alpha_i y^0 > 0$  for some row  $\alpha_i$  of  $A$ . Let  $\lambda^0$  be the largest  $\lambda \geq 0$  such that  $x^0 + \lambda y^0 \in \Omega$ , that is

$$\lambda^0 := \min \left\{ \frac{b_j - \alpha_j x^0}{\alpha_j y^0} \right\},$$

where the minimum is over  $j$  satisfying  $\alpha_j y^0 > 0$ .

Let  $j^*$  be the smallest index attaining this minimum. Let  $E^1$  be the set of numbers obtained from  $E^0$  if we replace  $i^*$  by  $j^*$ . Let  $A^1$  be the submatrix of  $A$  consisting of rows with numbers in  $E^1$  and  $b^1$  be a subvector of  $b$  corresponding to  $E^1$ . Let  $x^1 = x^0 + \lambda^0 y^0$ .

**LEMMA 4.2** (1)  $c^T x^1 \geq c^T x^0$  unless  $x^1 = x^0$ .

(2)  $A^1 x^1 = b^1$ .

(3)  $A^1$  is invertible.

**PROOF.** (1). Since  $c^T y^0 > 0$ , then  $c^T(x^1) = c^T(x^0 + \lambda^0 y^0) > c^T(x^0)$  unless  $x^1 = x^0$ .

(2). By the choice of  $j^*$  we have

$$\lambda^0 = \frac{b_{j^*} - \alpha_{j^*} x^0}{\alpha_{j^*} y^0}$$

This equation can be rewritten in the form

$$b_{j^*} = \alpha_{j^*}(x^0 + \lambda^0 y^0)$$

or  $b_{j^*} = \alpha_{j^*} x^1$ .

Let  $i \in E^1$  and  $i \neq j^*$ . By the definition of  $y^0$  we have  $\alpha_i y^0 = 0$  if  $i \in E^0$  and  $i \neq i^*$ . Hence  $\alpha_i y^0 = 0$  and

$$\begin{aligned} b_i &= \alpha_i x^0 = \alpha_i x^0 + \lambda^0(\alpha_i y^0) = \\ &\alpha_i(x^0 + \lambda^0 y^0) = \alpha_i x^1. \end{aligned}$$

We have proved

$$A^1 x^1 = b^1.$$

(3). We have:  $\alpha_i y^0 = 0$  for all rows of  $A^1$  except  $\alpha_{j^*}$  and  $\alpha_{j^*} y^0 > 0$ . Hence the row  $\alpha_{j^*}$  is not a linear combination of the other rows of  $A^1$ . The other rows of  $A^1$  are linearly independent because they are rows of the invertible matrix  $A^0$ . Hence the rows of  $A^1$  are linearly independent, so  $A^1$  is invertible. ■

Therefore we can start the process anew with  $A^0, x^0$  replaced by  $A^1, x^1$ . Denote by  $u^1$  and  $y^1$  the corresponding vectors.

**Remark.** (Geometric meaning). It can be verified that  $x^1$  is a vertex of  $\Omega$  and the line segment between  $x^0$  and  $x^1$  is an edge of  $\Omega$  (unless  $x^0 = x^1$ ).

Repeating the process we find

$$A^2, x^2, E^2, b^2, u^2, y^2; A^3, x^3, E^3, b^3, u^3, y^3; \dots$$

## Summary

In the  $j$ th iteration we start with  $A^{j-1}, x^{j-1}, E^{j-1}, b^{j-1}$  and find  $u^{j-1}$ .

If it is non-negative, we **stop**. In such a case the vector  $x^{j-1}$  is the answer to the problem.

If  $u^{j-1}$  has negative coordinates, we find  $y^{j-1}$ .

If it satisfies the condition  $\alpha_i y^{j-1} \leq 0$  for every  $i \in \{1, \dots, k\}$ , we **stop**. It means that the maximum in the problem is  $\infty$ .

If  $\alpha_i y^{j-1} > 0$  for some  $i \in \{1, \dots, k\}$ , then we find  $A^j, x^j, E^j, b^j$  and go to the  $(j + 1)$ th iteration.

From this description we see that the process terminates if in one of the iterations we get either Case 1 or Case 2a.

**THEOREM 4.2** *The process terminates.*

**PROOF.** Suppose that the process does not terminate. Since there are only finitely many submatrices of  $A$ , then there exist positive integers  $m, l$  such that  $m < l$  and  $A^m = A^l$ . Hence  $x^m = x^l$ .

By the observation above we have

$$c^T x^0 \leq c^T x^1 \leq \dots \leq c^T x^m \leq \dots \leq c^T x^l \leq \dots$$

and  $c^T x^i < c^T x^{i+1}$  unless  $x^i = x^{i+1}$ . It follows that

$$x^m = x^{m+1} = \dots = x^l.$$

Let  $r$  be the largest index for which  $\alpha_r$  has been removed from  $A^t$  in the iteration  $t + 1$ ,  $t = m, m + 1, \dots, l - 1$ , say in iteration  $p + 1$ . Since  $A^m = A^l$  we know that  $\alpha_r$  also has been added to  $A^q$  in some iteration  $q + 1$  with  $m \leq q < l$ .

Since  $r$  is the largest index for which  $\alpha_r$  has been removed from  $A^t$  for  $t = m, m + 1, \dots, l - 1$ , then for  $j > r$  we have:

$$(*) \quad j \in E^p \text{ if and only if } j \in E^t \text{ for every } t = m, m + 1, \dots, l.$$

By the choice of  $u^p$  and  $y^q$  we have  $(u^p)^T A y^q = c^T y^q > 0$ . Let  $(u^p)^T = [u_1^p, \dots, u_k^p]$ . It implies that that

$$u_j^p(\alpha_j y^q) > 0 \tag{**}$$

for at least one value of  $j$ . Now we shall show that it is impossible.

1) If  $j \notin E^p$  we have  $u_j^p = 0$  by the definition of  $u^p$ .

2) If  $j \in E^p$  and  $j < r$ , then

a) Since  $\alpha_r$  was removed from  $A^p$ , then  $r$  is the smallest index of a negative component of  $u^p$ . Hence  $u_j^p \geq 0$ ;

b) Observe that  $x^{q+1} = x^q$ . It implies that  $\lambda^q = 0$ . Since  $\alpha_r$  was added to  $A^q$  in the iteration  $q + 1$ , then  $r$  is the smallest index  $j$  that satisfies the conditions  $\alpha_j y^q > 0$  or  $\alpha_j x^q = b_j$  simultaneously. (Analyse the definition of  $\lambda_0$  and the choice of  $j^*$ .) It remains to observe that if  $\alpha_j x^q < b_j$ , then  $j$  is not in any of the sets  $E^m, E^{m+1}, \dots, E^l$  (we use the fact that  $x^m = \dots = x^l$  and the definition of the sets  $E^0, E^1, \dots$ ). Hence either  $\alpha_j y^q \leq 0$  or  $j \notin E^p$ . Since we consider  $j \in E^p$  the second case cannot occur. In the first case we have  $u_j^p \alpha_j y^q \leq 0$ , so the inequality (\*\*) does not take place.

3) If  $j \in E^p$  and  $j = r$ , then  $u_r^p < 0$  (see the description of the choice of the row that we remove) and  $\alpha_r y^q > 0$  (see the description of the choice of the row that we add).

So  $u_r^p \alpha_r y^q < 0$ .

4) If  $j \in E^p$  and  $j > r$ , then  $j \in E^q$  and is not removed from  $E^q$  in iteration  $q + 1$  (by (\*)). Hence  $\alpha_j y^q = 0$  (by the definition of  $y^q$ ).

We see that (\*\*) is impossible. Hence the process terminates. ■

Now we shall discuss the problem: how to find an initial vertex? It is worthwhile to emphasize that this problem is non-trivial.

We shall consider this problem for a standard-form LP problem. So we consider the problem:

$$\begin{aligned} & \text{minimize } c^T x \\ & \text{subject to } Ax = b, x \geq 0, \end{aligned}$$

where  $A$  is an  $m \times n$  matrix,  $c, x \in \mathbf{R}^n$ ,  $b \in \mathbf{R}^m$ .

Observe that without loss of generality we may assume that  $b \geq 0$ . In fact, if some  $b_i < 0$  we replace  $b_i$  by  $-b_i$  and replace  $\alpha_i$  (the  $i$ th row of  $A$ ) by  $-\alpha_i$ . We get an equivalent system of equations. After doing this for all negative  $b_i$ 's we get what we need.

To find a vertex we consider the following associated *artificial problem*:

$$\begin{aligned} & \text{minimize } y_1 + y_2 + \cdots + y_m \\ & \text{subject to } [A \ I] \begin{bmatrix} x \\ y \end{bmatrix} = b, \begin{bmatrix} x \\ y \end{bmatrix} \geq 0. \end{aligned}$$

Observe, that the polyhedron

$$\Omega_1 = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} : [A \ I] \begin{bmatrix} x \\ y \end{bmatrix} = b, \begin{bmatrix} x \\ y \end{bmatrix} \geq 0 \right\}$$

has an obvious vertex:

$$\begin{bmatrix} 0 \\ b \end{bmatrix}.$$

(The fact that it is a vertex follows from the proposition describing vertices of a feasible set of a standard-form LP problem.)

Therefore we can solve the artificial problem using the simplex method.

Observe that the minimum in this problem is  $\geq 0$ . It turns out that if the set  $\Omega = \{x : Ax = b, x \geq 0\}$  is non-empty, then the minimum is 0. In fact, let  $x^1 \in \Omega$ , then

$$\begin{bmatrix} x^1 \\ 0 \end{bmatrix} \in \Omega_1$$

and the value of the functional on this element is 0.

Hence, if the minimum in the artificial problem is  $> 0$ , then the original problem is ill-posed.

If the minimum is 0, then the optimal feasible solution that we find using the simplex method is of the form:

$$\begin{bmatrix} x^0 \\ 0 \end{bmatrix}.$$

Since it is a vertex of

$$\left\{ \begin{bmatrix} x \\ y \end{bmatrix} : [A, I] \begin{bmatrix} x \\ y \end{bmatrix} = b, \begin{bmatrix} x \\ y \end{bmatrix} \geq 0 \right\},$$

then the columns of  $A$ , that correspond to non-zero coordinates of  $x^0$  are linearly independent (see the corresponding proposition). By the same proposition  $x^0$  is a vertex of  $\Omega = \{x : Ax = b, x \geq 0\}$ .

For more information on the simplex algorithm see [3] and [11].

## 5 Solutions of some of the HW problems

1. (HW # 7, Part 2). One of the important results about determinants is the following identity due to Vandermonde (see [10] (p. 3)):

Let  $a_1, a_2, \dots, a_n$  be real numbers. Then

$$\det \begin{bmatrix} 1 & 1 & \dots & 1 \\ a_1 & a_2 & \dots & a_n \\ a_1^2 & a_2^2 & \dots & a_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^{n-1} & a_2^{n-1} & \dots & a_n^{n-1} \end{bmatrix} = \prod_{i>j} (a_i - a_j).$$

(The right-hand side is the product of all differences between  $a_i$  and  $a_j$  ( $i > j$ ).)

Use this identity to show that for any  $m$  and any  $n \geq m$  there exists an  $m \times n$  matrix all of whose  $m \times m$  submatrices are invertible.

**SOLUTION.** The identity implies the following:

**COROLLARY 5.1** *For any positive integer  $m$  and for any real numbers  $b_1, \dots, b_m$  such that  $b_i \neq b_j$  if  $i \neq j$ , we have:*

$$\det \begin{bmatrix} 1 & 1 & \dots & 1 \\ b_1 & b_2 & \dots & b_m \\ \vdots & \vdots & \ddots & \vdots \\ b_1^{m-1} & b_2^{m-1} & \dots & b_m^{m-1} \end{bmatrix} \neq 0.$$

Now, we consider a collection consisting of  $n$  real numbers  $a_1, \dots, a_n$  satisfying  $a_i \neq a_j$  if  $i \neq j$ . Such collection exists, for example, the collection of the first  $n$  positive integers satisfy this condition.

Let  $A$  be the  $m \times n$  matrix given by

$$A = \begin{bmatrix} 1 & 1 & \dots & 1 \\ a_1 & a_2 & \dots & a_n \\ a_1^2 & a_2^2 & \dots & a_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^{m-1} & a_2^{m-1} & \dots & a_n^{m-1} \end{bmatrix}.$$

Then any  $m \times m$  submatrix of  $A$  satisfy the condition of Corollary 5.1. Hence its determinant is non-zero, and it is invertible. ■

2. (HW # 3, Part 2). State the analogue of the following theorem for maximizers.

**Theorem.** Let  $f : \mathbf{R}^n \rightarrow \mathbf{R}$  be twice continuously differentiable.

1. If  $x^*$  is a local minimizer of  $f$  over  $\mathbf{R}^n$ , then  $Df(x^*) = 0$  and  $x^T D^2 f(x^*) x$  is positive semidefinite.

2. If  $Df(x^*) = 0$  and  $x^T D^2 f(x^*) x$  is positive definite, then  $x^*$  is a strict local minimizer of  $f$  over  $\mathbf{R}^n$ .

3. If  $Df(x^*) = 0$  and  $x^T D^2 f(z) x$  is positive semidefinite for every  $z \in \mathbf{R}^n$ , then  $x^*$  is a global minimizer of  $f$  over  $\mathbf{R}^n$ .

ANSWER.

**THEOREM 5.1** Let  $f : \mathbf{R}^n \rightarrow \mathbf{R}$  be twice continuously differentiable.

1. If  $x^*$  is a local maximizer of  $f$  over  $\mathbf{R}^n$ , then  $Df(x^*) = 0$  and  $x^T D^2 f(x^*) x$  is negative semidefinite.

2. If  $Df(x^*) = 0$  and  $x^T D^2 f(x^*) x$  is negative definite, then  $x^*$  is a strict local maximizer of  $f$  over  $\mathbf{R}^n$ .

3. If  $Df(x^*) = 0$  and  $x^T D^2 f(z) x$  is negative semidefinite for every  $z \in \mathbf{R}^n$ , then  $x^*$  is a global maximizer of  $f$  over  $\mathbf{R}^n$ .

■

3. (HW # 7, Part 1). **True or false?**

Let  $f : \mathbf{R}^n \rightarrow \mathbf{R}$  be a twice continuously differentiable function. Let  $x$  be a critical point of  $f$ , and  $D^2 f(x)$  be the Hessian at  $x$ .

(a) If  $D^2 f(x)$  is positive semidefinite, then  $x$  is a local minimizer.

(b) If  $D^2 f(x)$  is indefinite, then  $x$  is neither local minimizer nor local maximizer.

(c) If  $D^2 f(x)$  is indefinite, then the nature of  $x$  cannot be determined using the Hessian.

(d) If  $D^2 f(x)$  is negative definite, then  $x$  is a strict local maximizer.

(e) If  $D^2f(x)$  is positive semidefinite, but not positive definite, then the nature of  $x$  cannot be determined using the Hessian.

(f) If  $D^2f(x)$  is negative semidefinite, but not negative definite, then the nature of  $x$  cannot be determined using the Hessian.

ANSWER. (a) False; (b) True; (c) False; (d) True; (e) True; (f) True.

I find it useful to compile the following table:

Let  $x$  be a critical point of  $f : \mathbf{R}^n \rightarrow \mathbf{R}$ . Then

$D^2f(x)$	the nature of $x$
indefinite	neither local min., nor local maximizer
positive definite	strict local minimizer
negative definite	strict local maximizer
pos. semidef., but not pos. def.	cannot be determ. using the Hessian
neg. semidef., but not neg. def.	cannot be determ. using the Hessian

The statements contained in the 2nd and the 3rd lines were stated explicitly in Theorem 1.7 and in the theorem that you were expected to state in HW # 3, Part 2 (see Theorem 5.1 above). So we shall discuss the remaining statements only. The mentioned theorems contain the following statements:

- if  $x$  is a local minimizer, then  $D^2f(x)$  is positive semidefinite;
- if  $x$  is a local maximizer, then  $D^2f(x)$  is negative semidefinite.

If  $D^2f(x)$  is indefinite, it is neither positive semidefinite, nor negative semidefinite. Hence  $x$  is neither local minimizer, nor local maximizer.

The last two statements can be proved using one-dimensional examples. For example, if  $f(x) = x^3$ , then 0 is also a critical point, and the Hessian (in one dimensional case it coincides with the second derivative) at 0 is 0, but 0 is neither minimizer, nor maximizer (graph). If  $f(x) = x^4$ , then 0 is also a critical point, and the Hessian at 0 is also equal 0, but in this case 0 is a strict local minimizer. Observe that the  $1 \times 1$  matrix whose only

entry is equal to 0 is both negative semidefinite and positive semidefinite, but neither negative definite nor positive definite. This example proves the last two statements from the table. ■

4. (HW # 6) State an analogue of Theorem 5.2 for functions satisfying  $\lim_{\|x\| \rightarrow \infty} f(x) = -\infty$ .

**THEOREM 5.2** (a) Let  $q_1, q_2, \dots, q_n$  be linearly independent linear functionals on  $\mathbf{R}^n$  and  $P_i : \mathbf{R} \rightarrow \mathbf{R}$   $i = 1, \dots, n$  be such that

$$\lim_{z \rightarrow \infty} P_i(z) = \infty \quad \text{and} \quad \lim_{z \rightarrow -\infty} P_i(z) = \infty.$$

Then

$$\lim_{\|x\| \rightarrow \infty} \sum_{i=1}^n P_i(q_i(x)) = \infty.$$

(b) Suppose that  $\lim_{\|x\| \rightarrow \infty} f(x) = \infty$  and that  $g(x)$  is such that  $g(x) \geq C$  for some real number  $C$  and every  $x \in \mathbf{R}^n$ . Then  $\lim_{\|x\| \rightarrow \infty} (f(x) + g(x)) = \infty$ .

ANSWER.

**THEOREM 5.3** (a) Let  $q_1, q_2, \dots, q_n$  be linearly independent linear functionals on  $\mathbf{R}^n$  and  $P_i : \mathbf{R} \rightarrow \mathbf{R}$   $i = 1, \dots, n$  be such that

$$\lim_{z \rightarrow \infty} P_i(z) = -\infty \quad \text{and} \quad \lim_{z \rightarrow -\infty} P_i(z) = -\infty.$$

Then

$$\lim_{\|x\| \rightarrow \infty} \sum_{i=1}^n P_i(q_i(x)) = -\infty.$$

(b) Suppose that  $\lim_{\|x\| \rightarrow \infty} f(x) = -\infty$  and that  $g(x)$  is such that  $g(x) \leq C$  for some real number  $C$  and every  $x \in \mathbf{R}^n$ . Then  $\lim_{\|x\| \rightarrow \infty} (f(x) + g(x)) = -\infty$ .

■

5. (HW # 7, Part 1). Whether  $\lim_{\|x\| \rightarrow \infty} f(x) = \infty$ ,  $\lim_{\|x\| \rightarrow \infty} f(x) = -\infty$  or neither? Support you answer.

1.

$$f(x_1, x_2, x_3) = (x_1 + x_2)^4 - (x_1 + x_2)^3 + (x_2 + x_3)^8 + (x_1 + 2x_2 + x_3)^2 + e^{x_3^2}.$$

SOLUTION. The function  $f$  can be represented in the form  $\sum_{i=1}^4 P_i(q_i(x))$  where  $q_1, q_2, q_3, q_4$  are linear functionals on  $\mathbf{R}^3$ :

$$q_1(x) = x_1 + x_2,$$

$$q_2(x) = x_2 + x_3,$$

$$q_3(x) = x_1 + 2x_2 + x_3,$$

$$q_4(x) = x_3;$$

and  $P_i : \mathbf{R} \rightarrow \mathbf{R}$ ,  $i = 1, \dots, 4$ , are given by

$$P_1(z) = z^4 - z^3, \quad P_2(z) = z^8, \quad P_3(z) = z^2, \quad P_4(z) = e^{z^2}.$$

From the well-known properties of polynomial and exponential functions we get:

$$\lim_{z \rightarrow \infty} P_i(z) = \infty \quad \text{and} \quad \lim_{z \rightarrow -\infty} P_i(z) = \infty.$$

But the linear functionals  $q_i$  are not linearly independent (and it is impossible in principle to find 4 linearly independent functionals on  $\mathbf{R}^3$ ). In such a case (that is, when we have more functionals than the dimension of the space) we check whether it is possible to find 3 (in general case  $n$  = the dimension of the space) linearly independent functionals among  $q_1, \dots, q_4$ . In this example it is possible, for example  $q_1, q_2$  and  $q_4$  are linearly independent because

$$\det \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} = 1 \neq 0.$$

By Theorem 5.2(a) we get

$$\lim_{\|x\| \rightarrow \infty} P_1(q_1(x)) + P_2(q_2(x)) + P_4(q_4(x)) = \infty$$

Observe also, that the function  $P_3(q_3(x))$  is bounded from below, more precisely  $(x_1 + 2x_2 + x_3)^2 \geq 0$ . Applying Theorem 5.2(b) we get

$$\lim_{\|x\| \rightarrow \infty} f(x) = \infty. \quad \blacksquare$$

$$2. f(x_1, x_2, x_3) = (x_1 + x_2 + x_3)^2 + (x_1 + x_2)^4 + x_3^8$$

SOLUTION. In this case the function  $f$  can be represented in the form  $\sum_{i=1}^3 P_i(q_i(x))$  where  $P_i : \mathbf{R} \rightarrow \mathbf{R}$ ,  $i = 1, 2, 3$  are given by

$$P_1(z) = z^2, P_2(z) = z^4, P_3(z) = z^8;$$

$q_1, q_2, q_3$  are linear functionals,

$$q_1(x) = x_1 + x_2 + x_3, q_2(x) = x_1 + x_2, q_3(x) = x_3.$$

It is clear that  $P_1, P_2$  and  $P_3$  satisfy

$$\lim_{z \rightarrow \infty} P_i(z) = \infty \quad \text{and} \quad \lim_{z \rightarrow -\infty} P_i(z) = \infty.$$

But the functionals  $q_1, q_2, q_3$  are not linearly independent because

$$\det \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = 0.$$

And in this case (for obvious reasons) we cannot find 3 linearly independent functionals among  $q_1, q_2, q_3$ . In such cases the answer is: neither (the limit of  $f$  at infinity is neither  $+\infty$  nor  $-\infty$ ). We are not going to prove this result in general case, but only for the given example.

From the definitions of infinite limits we see that it is enough to show that for every real  $N$  there exists  $x \in \mathbf{R}^3$  such that  $\|x\| > N$  and  $f(x) = 0$ .

First we find a nonzero vector  $v$  such that  $q_1(v) = q_2(v) = q_3(v) = 0$ , for example

$$v = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$$

satisfies this condition.

With such choice of  $v$  we can show that (for example) the vector  $Nv$  satisfies these conditions. In fact,  $\|Nv\| = \sqrt{N^2 + (-N)^2 + 0^2} = \sqrt{2}N > N$ , and  $f(Nv) = (N - N + 0)^2 + (N - N)^4 + 0^8 = 0$ . ■

3.

$$f(x_1, x_2, x_3) = 100 \sin x_1 - 2x_1^2 + 4x_1 - (x_1 + x_2)^2 - e^{(x_2 + x_3)^2}.$$

SOLUTION. In this case the function  $f$  can be represented in the form  $\sum_{i=1}^3 P_i(q_i(x))$  where  $P_i : \mathbf{R} \rightarrow \mathbf{R}$ ,  $i = 1, 2, 3$  are given by

$$P_1(z) = 100 \sin z - 2z^2 + 4z, \quad P_2(z) = -z^2, \quad P_3(z) = -e^{z^2},$$

and  $q_1, q_2, q_3$  are linear functionals,  $q_1(x) = x_1$ ,  $q_2(x) = x_1 + x_2$ ,  $q_3(x) = x_2 + x_3$ . The functionals  $q_1, q_2, q_3$  are linearly independent, because

$$\det \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} = 1 \neq 0.$$

Let us show that the functions  $P_i(z)$  satisfy

$$\lim_{z \rightarrow \infty} P_1(z) = -\infty \quad \text{and} \quad \lim_{z \rightarrow -\infty} P_1(z) = -\infty.$$

For  $P_2$  it follows from properties of polynomials, for  $P_3$  it follows from the corresponding property of the exponential function.

Observe that

$$\lim_{|z| \rightarrow \infty} -2z^2 + 4z = -\infty$$

and that

$$100 \sin z \leq 100 \quad \forall z \in \mathbf{R}.$$

Hence, by Theorem 5.3(b) (applied in one-dimensional case) we get

$$\lim_{|z| \rightarrow \infty} P_1(z) = -\infty$$

By Theorem 5.3(a) we get  $\lim_{\|x\| \rightarrow \infty} f(x) = -\infty$ . ■

6. (HW # 8, Part 2). Let

$$A = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in \mathbf{R}^3 : x_1 + x_2 + x_3 \leq 1 \right\}$$

Show that the polyhedron  $A$  does not have vertices.

SOLUTION. We find a nonzero vector

$$v = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$$

satisfying  $v_1 + v_2 + v_3 = 0$ . One of the possible choices is

$$v = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

We need to show that any point in  $A$  is not a vertex. So let

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \text{ be any vector from } A.$$

Then  $x + v$  and  $x - v$  are also in  $A$ . In fact, we have to verify that

$$(x_1 + v_1) + (x_2 + v_2) + (x_3 + v_3) \leq 1$$

and

$$(x_1 - v_1) + (x_2 - v_2) + (x_3 - v_3) \leq 1$$

provided  $x_1 + x_2 + x_3 \leq 1$ .

But  $(x_1 + v_1) + (x_2 + v_2) + (x_3 + v_3) = x_1 + x_2 + x_3 + v_1 + v_2 + v_3 = x_1 + x_2 + x_3 + 1 - 1 + 0 = x_1 + x_2 + x_3 \leq 1$ . Similarly  $(x_1 - v_1) + (x_2 - v_2) + (x_3 - v_3) = x_1 + x_2 + x_3 - v_1 - v_2 - v_3 = x_1 + x_2 + x_3 - 1 - (-1) - 0 = x_1 + x_2 + x_3 \leq 1$ .

Also  $x = \frac{1}{2}(x + v) + \frac{1}{2}(x - v)$  and  $x + v \neq x - v$  (because  $v$  is nonzero). Hence  $x$  is not vertex of  $A$ . ■

**General Remark.** A similar approach can be used to show that a polyhedron  $\Omega = \{x \in \mathbf{R}^n : Hx \leq b\}$  does not have vertices if (and only if) the system  $Hx = 0$  has a nonzero solution.

7. (HW # 8, Part 1). Use the algorithm:

*Find all basic feasible solutions and choose the ones with the minimal value of the function*

to minimize  $f(x) = x_1 + 2x_2 + 3x_3 + 4x_4$  subject to  $x \geq 0$  and  $Ax = b$ , where  $x \in \mathbf{R}^4$  and

$$A = \begin{bmatrix} 3 & 2 & 1 & 2 \\ 2 & 1 & 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

**SOLUTION.** All elements of  $A$  are nonnegative, and each column of  $A$  contains a strictly positive entry. Hence we may use the algorithm. But in this case we do not need to consider all invertible  $2 \times 2$  submatrices of  $A$ , because we can immediately prove: the system  $Ax = b$  does not have feasible solutions. In fact, assume the contrary, let

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

be a feasible solution. Then  $x_1, x_2, x_3, x_4 \geq 0$  and

$$3x_1 + 2x_2 + x_3 + 2x_4 = 2x_1 + x_2 + x_4 = 1$$

On the other hand for nonnegative  $x_1, x_2, x_3, x_4$  we have

$$3x_1 + 2x_2 + x_3 + 2x_4 \geq 2x_1 + x_2 + x_4,$$

and the inequality is strict unless  $x_1 = x_2 = x_3 = x_4 = 0$ . But the zero vector does not satisfy  $Ax = b$ . Hence  $x$  is not a feasible solution. This contradiction proves the statement.

**Answer.** The feasible set is empty. ■

## References

- [1] A. Brøndsted, *An introduction to convex polytopes*, Springer-Verlag, 1983.
- [2] E. K. P. Chong and S. H. Zak, *An Introduction to Optimization*, John Wiley & Sons, Inc., 1996.
- [3] G. B. Dantzig, *Linear programming and extensions*, Princeton, N.J., Princeton University Press, 1963 (11th paperback printing, 1998).
- [4] H. Helson, *Honors Calculus*, Second Edition, GRT Book Printing, Oakland, CA, 1995.
- [5] R. Horn and C. Johnson, *Matrix analysis*, Cambridge University Press, 1985.
- [6] R. E. Larson, R. P. Hostetler and B. H. Edwards, *Calculus with Analytic Geometry*, Houghton Mifflin Company, Boston New York, 1998.
- [7] J. E. Marsden, A. J. Tromba and A. Weinstein, *Basic Multivariable Calculus*, Springer & W. H. Freeman and Company, 1993.
- [8] A. L. Peressini, F. E. Sullivan and J. J. Uhl, Jr., *The Mathematics of Nonlinear Programming*, Springer-Verlag, 1988.
- [9] E. Polak, *Optimization. Algorithms and consistent approximations*, Applied Mathematical Sciences, **124**. Springer-Verlag, New York, 1997. xx+779 pp. ISBN: 0-387-94971-2 (MR98g:49001).
- [10] V. V. Prasolov, *Problems and theorems in linear algebra*, American Mathematical Society, Providence, RI, 1994.
- [11] A. Schrijver, *Theory of integer and linear programming*, John Wiley & Sons, 1986.
- [12] R. K. Sundaram, *A first course in optimization theory*, Cambridge University Press, Cambridge, 1996. xviii+357 pp. ISBN: 0-521-49719-1 (MR97g:90002).
- [13] R. Webster, *Convexity*, Oxford University Press, 1994.