Chapter 3

Simple Problems

1 Rank

We prove the Rank Theorem and discuss its consequences for the theory of linear equations and geometry of linear subspaces. Perhaps, the reader will find most of the arguments here rather shallow. This is because the key effort was made earlier, in Chapter 2, in the lemma on linear dependence of any $n+1$ vectors of $\mathbb{K}^n$.

The Rank Theorem

In the subsection *Bases and Dimension*, we constructed a basis of a finite dimensional vector space by starting from a finite set that spans it and removing unnecessary vectors. Alternatively, one can construct a basis by starting from any linearly independent set and adding, one by one, new vectors linearly independent from the previous ones. Since the number of such vectors cannot exceed the dimension of the space, the process will stop when the vectors span the whole space and form therefore a basis. Thus we have proved that in a finite dimensional vector space, ***every linearly independent set of vectors can be completed to a basis.***[1] We are going to use this in the proof of the Rank Theorem.

The **rank** of a linear map $A : \mathcal{V} \to \mathcal{W}$ is defined as the dimension of its range: $\operatorname{rk} A := \dim A(\mathcal{V})$.

[1] Using the so called *transfinite induction* one can prove the same for infinite dimensional vector spaces as well.

`Example.` Consider the map $E_r : \mathbb{K}^n \to \mathbb{K}^m$ given by the block matrix $E_r = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}$, of size $m \times n$, where the left upper block is the identity matrix I_r of size $r \times r$, and the other three blocks are zero matrices of appropriate sizes. In standard coordinates $(x_1, \dots, x_n)$ in $\mathbb{K}^n$ and $(y_1, \dots, y_m)$ in $\mathbb{K}^m$, the map E_r is given by the formulas $y_1 = x_1, \dots, y_r = x_r, y_{r+1} = 0, \dots, y_m = 0$. The range of E_r is the subspace of dimension r in $\mathbb{K}^m$ given by the $m - r$ equations $y_{r+1} = \dots = y_m$. Thus $\operatorname{rk} E_r = r$. The kernel of E_r is the subspace of dimension $n - r$ in $\mathbb{K}^n$ given by r equations $x_1 = \dots = x_r = 0$. The map can be viewed geometrically as the projection along the kernel onto the range.

`The Rank Theorem.` ***A linear map*** $A : \mathcal{V} \to \mathcal{W}$ ***of rank*** r ***between two vector spaces of dimensions*** n ***and*** m ***is given by the matrix*** E_r ***in suitable bases of the spaces*** $\mathcal{V}$ ***and*** $\mathcal{W}$***.***

`Proof.` Let $\mathbf{f}_1, \dots, \mathbf{f}_r$ be any basis in the range $A(\mathcal{V}) \subset \mathcal{W}$ (Figure 32). Complete it to a basis of $\mathcal{W}$ by choosing vectors $\mathbf{f}_{r+1}, \dots, \mathbf{f}_m$ as explained above. Pick vectors $\mathbf{e}_1, \dots, \mathbf{e}_r \in \mathcal{V}$ such that $A\mathbf{e}_i = \mathbf{f}_i$. (They exist because $\mathbf{f}_i$ lie in the range of A.) Take vectors $\mathbf{e}_{r+1}, \mathbf{e}_{r+2}, \dots$ to form a basis in the kernel of A. We claim that $\mathbf{e}_1, \dots, \mathbf{e}_r, \mathbf{e}_{r+1} \dots$ form a basis in $\mathcal{V}$ (and in particular the total number of these vectors is equal to n). The theorem follows from this, since $A\mathbf{e}_i = \mathbf{f}_i$ for $i = 1, \dots r$, and $A\mathbf{e}_i = \mathbf{0}$ for $i = r + 1, \dots, n$, and hence the matrix of A in these bases coincides with E_r.

To justify the claim, we will show that every vector $\mathbf{x} \in \mathcal{V}$ is uniquely written as a linear combination of $\mathbf{e}_i$. Indeed, we have: $A\mathbf{x} = \alpha_1\mathbf{f}_1 + \dots + \alpha_r\mathbf{f}_r$ since $A\mathbf{x}$ lies in the range of A. Then $A(\mathbf{x} - \alpha_1\mathbf{e}_1 - \dots - \alpha_r\mathbf{e}_r) = \mathbf{0}$, and hence $\mathbf{x} - \alpha_1\mathbf{e}_1 - \dots - \alpha_r\mathbf{e}_r$ lies in the kernel of A. Therefore $\mathbf{x} = \alpha_1\mathbf{e}_1 + \dots + \alpha_r\mathbf{e}_r + \alpha_{r+1}\mathbf{e}_{r+1} + \dots$, i.e. the vectors $\mathbf{v}_i$ span $\mathcal{V}$. On the other hand, if in the last equality we have $\mathbf{x} = \mathbf{0}$, then $A\mathbf{x} = \alpha_1\mathbf{f}_1 + \dots + \alpha_r\mathbf{f}_r = \mathbf{0}$ and hence $\alpha_1 = \dots = \alpha_r = 0$, since $\mathbf{f}_i$ are linearly independent in $\mathcal{W}$. Finally, $\mathbf{0} = \alpha_{r+1}\mathbf{e}_{r+1} + \alpha_{r+2}\mathbf{e}_{r+2} + \dots$ implies that $\alpha_{r+1} = \alpha_{r+2} = \dots = 0$ since $\mathbf{e}_{r+1}, \mathbf{e}_{r+2}, \dots$ are linearly independent in $\mathcal{V}$. □

Let A be an $m \times n$ matrix. It defines a linear map $\mathbb{K}^n \to \mathbb{K}^m$. The rank of this map is the dimension of the subspace in $\mathbb{K}^m$ spanned by columns of A. It is called the **rank** of the matrix A. Applying the Rank Theorem to this linear map, we obtain the following result.

Corollary. ***For every $m \times n$-matrix A of rank r there exist invertible matrices D and C of sizes $m \times m$ and $n \times n$ respectively such that $D^{-1}AC = E_r$.***

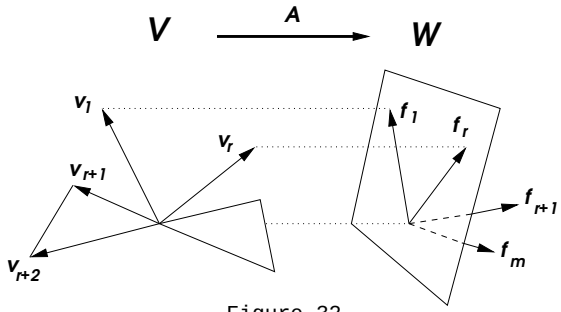

Figure 32

The Rank Theorem has the following reformulation. Let $A : \mathcal{V} \to \mathcal{W}$ and $A' : \mathcal{V}' \to \mathcal{W}'$ be two linear maps. They are called **equivalent** if there exist isomorphisms $C : \mathcal{V}' \xrightarrow{\cong} V$ and $D : \mathcal{W}' \xrightarrow{\cong} W$ such that $DA' = AC$. One expresses the last equality by saying that the following **square is commutative**:

$$\begin{array}{ccc} \mathcal{V}' & \xrightarrow{A'} & \mathcal{W}' \\ C\downarrow\cong & & \cong\downarrow D \\ \mathcal{V} & \xrightarrow{A} & \mathcal{W} \end{array} .$$

The Rank Theorem′. ***Linear maps between finite dimensional spaces are equivalent if and only if they have the same rank.***

Indeed, when $A' = D^{-1}AC$, the ranges of A and A' must have the same dimension (since C and D are isomorphisms). Conversely, when $\operatorname{rk} A = r = \operatorname{rk} A'$, each A and A' is equivalent to $E_r : \mathbb{K}^n \to \mathbb{K}^m$ by the Rank Theorem.

Below we discuss further corollaries and applications of the Rank Theorem.

EXERCISES

242. Prove that columns of an invertible $n \times n$-matrix form a basis in $\mathbb{K}^n$, and *vice versa: every basis in $\mathbb{K}^n$ is thus obtained.*

243. Professor Dumbel writes his office and home phone numbers as a 7×1-matrix O and 1×7-matrix H respectively. Help him compute $\operatorname{rk}(OH)$. ϟ✓

244. Prove that $\operatorname{rk}(A+B) \le \operatorname{rk} A + \operatorname{rk} B$. ϟ

245. Following the proof of the Rank Theorem, find bases in the domain and the target spaces in which the following linear map $A : \mathbb{K}^3 \to \mathbb{K}^3$

$$\begin{aligned} y_1 &= 2x_1 - x_2 - x_3 \\ y_2 &= -x_1 + 2x_2 - x_3 \\ y_3 &= -x_1 - x_2 + 2x_3 \end{aligned}$$

has the matrix E_2. For which $\mathbf{b} \in \mathbb{K}^3$ the system $A\mathbf{x} = \mathbf{b}$ is consistent? ✓

246.* Given a linear map $A : \mathcal{V} \to \mathcal{W}$, its **right inverse** (respectively, **left inverse**) is defined as a linear map $B : \mathcal{W} \to \mathcal{V}$ such that $AB = \mathrm{id}_{\mathcal{W}}$ (respectively, $BA = \mathrm{id}_{\mathcal{V}}$), where $\mathrm{id}_{\mathcal{V}}$ and $\mathrm{id}_{\mathcal{W}}$ denote the identity transformations on $\mathcal{V}$ and $\mathcal{W}$. Prove that a right (left) inverse to A exists if and only if $\operatorname{rk} A = \dim \mathcal{W}$ ($\operatorname{rk} A = \dim \mathcal{V}$), and that neither is unique unless $\dim V = \dim W$.

247. Prove that $\operatorname{rk} A$ does not change if to a row of A a linear combination of other rows of A is added.

Adjoint Maps

Recall from Section 1 of Chapter 2 that to a linear map $A : \mathcal{V} \to \mathcal{W}$, one can associate its **adjoint map**, acting between *dual* spaces in the opposite direction: $A^t : \mathcal{W}^* \to \mathcal{V}^*$. Namely, to a linear function $\mathcal{W} \xrightarrow{\mathbf{a}} \mathbb{K}$, the adjoint map A^t assigns the composition $\mathcal{V} \xrightarrow{A} \mathcal{W} \xrightarrow{\mathbf{a}} \mathbb{K}$, i.e.:

$$(A^t\mathbf{a})(\mathbf{x}) = \mathbf{a}(A\mathbf{x}) \quad \text{for all } \mathbf{x} \in \mathcal{V} \text{ and } \mathbf{a} \in \mathcal{W}^*.$$

In coordinates, suppose that $A : \mathbb{K}^n \to \mathbb{K}^m$ is given by m linear functions in n variables:

$$\begin{aligned} y_1 &= a_{11}x_1 + \cdots + a_{1n}x_n \\ &\cdots \\ y_m &= a_{m1}x_1 + \cdots + a_{mn}x_n. \end{aligned}$$

These equalities show that the elements y_i of the dual basis in $(\mathbb{K}^m)^*$ are mapped by A^t to the linear combinations $a_{i1}x_1 + \cdots + a_{in}x_n$ of the elements x_j of the dual basis in $(\mathbb{K}^n)^*$. Therefore *columns* of the matrix representing the map A^t in these bases are *rows* of A. Thus, ***matrices of adjoint maps with respect to dual bases are transposed to each other***.

`Corollary 1.` ***Adjoint linear maps have the same rank.***

Indeed, when a map $A : \mathcal{V} \to \mathcal{W}$ has the matrix E_r in suitable bases of $\mathcal{V}$ and $\mathcal{W}$, the map A^t has the matrix E_r^t in respectively dual bases of $\mathcal{W}^*$ and $\mathcal{V}^*$. Thus $\operatorname{rk} A^t = \operatorname{rk} E_r^t = r$.

`Remark.` Here is a more geometric way to understand this fact. According to the homomorphism theorem, the range of $A : \mathcal{V} \to \mathcal{W}$ is a subspace in $\mathcal{W}$ canonically isomorphic to $\mathcal{V}/\operatorname{Ker} A$. The range of A^t is exactly the dual space $(\mathcal{V}/\operatorname{Ker} A)^*$ considered as the subspace in $\mathcal{V}^*$ which consists of all those linear functions on $\mathcal{V}$ that vanish on $\operatorname{Ker} A$. Since dual spaces have the same dimension, we conclude once again that $\operatorname{rk} A = \operatorname{rk} A^t$.

The range of the linear map $A : \mathbb{K}^n \to \mathbb{K}^m$ is spanned by columns of the matrix A. Therefore the rank of A is equal to the maximal number of linearly independent columns of A.

`Corollary 2.` ***The maximal number of linearly independent rows of a matrix is equal to the maximal number of linearly independent columns of it.***

Ranks and Determinants

`Corollary 3.` ***The rank of a matrix is equal to the maximal size k for which there exists a $k \times k$-submatrix with a non-zero determinant.***

`Proof.` Suppose that a given $m \times n$-matrix has rank r. Then there exists a set of r linearly independent columns of it. These columns form an $m \times r$-matrix of rank r. Therefore there exists a set of r linearly independent rows of it. These rows form an $r \times r$ matrix M of rank r. By Corollary of the Rank Theorem, this matrix can be written as the product $M = DE_rC^{-1}$ where D and C are invertible $r \times r$-matrices, and $E_r = I_r$ is the identity matrix of size r. Since invertible matrices have non-zero determinants, we conclude that $\det M = (\det D)/(\det C) \neq 0$.

On the other hand, let M' be a $k \times k$-submatrix of the given matrix, such that $k > r$. Then columns of M' are linearly dependent. Therefore one of them can be represented as a linear combination of the others. Since determinants don't change when from one of the columns, a linear combination of other columns is subtracted, we conclude that $\det M' = 0$.

EXERCISES

248. Find the rank of the following matrix, using determinants:

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

249. Find all values of λ for which the matrix $\begin{bmatrix} \lambda & 1 \\ 2 & \lambda+1 \end{bmatrix}$ has rank 1.

250.* To a 2-dimensional subspace in $\mathbb{K}^4$, associate a 1-dimensional subspace in $\mathbb{K}^6$ in the following way. Take the 4×2-matrix whose columns represent a basis in the subspace, and form a 6-column from the 2×2-minors of that matrix. (These minors are called **Plücker coordinates** of the plane.) Prove that the column is non-zero, and that the 1-dimensional subspace spanned by it does not change when another basis is chosen in the same plane.

Systems of Linear Equations — Theory

Let $A\mathbf{x} = \mathbf{b}$ be a system of m linear equations in n unknowns $\mathbf{x}$ with the coefficient matrix A and the right hand side $\mathbf{b}$. Let $r = \operatorname{rk} A$.

`Corollary 4.` (1) ***The solution set to the*** **homogeneous system** $A\mathbf{x} = \mathbf{0}$ ***is a linear subspace in*** $\mathbb{K}^n$ ***of dimension*** $n - r$ (namely, the kernel of $A : \mathbb{K}^n \to \mathbb{K}^m$).

(2) ***The system*** $A\mathbf{x} = \mathbf{b}$ ***is*** **consistent** (i.e. has at least one solution) ***only when*** $\mathbf{b}$ ***lies in a certain subspace of dimension*** r (namely, in the range of A).

(3) ***When it does, the solution set to the system*** $A\mathbf{x} = \mathbf{b}$ ***is an affine subspace in*** $\mathbb{K}^n$ ***of dimension*** $n-r$ ***parallel to the kernel of*** A***.***

Indeed, this is obviously true in the special case $A = E_r$, and is therefore true in general due to the Rank Theorem.

A subspace (affine or linear) of dimension $n - r$ in a space of dimension n is said to have **codimension** r. Thus, rephrasing Corollary 4, we can say that the solution space to a system $A\mathbf{x} = \mathbf{b}$ is either empty (when the column $\mathbf{b}$ does not lie in a subspace spanned by columns of A), or is an affine subspace of codimension r parallel to the solution spaces of the corresponding homogeneous system $A\mathbf{x} = \mathbf{0}$. One calls the rank r of the matrix A also the **rank of the system** $A\mathbf{x} = \mathbf{b}$.

Consider now the case when the number of equations is equal to the number of unknowns.

Corollary 5. ***When*** $\det A \neq 0$***, the linear system*** $A\mathbf{x} = \mathbf{b}$ ***of*** n ***linear equations with*** n ***unknowns has a unique solution for every*** $\mathbf{b}$***, and when*** $\det A = 0$***, solutions are non-unique for some (but not all)*** $\mathbf{b}$ ***and do not exist for all others.***

Indeed, when $\det A \neq 0$, the matrix A is invertible, and $\mathbf{x} = A^{-1}\mathbf{b}$ is the unique solution. When $\det A = 0$, the rank r of the system is smaller than n. Then the range of A has positive codimension, and the kernel has positive dimension, both equal to $n - r$ (Figure 33).

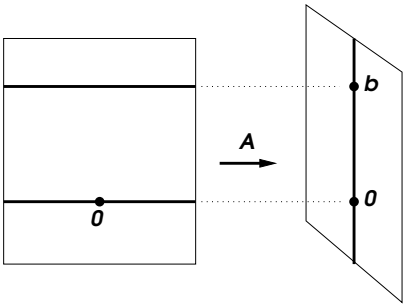

Figure 33

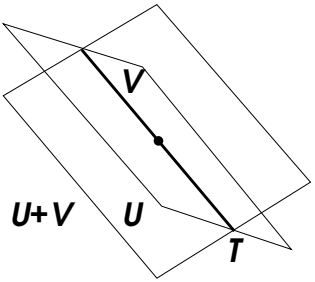

Figure 34

EXERCISES

251. Suppose that a system $A\mathbf{x} = \mathbf{b}$ of m linear equations in 2014 unknowns has a unique solution for $\mathbf{b} = (1, 0, \dots, 0)^t$. Does this imply that: (a) $\operatorname{Ker} A = \{\mathbf{0}\}$, (b) $\operatorname{rk} A = 2014$, (c) $m \geq 2014$, (d) A^{-1} exists, (e) $A^t A$ is invertible, (f) $\det(AA^t) \neq 0$, (g) rows of A are linearly independent, (h) columns of A are linearly independent? ✓

252.* Given $A : \mathbb{K}^n \to \mathbb{K}^m$, consider two **adjoint systems**: $A\mathbf{x} = \mathbf{f}$ and $A^t\mathbf{a} = \mathbf{b}$. Prove that one of them (say, $A\mathbf{x} = \mathbf{f}$) is consistent for a given right hand side vector ($\mathbf{f} \in \mathbb{K}^m$) if and only if this vector is annihilated (i.e. $\mathbf{a}(\mathbf{f}) = 0$) by all linear functions ($\mathbf{a} \in (\mathbb{K}^m)^*$) satisfying the adjoint homogeneous system ($A^t\mathbf{a} = \mathbf{0}$). ϟ

Dimension Counting

In 3-space, two distinct planes intersect in a line, and a line meets a plane at a point. How do these geometric statements generalize to higher dimensions?

It follows from the Rank Theorem, that dimensions of the range and kernel of a linear map add up to the dimension of the domain space. We will use this fact to answer the above question.

Corollary 6. ***If linear subspaces of dimensions k and l span together a subspace of dimension n, then their intersection is a linear subspace of dimension $k+l-n$.***

Proof. Let $\mathcal{U}, \mathcal{V} \subset \mathcal{W}$ be linear subspaces of dimensions k and l in a vector space $\mathcal{W}$, and $\mathcal{T} = \mathcal{U} \cap \mathcal{V}$ be their intersection (Figure 34). Denote by $\mathcal{U}+\mathcal{V} \subset \mathcal{W}$ the subspace of dimension n spanned by vectors of $\mathcal{U}$ and $\mathcal{V}$. Define a linear map $A : \mathcal{U} \oplus \mathcal{V} \to \mathcal{W}$, where $\mathcal{U} \oplus \mathcal{V} = \{(\mathbf{u}, \mathbf{v}) | \mathbf{u} \in \mathcal{U},\ \mathbf{v} \in \mathcal{V}\}$ is the direct sum, by $A(\mathbf{u}, \mathbf{v}) = \mathbf{u} - \mathbf{v}$. The range of A coincides with $\mathcal{U}+\mathcal{V}$. The kernel of A consists of all those pairs $(\mathbf{u}, \mathbf{v})$, where $\mathbf{u} \in \mathcal{U}$ and $\mathbf{v} \in \mathcal{V}$, for which $\mathbf{u} = \mathbf{v}$. Therefore $\operatorname{Ker} A = \{(\mathbf{t}, \mathbf{t}) | \mathbf{t} \in \mathcal{T}\} \cong \mathcal{T}$. Thus $\dim(\mathcal{U}+\mathcal{V}) + \dim \mathcal{T} = \dim(\mathcal{U} \oplus \mathcal{V}) = \dim \mathcal{U} + \dim \mathcal{V}$. We conclude that $\dim \mathcal{T} = k + l - n$.

EXERCISES

253. Prove that two affine planes lying in a vector space are contained in an affine subspace of dimension ≤ 5.

254. The solution set of a single non-trivial linear equation $\mathbf{a}(\mathbf{x}) = b$ is called a **hyperplane** (affine if $b \neq 0$ and linear if $b = 0$). Show that a hyperplane is an (affine or linear) subspace of codimension 1.

255. Find possible codimensions of intersections of k linear hyperplanes. ✓

256. Prove that every subspace in $\mathbb{K}^n$ can be described as: (a) the range of a linear map; (b) the kernel of a linear map.

257. Classify linear subspaces in $\mathbb{K}^n$ up to linear transformations of $\mathbb{K}^n$. ✓

258.* Classify *pairs* of subspaces in $\mathbb{K}^n$ up to linear transformations. ✓

259.* Let $\mathbb{K}$ be a *finite* field of q elements. Compute the number of: (a) vectors in a $\mathbb{K}$-vector space of dimension n, (b) bases in $\mathbb{K}^n$, (c) $n \times r$-matrices of rank r, (d) subspaces of dimension r in $\mathbb{K}^n$. ✓

2 Gaussian Elimination

Evaluating the determinant of a 20×20-matrix directly from the definition of determinants requires 19 multiplications for each of the $20! > 2 \cdot 10^{18}$ elementary products. On a typical PC that makes 1 *gigaflops* (i.e. 10^9 **FL**oating point arithmetic **O**perations **P**er **S**econd), this would take about $4 \cdot 10^{10}$ seconds, which is a little longer than 1000 years. Algorithms based on Gaussian elimination allow your PC to evaluate much larger determinants in tiny fractions of a second.

Row Reduction

Usually, solving a system of linear algebraic equations with coefficients given numerically we, using one of the equations, express the 1st unknown via the other unknowns and eliminate it from the remaining equations, then express the 2nd unknown from one of the remaining equations, etc., and finally arrive to an equivalent algebraic system which is easy to solve starting from the last equation and working backward. This computational procedure called **Gaussian elimination** can be conveniently organized as a sequence of operations with rows of the coefficient matrix of the system. Namely, we use three **elementary row operations**:

- transposition of two rows;
- division of a row by a non-zero scalar;
- subtraction of a multiple of one row from another one.

`Example 1.` Solving the system

$$\begin{array}{rcrcrcr} & & x_2 & + & 2x_3 & = & 3 \\ 2x_1 & + & 4x_2 & & & = & -2 \\ 3x_1 & + & 5x_2 & + & x_3 & = & 0 \end{array}$$

by Gaussian elimination, we pull the 2nd equation up (since the 1st equation does not contain x_1), divide it by 2 (in order to express x_1 via x_2) and subtract it 3 times from the 3rd equation in order to get rid of x_1 therein. Then we use the 1st equation (which has become the 2nd one in our pile) in order to eliminate x_2 from the 3rd equation. The coefficient matrix of the system is subject to the elementary row transformations:

$$\left[\begin{array}{ccc|c} 0 & 1 & 2 & 3 \\ 2 & 4 & 0 & -2 \\ 3 & 5 & 1 & 0 \end{array}\right] \mapsto \left[\begin{array}{ccc|c} 2 & 4 & 0 & -2 \\ 0 & 1 & 2 & 3 \\ 3 & 5 & 1 & 0 \end{array}\right] \mapsto \left[\begin{array}{ccc|c} 1 & 2 & 0 & -1 \\ 0 & 1 & 2 & 3 \\ 3 & 5 & 1 & 0 \end{array}\right]$$

$$\mapsto \left[\begin{array}{ccc|c} 1 & 2 & 0 & -1 \\ 0 & 1 & 2 & 3 \\ 0 & -1 & 1 & 3 \end{array}\right] \mapsto \left[\begin{array}{ccc|c} 1 & 2 & 0 & -1 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 3 & 6 \end{array}\right] \mapsto \left[\begin{array}{ccc|c} 1 & 2 & 0 & -1 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 2 \end{array}\right].$$

The final "triangular" shape of the coefficient matrix is an example of the **row echelon form**. If read from bottom to top, it represents the system $x_3 = 2,\ x_2 + 2x_3 = 3,\ x_1 + 2x_2 = -1$ which is ready to be solved by **back substitution**: $x_3 = 2,\ x_2 = 3 - 2x_3 = 3 - 4 = -1,\ x_1 = -1 - 2x_2 = -1 + 2 = 1$. The process of back substitution, expressed in the matrix form, consists of a sequence of elementary row operations of the third type:

$$\left[\begin{array}{ccc|c} 1 & 2 & 0 & -1 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 2 \end{array}\right] \mapsto \left[\begin{array}{ccc|c} 1 & 2 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 2 \end{array}\right] \mapsto \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 2 \end{array}\right].$$

The last matrix is an example of the **reduced row echelon form** and represents the system $x_1 = 1,\ x_2 = -1,\ x_3 = 2$ which is "already solved".

In general, Gaussian elimination is an algorithm of reducing an **augmented matrix** to a row-echelon form by means of elementary row operations. By an augmented matrix we mean simply a matrix subdivided into two blocks $[A|B]$. The augmented matrix of a linear system $A\mathbf{x} = \mathbf{b}$ in n unknowns is $[\mathbf{a}_1, ..., \mathbf{a}_n|\mathbf{b}]$ where $\mathbf{a}_1, ..., \mathbf{a}_n$ are columns of A, but we will also make use of augmented matrices with B consisting of several columns. Operating with a row $[a_1, ..., a_n|b_1, ...]$ of augmented matrices we will refer to the leftmost non-zero entry among a_j as the **leading entry**[2] of the row. We say that the augmented matrix $[A|B]$ is in the **row echelon form of rank** r if the $m \times n$-matrix A satisfies the following conditions:

- each of the first r rows has the leading entry equal to 1;
- leading entries of the rows $1, 2, ..., r$ are situated respectively in the columns with indices $j_1, ..., j_r$ satisfying $j_1 < j_2 < ... < j_r$;
- all rows of A with indices $i > r$ are zero.

Notice that a matrix in a row echelon form has zero entries everywhere below and to the left of each leading entry. A row echelon form is called **reduced** (Figure 35) if all the entries in the columns $j_1, ..., j_r$ above the leading entries are also equal to zero.

[2]Also called **leading coefficient**, or **pivot**.

If the matrix A of a linear system is in the row echelon form and indeed has one or several zero rows on the bottom, then the system contains equations of the form $0x_1 + ... + 0x_n = b$. If at least one of such b is non-zero, the system is **inconsistent** (i.e. has no solutions). If all of them are zeroes, the system is consistent and ready to be solved by back substitution.

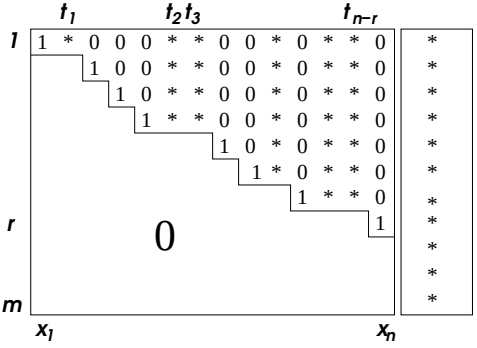

Figure 35

Example 2. The following augmented matrix is in the row echelon form of rank 2:

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

It corresponds to the system $x_1 + 2x_2 + 3x_3 = 0,\ x_3 = 2,\ 0 = 0$. The system is consistent: $x_3 = 2,\ x_2 = t,\ x_1 = -3x_3 - 2x_2 = -6 - 2t$ satisfy the system *for any value of the parameter* t.

We see that presence of leading entries in the columns $j_1, ..., j_r$ of the row echelon form allows one to express the unknowns $x_{j_1}, ..., x_{j_r}$ in terms of the unknowns x_j with $j \neq j_1, ..., j_r$, while the values $t_1, ..., t_{n-r}$ of the unknowns $x_j,\ j \neq j_1, ..., j_r$ remain completely ambiguous. In this case, ***solutions to the linear system depend on the*** $n - r$ ***parameters*** $t_1, ..., t_{n-r}$**.**

The algorithm of row reduction of an augmented matrix $[A|B]$ to the row echelon form can be described by the following instructions. Let n be the number of columns of A. Then the algorithm consists of n steps. At the step $l = 1, ..., n$, we assume that the matrix formed by the columns of A with the indices $j = 1, ..., l-1$ is already in the row echelon form of some rank $s < l$, with the leading entries located in some columns $j_1 < ... < j_s < l$. The l-th step begins with locating

the first non-zero entry in the column l below the row s. If none is found, the l-th step is over, since the columns $1, ..., l$ are already in the row echelon form of rank s. Otherwise the first non-zero entry is located in a row $i(> s)$, and the following operations are performed:

(i) transposing the rows i and $s+1$ of the augmented matrix,

(ii) dividing the whole row $s+1$ of the augmented matrix by the leading entry, which is now $a_{s+1,l}(\neq 0)$,

(iii) annihilating all the entries in the column l below the leading entry of the $s+1$-st row by subtracting suitable multiples of the $s+1$-st row of the augmented matrix from all rows with indices $i > s+1$.

After that, the l-th step is over since the columns $1, ..., l$ are now in the row echelon form of rank $s+1$.

When an augmented matrix $[A|B]$ has been reduced to a row echelon form with the leading entries $a_{1,j_1} = ... = a_{r,j_r} = 1$, the back substitution algorithm, which reduces it further to a reduce row echelon form, consists of r steps which we number by $l = r, r-1, ..., 1$ and perform in this order. On the l-th step, we subtract from each of the rows $i = 1, ..., l-1$ of the augmented matrix, the l-th row multiplied by a_{i,j_l}, and thus annihilate all the entries of the column j_l above the leading entry.

EXERCISES

260. Solve systems of linear equations: ✓

$$\begin{array}{l} 2x_1 - x_2 - x_3 = 4 \\ 3x_1 + 4x_2 - 2x_3 = 11 \\ 3x_1 - 2x_2 + 4x_3 = 11 \end{array} \qquad \begin{array}{l} x_1 + x_2 - 3x_3 = -1 \\ 2x_1 + x_2 - 2x_3 = 1 \\ x_1 + x_2 + x_3 = 3 \\ x_1 + 2x_2 - 3x_3 = 1 \end{array} \qquad \begin{array}{l} 2x_1 + x_2 + x_3 = 2 \\ x_1 + 3x_2 + x_3 = 5 \\ x_1 + x_2 + 5x_3 = -7 \\ 2x_1 + 3x_2 - 3x_3 = 14 \end{array}$$

$$\begin{array}{l} x_1 - 2x_2 + x_3 + x_4 = 1 \\ x_1 - 2x_2 + x_3 - x_4 = -1 \\ x_1 - 2x_2 + x_3 + 5x_4 = 5 \end{array} \qquad \begin{array}{l} x_1 - 2x_2 + 3x_3 - 4x_4 = 4 \\ x_2 - x_3 + x_4 = -3 \\ x_1 + 3x_2 - 3x_4 = 1 \\ -7x_2 + 3x_3 + x_4 = -3 \end{array}$$

$$\begin{array}{l} 2x_1 + 3x_2 - x_3 + 5x_4 = 0 \\ 3x_1 - x_2 + 2x_3 - 7x_4 = 0 \\ 4x_1 + x_2 - 3x_3 + 6x_4 = 0 \\ x_1 - 2x_2 + 4x_3 - 7x_4 = 0 \end{array} \qquad \begin{array}{l} 3x_1 + 4x_2 - 5x_3 + 7x_4 = 0 \\ 2x_1 - 3x_2 + 3x_3 - 2x_4 = 0 \\ 4x_1 + 11x_2 - 13x_3 + 16x_4 = 0 \\ 7x_1 - 2x_2 + x_3 + 3x_4 = 0 \end{array}$$

$$\begin{array}{l} x_1 + x_2 + x_3 + x_4 + x_5 = 7 \\ 3x_1 + 2x_2 + x_3 + x_4 - 3x_5 = -2 \\ x_2 + 2x_3 + 2x_4 + 6x_5 = 23 \\ 5x_1 + 4x_2 + 3x_3 + 3x_4 - x_5 = 12 \end{array} .$$

261.* Find those λ for the system is consistent: ✓

$$\begin{array}{l} 2x_1 - x_2 + x_3 + x_4 = 1 \\ x_1 + 2x_2 - x_3 + 4x_4 = 2 \\ x_1 + 7x_2 - 4x_3 + 11x_4 = \lambda \end{array} .$$

Applications

Row reduction algorithms allow one to compute efficiently determinants and inverses of square matrices given numerically, and to find a basis in the **null space**, **column space** and **row space** of a given rectangular matrix (i.e., speaking geometrically, in the kernel of the matrix, its range, and the range of the transposed matrix).

`Proposition 1.` ***Suppose that an*** $m \times n$***-matrix*** A ***has been reduced by elementary row operations to a row echelon form*** A' ***of rank*** r ***with the leading entries*** $a_{1,j_1} = ... = a_{r,j_r} = 1$***,*** $j_1 < ... < j_r$***. Then***

(1) $\operatorname{rk} A = \operatorname{rk} A' = r$,

(2) ***rows*** $1, \dots, r$ ***of*** A' ***form a basis in the row space of*** A***,***

(3) ***the columns of*** A ***with indices*** $j_1, ..., j_r$ ***form a basis in the column space of*** A***.***

`Proof.` Elementary row operations do not change the space spanned by rows of the matrix. The non-zero rows of a row echelon matrix are linearly independent and thus form a basis in the row space. In particular, $\operatorname{rk} A = \operatorname{rk} A' = r$.

The row operations change columns $\mathbf{a}_1, ..., \mathbf{a}_n$ of the matrix A, but preserve linear dependencies among them: $\alpha_1\mathbf{a}_1 + ... + \alpha_n\mathbf{a}_n = \mathbf{0}$ if and only if $\alpha_1\mathbf{a}'_1 + ... + \alpha_n\mathbf{a}'_n = \mathbf{0}$. The r columns $\mathbf{a}'_{j_1}, ..., \mathbf{a}'_{j_r}$ of the matrix A' in the row echelon form which contain the leading entries are linearly independent. Therefore columns $\mathbf{a}_{j_1}, ..., \mathbf{a}_{j_r}$ of the matrix A are linearly independent too and hence form a basis in the column space of A. □

`Example 3.` The following row reduction

$$\begin{bmatrix} 1 & 2 & 3 & -1 \\ 2 & 4 & 5 & 1 \\ 3 & 6 & 8 & 0 \end{bmatrix} \mapsto \begin{bmatrix} 1 & 2 & 3 & -1 \\ 0 & 0 & -1 & 3 \\ 0 & 0 & -1 & 3 \end{bmatrix} \mapsto \begin{bmatrix} 1 & 2 & 3 & -1 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

shows that the matrix has rank 2, rows $(1,2,3,-1)$, $(0,0,1,-3)$ form a basis in the row space, and columns $(1,2,3)^t$, $(3,5,8)^t$ a basis in the column space.

Suppose that the augmented matrix $[A|\mathbf{b}]$ of the system $A\mathbf{x}=\mathbf{b}$ has been transformed to a *reduced* row echelon form $[A'|\mathbf{b}']$ with the leading entries positioned in the columns $j_1 < j_2 < ... < j_r$. These columns are the unit coordinate vectors $\mathbf{e}_1, ..., \mathbf{e}_r$, and the system is consistent only if $\mathbf{b}'$ is their linear combination, $\mathbf{b}' = b'_1\mathbf{e}_1 + ... + b'_r\mathbf{e}_r$. Assuming that this is the case we can assign arbitrary values $t_1, ..., t_{n-r}$ to the unknowns x_j, $j \neq j_1, ..., j_r$, and express $x_{j_1}, ..., x_{j_r}$ as linear inhomogeneous functions of $t_1, ..., t_{n-r}$. The general solution to the system will have the form $\mathbf{x} = \mathbf{v}_0 + t_1\mathbf{v}_1 + ... + t_{n-r}\mathbf{v}_{n-r}$ of a linear combination of some n-dimensional vectors $\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_{n-r}$. We claim that ***the vectors*** $\mathbf{v}_1, ..., \mathbf{v}_{n-r}$ ***form a basis in the null space of the matrix*** A. Indeed, substituting $t=0$ we conclude that $\mathbf{v}_0$ satisfies the equation $A\mathbf{v}_0 = \mathbf{b}$. Therefore $\mathbf{x}-\mathbf{v}_0 = t_1\mathbf{v}_1+...+t_{n-r}\mathbf{v}_{n-r}$ form the general solution to the homogeneous system $A\mathbf{x}=\mathbf{0}$, i.e. the null space of A. In addition, we see that ***the solution set to the inhomogeneous system is the affine subspace in*** $\mathbb{K}^n$ ***obtained from the null space by the translation through the vector*** $\mathbf{v}_0$.

`Example 4.` Consider the system $A\mathbf{x}=\mathbf{0}$ with the matrix A from Example 3. Transform the matrix to the reduced row echelon form:

$$... \mapsto \begin{bmatrix} 1 & 2 & 3 & -1 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \mapsto \begin{bmatrix} 1 & 2 & 0 & 8 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The general solution to the system assumes the form

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -2t_1 - 8t_2 \\ t_1 \\ 3t_2 \\ t_2 \end{bmatrix} = t_1 \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t_2 \begin{bmatrix} -8 \\ 0 \\ 3 \\ 1 \end{bmatrix}.$$

The columns $(-2,1,0,0)^t$ and $(-8,0,3,1)^t$ form therefore a basis in the null space of the matrix A.

`Proposition 2.` ***Suppose that in the process of row reduction of an*** $n \times n$ ***matrix*** A ***to a row echelon form*** A' ***row transpositions occurred*** k ***times, and the operations of division by leading entries*** $\alpha_1, ..., \alpha_r$ ***were performed. If*** $\operatorname{rk} A' < n$ ***then*** $\det A = 0$. ***If*** $\operatorname{rk} A' = n$ ***then*** $\det A = (-1)^k \alpha_1 ... \alpha_n$.

Indeed, each transposition of rows reverses the sign of the determinant, divisions of a row by α divides the determinant by α, and subtraction of a multiple of one row from another one does not change the determinant. Thus $\det A = (-1)^k \alpha_1 ... \alpha_r \det A'$. The row echelon matrix is upper triangular. When $\operatorname{rk} A' = n$, it has n leading 1's on the diagonal, and hence $\det A' = 1$. When $r < n$ we have $\det A' = 0$.

`Proposition 3.` ***Given an*** $n \times n$***-matrix*** A***, introduce the augmented matrix*** $[A|I_n]$ (where I_n is the identity matrix) ***and transform it to the reduced row-echelon form*** $[A'|B]$ ***by elementary row operations. If*** $A' = I_n$ ***then*** $B = A^{-1}$***.***

Indeed, the equality $A' = I_n$ means that $\operatorname{rk} A = n$ and thus A^{-1} exists. Then the system $A\mathbf{x} = \mathbf{b}$ has a unique solution for any $\mathbf{b}$, and for $\mathbf{b} = \mathbf{e}_1, ..., \mathbf{e}_n$ the corresponding solutions $\mathbf{x} = A^{-1}\mathbf{e}_1, ..., A^{-1}\mathbf{e}_n$ are the columns of the inverse matrix A^{-1}. These solutions can be found by simultaneous row reduction of the augmented matrices $[A|\mathbf{e}_1], ..., [A|\mathbf{e}_n]$ and thus coincide with the columns of the matrix B in the reduced row-echelon form $[I_n|B]$.

`Example 5.` Let us compute $\det A$ and A^{-1} for the matrix of Example 1. We have:

$$\left[\begin{array}{ccc|ccc} 0 & 1 & 2 & 1 & 0 & 0 \\ 2 & 4 & 0 & 0 & 1 & 0 \\ 3 & 5 & 1 & 0 & 0 & 1 \end{array}\right] \mapsto \left[\begin{array}{ccc|ccc} 1 & 2 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 2 & 1 & 0 & 0 \\ 3 & 5 & 1 & 0 & 0 & 1 \end{array}\right] \mapsto$$

$$\left[\begin{array}{ccc|ccc} 1 & 2 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 2 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 & -\frac{3}{2} & 1 \end{array}\right] \mapsto \left[\begin{array}{ccc|ccc} 1 & 2 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & \frac{1}{3} & -\frac{1}{2} & \frac{1}{3} \end{array}\right].$$

Here one transposition of rows and divisions by 2 and by 3 were applied. Thus $\det A = (-1) \cdot 2 \cdot 3 = -6$, and the matrix is invertible. Back substitution eventually yields the inverse matrix:

$$\mapsto \left[\begin{array}{ccc|ccc} 1 & 2 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & \frac{1}{3} & 1 & -\frac{2}{3} \\ 0 & 0 & 1 & \frac{1}{3} & -\frac{1}{2} & \frac{1}{3} \end{array}\right] \mapsto \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -\frac{2}{3} & -\frac{3}{2} & \frac{4}{3} \\ 0 & 1 & 0 & \frac{1}{3} & 1 & -\frac{2}{3} \\ 0 & 0 & 1 & \frac{1}{3} & -\frac{1}{2} & \frac{1}{3} \end{array}\right].$$

`Remark.` Gaussian elimination algorithms are unlikely to work well for matrices depending on parameters. To see why, try row reduction in order to solve a linear system of the form $(\lambda I - A)\mathbf{x} = \mathbf{0}$ depending on the parameter λ, or (even better!) apply Gaussian

elimination to the system $a_{11}x_1 + a_{12}x_2 = b_1$, $a_{21}x_1 + a_{22}x_2 = b_2$ depending on 6 parameters $a_{11}, a_{12}, a_{21}, a_{22}, b_1, b_2$.

EXERCISES

262. For each of the following matrices, find the rank and bases in the null, column, and row spaces: ✓

$$(a)\begin{bmatrix} 0 & 4 & 10 & 1 \\ 4 & 8 & 18 & 7 \\ 10 & 18 & 40 & 17 \\ 1 & 7 & 17 & 3 \end{bmatrix} \quad (b)\begin{bmatrix} 14 & 12 & 6 & 8 & 2 \\ 6 & 104 & 21 & 9 & 17 \\ 7 & 6 & 3 & 4 & 1 \\ 35 & 30 & 15 & 20 & 5 \end{bmatrix}$$

$$(c)\begin{bmatrix} 1 & 0 & 0 & 1 & 4 \\ 0 & 1 & 0 & 2 & 5 \\ 0 & 0 & 1 & 3 & 6 \\ 1 & 2 & 3 & 14 & 32 \\ 4 & 5 & 6 & 32 & 77 \end{bmatrix} \quad (d)\begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 \\ 1 & 1 & 4 & 1 \\ 1 & 1 & 1 & 5 \\ 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad (e)\begin{bmatrix} 2 & 1 & 3 & -1 \\ 3 & -1 & 2 & 0 \\ 1 & 3 & 4 & -2 \\ 4 & -3 & 1 & 1 \end{bmatrix}.$$

263. For each of the following matrices, compute the determinant and the inverse matrix, and an LPU decomposition: ✓

$$(a)\begin{bmatrix} 2 & 2 & -3 \\ 1 & -1 & 0 \\ -1 & 2 & 1 \end{bmatrix} \quad (b)\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \quad (c)\begin{bmatrix} 2 & 1 & 0 & 0 \\ 3 & 2 & 0 & 0 \\ 1 & 1 & 3 & 4 \\ 2 & -1 & 2 & 3 \end{bmatrix}.$$

LPU Decomposition

Gaussian elimination is an algorithm. The fact that it always works is a theorem. In this and next subsections, we reformulate this theorem (or rather an important special case of it) first in the language of matrix algebra, and then in geometric terms.

Recall that a square matrix A is called **upper triangular** (respectively **lower triangular**) if $a_{ij} = 0$ for all $i > j$ (respectively $i < j$). We call P a **permutation matrix** if it is obtained from the identity matrix I_n by a permutation of columns. Such P is indeed the matrix of a linear transformation in $\mathbb{K}^n$ defined as the permutation of coordinate axes.

`Theorem.` ***Every invertible matrix M can be factored as the product $M = LPU$ of a lower triangular matrix L, a permutation matrix P, and an upper triangular matrix U.***

The proof of this theorem is based on interpretation of elementary row operations in terms of matrix multiplication. Consider the following $m \times m$-matrices:

• T_{ij} $(i \neq j)$, a **transposition matrix**, obtained by transposing the i-th and j-th columns of the identity matrix;

• $D_i(d)$ $(d \neq 0)$, a diagonal matrix, all of whose diagonal entries are equal to 1 except the i-th one, which is equal to $1/d$;

• $L_{ij}(\alpha)$ $(i > j)$, a lower triangular matrix, all of whose diagonal entries are equal to 1, and all off-diagonal equal to 0 except the entry in i-th row and j-th column, which is equal to $-\alpha$.

Here are examples T_{13}, $D_2(3)$, and $L_{42}(-2)$ of size $m = 4$:

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \end{bmatrix}.$$

Elementary row operations on a given $m \times n$-matrix can be described as multiplication on the left by T_{ij}, $D_i(d)$, or $L_{ij}(\alpha)$, which results respectively in transposing the i-th and j-th rows, dividing the i-th row by d, and subtracting the j-th row times α from the i-th row. Note that inverses of elementary row operations are also elementary row operations. Thus, Gaussian elimination allows one to represent any matrix A as the product of a row echelon matrix with matrices of elementary row operations. In order to prove the LPU decomposition theorem, we will combine this idea with a modification of the Gaussian elimination algorithm.

Let M be an *invertible* $n \times n$-matrix. We apply the row reduction process to it, temporarily refraining from using permutations and divisions of rows and using only the row operations equivalent to left multiplication by $L_{kl}(\alpha)$. On the i-th step of the algorithm, if the i-th row does not contain a non-zero entry where expected, we don't swap it with the next row. Instead, we locate in *this* row the leading (i.e. leftmost non-zero) entry, which must exist since the matrix is invertible. When it is found in a column j, we subtract multiples of the row i from rows $i+1, \ldots, n$ with such coefficients that all entries of these rows in the column j become annihilated.

`Example 7.` Let us illustrate the modified row reduction with the matrix taken from Example 1, and at the same time represent the process as matrix factorization. On the first step, we subtract the 1st row from the 2nd and 3rd 4 and 5 times respectively, and on the second step subtract the 2nd row times $\frac{3}{2}$ from the 3rd. The

lower triangular factors shown are *inverses* of the matrices $L_{21}(4)$, $L_{31}(5)$ and $L_{32}(\frac{3}{2})$. The leading entries are boldfaced:

$$\begin{bmatrix} 0 & \mathbf{1} & 2 \\ 2 & 4 & 0 \\ 3 & 5 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 5 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & \mathbf{1} & 2 \\ \mathbf{2} & 0 & -8 \\ 3 & 0 & -9 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 5 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{3}{2} & 1 \end{bmatrix} \begin{bmatrix} 0 & \mathbf{1} & 2 \\ \mathbf{2} & 0 & -8 \\ 0 & 0 & \mathbf{3} \end{bmatrix}.$$

Products of lower triangular matrices are lower triangular. As a result, applying the modified row reduction we obtain a factorization of the form $M = LM'$, where L is a lower triangular matrix with all diagonal entries equal to 1. Note that on the ith step of the algorithm, when a leading entry in the row i is searched, the leading entries of the previous rows can be in any columns $j_1, \dots, j_{i-1}$. The entries of the i-th row situated in these columns have been annihilated at the previous steps. Therefore the leading entry of the i-th row will be found in a new column. This shows that the columns $j_1, \dots, j_n$ of the leading entries found in the rows $1, \dots, n$ are all distinct and thus form a permutation of $\{1, \dots, n\}$. We now write $M' = PU$, where P is the matrix of the permutation $\left(\begin{smallmatrix} 1,\dots,n \\ j_1,\dots,j_n \end{smallmatrix}\right)$. The operation $M' \mapsto U = P^{-1}M'$ permutes *rows* of M' and places all leading entries of the resulting matrix U on the diagonal. Here is how this works in our example:

$$\begin{bmatrix} 0 & \mathbf{1} & 2 \\ \mathbf{2} & 0 & -8 \\ 0 & 0 & \mathbf{3} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{2} & 0 & -8 \\ 0 & \mathbf{1} & 2 \\ 0 & 0 & \mathbf{3} \end{bmatrix}.$$

Since leading entries are leftmost non-zero entries in their rows, the matrix U turns out to be *upper* triangular. This completes the proof.

`Remarks.` (1) As it is seen from the proof, the matrix L in the LPU factorization can be required to have all diagonal entries equal to 1. Triangular matrices with this property are called **unipotent**. Alternatively, one can require that U is unipotent.

(2) When the permutation matrix $P = I_n$, one obtains the **LU decomposition**, $M = LU$, of *some* invertible matrices. Which ones? One can work backward and consider products LDU of a lower and upper triangular unipotent matrices L and U with an invertible diagonal matrix D in between (the so called **LDU decomposition**).

A choice of such matrices depends on a number of arbitrary parameters: $n(n-1)/2$ for each L and U, and n non-zero parameters for D, i.e. totally n^2. This is equal to the total number of matrix elements, suggesting that a *typical* $n \times n$-matrix admits an LDU factorization.

(3) This heuristic claim can be made precise. As illustrated by the above example of LPU decomposition, when $P \neq I_n$, certain entries of the resulting factor U come out equal to 0. This is because some entries of the matrix M' on the *right* of leading ones are annihilated at previous steps of row reduction. As a result, such decompositions involve fewer than n^2 arbitrary parameters, and hence cover a positive codimension locus in the matrix space. Thus, the bulk of the space is covered by factorizations LPU with $P = I_n$.

EXERCISES264. Prove that $D_i^{-1}(d) = D_i(d^{-1})$ and $L_{ij}^{-1}(\alpha) = L_{ij}(-\alpha)$.

265. Prove that the inverse of a permutation matrix P is P^t.

266. Prove that every invertible matrix M has an **LUP decomposition** $M = LUP$ where L is lower triangular, U upper triangular, and P is a permutation matrix, and compute such factorizations for the matrices from the above exercise on LPU decomposition. ϟ

267. Prove that every invertible matrix M has an **PLU decomposition** $M = PLU$. ϟ

268. Prove that every invertible matrix has factorizations of the form UPL, PUL, and ULP, where L, U, and P stand for lower triangular, upper triangular, and permutation matrices respectively. ϟ

Flags and Bruhat cells

A sequence $\mathcal{V}_1 \subset \mathcal{V}_2 \subset \cdots \subset \mathcal{V}_n$ of nested subspaces is said to form a **flag** in the space $\mathcal{V} = \mathcal{V}_n$. When $\dim \mathcal{V}_k = k$ for all $k = 1, \ldots, n$, the flag is called **complete**.

Given a basis $\mathbf{f}_1, \ldots, \mathbf{f}_n$ in $\mathcal{V}$, one can associate to it the **standard coordinate flag** (Figure 36)

$$\mathrm{Span}(\mathbf{f}_1) \subset \mathrm{Span}(\mathbf{f}_1, \mathbf{f}_2) \subset \cdots \subset \mathrm{Span}(\mathbf{f}_1, \ldots, \mathbf{f}_n) = \mathcal{V}.$$

Let $U : \mathcal{V} \to \mathcal{V}$ be an invertible linear transformation *preserving* the standard coordinate flag. Then the matrix of U in the given basis is upper triangular. Indeed, since $U(\mathcal{V}_k) \subset \mathcal{V}_k$, the vector $U\mathbf{f}_k$ is a linear combination of $\mathbf{f}_1, \ldots, \mathbf{f}_k$, i.e. $U\mathbf{f}_k = \sum_{i \leq k} u_{ik}\mathbf{f}_i$. Since this is true for all k, the matrix $[u_{ik}]$ is upper triangular. Reversing this argument, we find that if the matrix of U is upper triangular, then U preserves the flag.

Conversely, given a complete flag $\mathcal{V}_1 \subset \cdots \subset \mathcal{V}_n = \mathbb{K}^n$, one can pick a basis $\mathbf{f}_1$ in the line $\mathcal{V}_1$, then complete it to a basis $\mathbf{f}_1, \mathbf{f}_2$ in the plane $\mathcal{V}_2$, and so on, until a basis $\mathbf{f}_1, \dots, \mathbf{f}_n$ in the whole space $\mathcal{V}_n$ is obtained, such that $\mathcal{V}_k = \mathrm{Span}(\mathbf{f}_1, \dots, \mathbf{f}_k)$ for each k. This shows that ***every complete flag in*** $\mathbb{K}^n$ ***can be obtained from any other by an invertible linear transformation.*** Indeed, the linear transformation, defined in terms of the standard basis $\mathbf{e}_1, \dots, \mathbf{e}_n$ of $\mathbb{K}^n$ by $\sum x_i \mathbf{e}_i \mapsto \sum x_i \mathbf{f}_i$, transforms the standard coordinate flag $\mathrm{Span}(\mathbf{e}_1) \subset \cdots \subset \mathrm{Span}(\mathbf{e}_1, \dots, \mathbf{e}_n)$ into the given flag $\mathcal{V}_1 \subset \cdots \subset \mathcal{V}_n$.

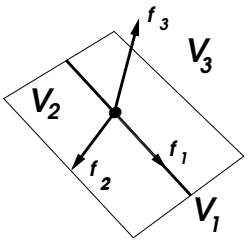

Figure 36

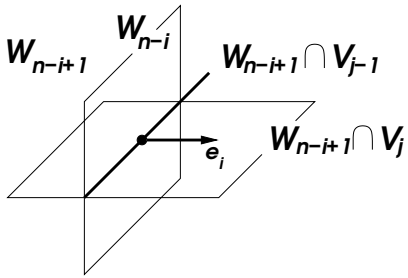

Figure 37

Example 8. Let $P_\sigma : \mathbb{K}^n \to \mathbb{K}^n$ act by a permutation $\sigma = \binom{1\ \dots\ n}{i_1\ \dots\ i_n}$ of coordinate axes. It transforms the *standard* coordinate flag into *a* **coordinate flag**

$$\mathbf{F}_\sigma : \mathrm{Span}(\mathbf{e}_{i_1}) \subset \mathrm{Span}(\mathbf{e}_{i_1}, \mathbf{e}_{i_2}) \subset \cdots \subset \mathrm{Span}(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_n}) = \mathbb{K}^n,$$

called so because all the spaces are spanned by vectors of the standard basis. There are $n!$ such flags, one for each permutation. For instance, $\mathbf{F}_{\mathrm{id}}$ is the standard coordinate flag. When $\sigma = \binom{1\ \dots\ n}{n\ \dots\ 1}$, the flag **opposite** to the standard one is obtained:

$$\mathrm{Span}(\mathbf{e}_n) \subset \mathrm{Span}(\mathbf{e}_n, \mathbf{e}_{n-1}) \subset \cdots \subset \mathrm{Span}(\mathbf{e}_n, \dots, \mathbf{e}_1).$$

Transformations of $\mathbb{K}^n$ defined by *lower* triangular matrices are exactly those that preserve this flag.

Theorem. ***Every complete flag in*** $\mathbb{K}^n$ ***can be transformed into exactly one of the*** $n!$ ***coordinate flags by invertible linear transformations preserving one of them (e.g. the flag opposite to the standard one).***

Proof. Let $\mathbf{F}$ be a given complete flag in $\mathbb{K}^n$, $\mathbf{F}_{\mathrm{id}}$ the standard coordinate flag, M a linear transformation such that $\mathbf{F} = M(\mathbf{F}_{\mathrm{id}})$,

and $M = LP_\sigma U$ its LPU decomposition. Since $U(\mathbf{F}_{\mathrm{id}}) = \mathbf{F}_{\mathrm{id}}$, and $P_\sigma(\mathbf{F}_{\mathrm{id}}) = \mathbf{F}_\sigma$, we find that $\mathbf{F} = L(\mathbf{F}_\sigma)$. Therefore the given flag $\mathbf{F}$ is transformed into a coordinate flag $\mathbf{F}_\sigma$ by L^{-1}, which is lower triangular and thus preserves the flag opposite to $\mathbf{F}_{\mathrm{id}}$.

It remains to show that the same flag $\mathbf{F}$ cannot be transformed this way into two different coordinate flags $\mathbf{F}_\sigma$. Let $\mathcal{V}_j$, $\dim \mathcal{V}_j = j$, denote the spaces of the flag $\mathbf{F}$, and $\mathcal{W}_{n-i} = \mathrm{Span}(\mathbf{e}_n, \dots, \mathbf{e}_{i+1})$ the spaces of the flag opposite to the standard one, $\operatorname{codim} \mathcal{W}_{n-i} = i$. Invertible transformations preserving the spaces $\mathcal{W}_{n-i}$ can change their intersections with $\mathcal{V}_j$, but cannot change the dimensions of these intersections. Thus it suffices to show that, in the case of the flag $\mathbf{F}_\sigma$, the permutation σ is uniquely determined by these dimensions.

Note that $\mathbf{e}_i \in \mathcal{W}_{n-i+1}$ but $\mathbf{e}_i \notin \mathcal{W}_{n-i}$ (Figure 37). Suppose that in the flag $\mathbf{F}_\sigma$, the vector $\mathbf{e}_i$ first occurs in the subspace $\mathcal{V}_j$, i.e. $\sigma(j) = i$. Consider the increasing sequence of spaces $\mathcal{V}_1 \subset \cdots \subset \mathcal{V}_n = \mathbb{K}^n$, and examine the sequence of differences

$$\dim(\mathcal{V}_k \cap \mathcal{W}_{n-i+1}) - \dim(\mathcal{V}_k \cap \mathcal{W}_{n-i}), \quad k = 1, \dots, n.$$

We find find the sequence of $j-1$ zeroes followed by $n-j$ ones. Thus $j = \sigma^{-1}(i)$ is determined by the flag. □

`Remark.` The theorem solves the following *classification* problem: In a vector space equipped with a fixed complete flag $\mathcal{W}_1 \subset \cdots \subset \mathcal{W}_n$, classify all complete flags up to invertible linear transformations preserving the fixed flag. According to the theorem, there are $n!$ equivalence classes determined by dimensions of intersections of spaces of the flags with the spaces of the fixed flag. The equivalence classes are known as **Bruhat cells**. This formulation also shows that the Gaussian elimination algorithm can be understood as a solution to a *simple* geometric classification problem. One can give a purely geometric proof of the above theorem (and hence a new proof of the LPU decomposition) by refining the argument in the proof of the Rank Theorem.

EXERCISES

269. List all *coordinate* complete flags in $\mathbb{K}^3$.

270. For each permutation matrix P of size 4×4, describe all upper triangular matrices U which can occur as a result of the modified Gaussian algorithm from the proof of the LPU decomposition theorem. For each permutation, find the maximal number of non-zero entries of U.

271.* Compute the **dimension** of each Bruhat cell, i.e. the number of parameters on which flags in the equivalence class of $\mathbf{F}_\sigma$ depend. ✓

272.* When $\mathbb{K}$ is a finite field of q elements, find the number of *all* complete flags in $\mathbb{K}^n$. ✓

273.* Prove that the number of Bruhat cells of dimension l is equal to the coefficient at q^l in the product (called **q-factorial**) ϟ

$$[n]_q! := (1+q)(1+q+q^2)\cdots(1+q+q^2+\cdots+q^{n-1}).$$

3 The Inertia Theorem

We study here classification of quadratic forms and some generalizations of this problem. The answer actually depends on properties of the field of scalars. This section focuses on the cases $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, while some other cases are delegated to Supplement F. We begin, however, with a key argument that remains valid in general. [3]

Orthogonal Bases

In section "Matrices" of Chapter 2 we established a one-to-one correspondence between symmetric bilinear forms and quadratic forms. To recall, a **symmetric bilinear form** on a vector space $\mathcal{V}$ is a function $Q : \mathcal{V} \times \mathcal{V} \to \mathbb{K}$, $(\mathbf{x}, \mathbf{y}) \mapsto Q(\mathbf{x}, \mathbf{y})$, which is linear in each vector variable $\mathbf{x}$ and $\mathbf{y}$, and symmetric, i.e. $Q(\mathbf{y}, \mathbf{x}) = Q(\mathbf{x}, \mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{V}$. Taking the arguments in a symmetric bilinear form equal to each other, one obtains the corresponding quadratic form, which we will denote by the same letter: $Q : \mathcal{V} \to \mathbb{K}$. Thus, $Q(\mathbf{x}) := Q(\mathbf{x}, \mathbf{x})$. (This should not cause confusion: whenever there are two arguments, it is the bilinear form, and when there is only one, it is the corresponding quadratic form.) The symmetric bilinear form is reconstructed from the corresponding quadratic form as

$$Q(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\left[Q(\mathbf{x} + \mathbf{y}) - Q(\mathbf{x}) - Q(\mathbf{y})\right].$$

In coordinates, if $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ is a basis of $\mathcal{V}$, we have

$$Q(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i q_{ij} y_j = \mathbf{x}^t Q \mathbf{y}, \text{ where } q_{ij} = Q(\mathbf{e}_i, \mathbf{e}_j) = q_{ji},$$

and respectively $Q(\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i q_{ij} x_j = \mathbf{x}^t Q \mathbf{x}$.

Under a linear change of coordinates $\mathbf{x} = C\mathbf{x}', \mathbf{y} = C\mathbf{y}'$, the symmetric coefficient matrix $Q = [q_{ij}]$ changes according to the transformation rule $Q \mapsto C^t Q C$.

A basis $\{\mathbf{f}_1, \dots, \mathbf{f}_n\}$ in the space $\mathcal{V}$ is called **Q-orthogonal** if $Q(\mathbf{f}_i, \mathbf{f}_j) = 0$ for all $i \neq j$, i.e. if the symmetric coefficient matrix of the quadratic form with respect to this basis is diagonal.

Lemma. ***Every quadratic form in a finite dimensional vector space has an orthogonal basis.***

[3] More precisely, whenever $\mathbb{K}$ does not contain $\mathbb{Z}_2$, so that $1/2$ exists.

Proof. We use induction on the dimension $n = \dim \mathcal{V}$ of the vector space. For $n = 1$ the requirement is empty. Let us construct a Q-orthogonal basis in $\mathcal{V}$ assuming that every quadratic form in space of dimension $n-1$ has an orthogonal basis. If the given quadratic form Q is identically zero, the corresponding symmetric bilinear form is identically zero too, and so any basis is Q-orthogonal. If the quadratic form is not identically zero, then there exists a vector $\mathbf{f}_1$ such that $Q(\mathbf{f}_1) \neq 0$. Let $\mathcal{W}$ be the subspace in $\mathcal{V}$ consisting of all vectors Q-orthogonal to $\mathbf{f}_1$: $\mathcal{W} = \{\mathbf{x} \in \mathcal{V} \mid Q(\mathbf{f}_1, \mathbf{x}) = 0.\}$. This subspace does *not* contain $\mathbf{f}_1$ and is given by 1 linear equation. Thus $\dim \mathcal{W} = n-1$. Let $\{\mathbf{f}_2, \ldots, \mathbf{f}_n\}$ be a basis in $\mathcal{W}$ orthogonal with respect to the symmetric bilinear form obtained by restricting Q to this subspace. Such a basis exists by the induction hypothesis. Therefore $Q(\mathbf{f}_i, \mathbf{f}_j) = 0$ for all $1 < i < j$. Besides, $Q(\mathbf{f}_1, \mathbf{f}_i) = 0$ for all $i > 1$, since $\mathbf{f}_i \in \mathcal{W}$. Then $Q(\mathbf{f}_i, \mathbf{f}_j) = 0$ for all $i > j$ by the symmetry of Q. Thus $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\}$ is a Q-orthogonal basis of $\mathcal{V}$. □

Corollary. ***For every symmetric $n \times n$-matrix Q with entries from $\mathbb{K}$ there exists an invertible matrix C such that $C^t Q C$ is diagonal.***

The diagonal entries here are the values $Q(\mathbf{f}_1), \ldots, Q(\mathbf{f}_n)$.

Inertia Indices

Consider the case $\mathbb{K} = \mathbb{R}$.

Given a quadratic form Q in $\mathbb{R}^n$, we pick a Q-orthogonal basis $\{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$ and then *rescale* those of the basis vectors for which $Q(\mathbf{f}_i) \neq 0$: $\mathbf{f}_i \mapsto \tilde{\mathbf{f}}_i = |Q(\mathbf{f}_i)|^{-1/2} \mathbf{f}_i$. After such rescaling, the non-zero coefficients $Q(\tilde{\mathbf{f}}_i)$ of the quadratic form will become ± 1. Reordering the basis so that the terms with positive coefficients come first, and negative next, we transform Q to the normal form:

$$Q = X_1^2 + \cdots + X_p^2 - X_{p+1}^2 - \cdots - X_{p+q}^2, \quad p + q \leq n.$$

Note that by restricting Q to the subspace $X_{p+1} = \cdots = X_n = 0$ of dimension p we obtain a quadratic form on this subspace which is **positive** (or **positive definite**), i.e takes on positive values everywhere outside the origin.

Proposition. ***The numbers p and q of positive and negative squares in the normal form are equal to the maximal dimensions of the subspaces in $\mathbb{R}^n$ where the quadratic form Q (respectively, $-Q$) is positive.***

Proof. The quadratic form $Q = X_1^2+...+X_p^2-X_{p+1}^2-...-X_{p+q}^2$ is *non*-positive everywhere on the subspace $\mathcal{W}$ of dimension $n-p$ given by the equations $X_1 = ... = X_p = 0$. Let us show that the existence of a subspace $\mathcal{V}$ of dimension $p+1$ where the quadratic form is positive leads to a contradiction. Indeed, the subspaces $\mathcal{V}$ and $\mathcal{W}$ would intersect in a subspace of dimension at least $(p+1)+(n-p)-n = 1$, containing therefore non-zero vectors $\mathbf{x}$ with $Q(\mathbf{x}) > 0$ and $Q(\mathbf{x}) \le 0$. Thus, Q is positive on some subspace of dimension p and cannot be positive on any subspace of dimension $> p$. Likewise, $-Q$ is positive on some subspace of dimension q and cannot be positive on any subspace of dimension $> q$. □

The maximal dimensions of positive subspaces of Q and $-Q$ are called respectively **positive** and **negative inertia indices** of a quadratic form in question. By definition, inertia indices of a quadratic form do not depend on the choice of a coordinate system. Our Proposition implies that the normal forms with different pairs of values of p and q are pairwise non-equivalent. This establishes the Inertia Theorem (as stated in Section 4 of Chapter 1).

Theorem. ***Every quadratic form in*** $\mathbb{R}^n$ ***by a linear change of coordinates can be transformed to exactly one of the normal forms:***

$$X_1^2 + ... + X_p^2 - X_{p+1}^2 - ... - X_{p+q}^2, \quad \textbf{\textit{where}}\ 0 \le p+q \le n.$$

The matrix formulation of the Inertia Theorem reads:

Every real symmetric matrix Q can be transformed to exactly one of the diagonal forms $\begin{bmatrix} I_p & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -I_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$ *by transformations of the form $Q \mapsto C^tQC$ defined by invertible real matrices C.*

EXERCISES

274. Find orthogonal bases and inertia indices of quadratic forms: ✓

$$x_1x_2 + x_2^2, \quad x_1^2 + 4x_1x_2 + 6x_2^2 - 12x_2x_3 + 18x_3^2, \quad x_1x_2 + x_2x_3 + x_3x_1.$$

275. Prove that $Q = \sum_{1\le i\le j\le n} x_ix_j$ is positive definite.

276. A minor of a square matrix formed by rows and columns with the same indices is called **principal**. Prove that all principal minors of the coefficient matrix of a positive definite quadratic form are positive.

277.* Let $\mathbf{a}_1, \dots, \mathbf{a}_p$ and $\mathbf{b}_1, \dots, \mathbf{b}_q$ be linear forms in $\mathbb{R}^n$, and let $Q(\mathbf{x}) = \mathbf{a}_1^2(\mathbf{x}) + \dots + \mathbf{a}_p^2(\mathbf{x}) - \mathbf{b}_1^2(\mathbf{x}) - \dots - \mathbf{b}_q^2(\mathbf{x})$. Prove that the positive and negative inertia indices of Q do not exceed p and q respectively. ϟ

Complex Quadratic Forms

Consider the case $\mathbb{K} = \mathbb{C}$.

Theorem. ***Every quadratic form in $\mathbb{C}^n$ can be transformed by linear changes of coordinates to exactly one of the normal forms:***

$$z_1^2 + \dots + z_r^2, \quad \textbf{\textit{where}}\ 0 \le r \le n.$$

Proof. Given a quadratic form Q, pick a Q-orthogonal basis in $\mathbb{C}^n$, order it in such a way that vectors $\mathbf{f}_1, \dots, \mathbf{f}_r$ with $Q(\mathbf{f}_i) \neq 0$ come first, and then rescale these vectors by $\mathbf{f}_i \mapsto Q(\mathbf{f}_i)^{-1/2}\mathbf{f}_i$.

In particular, we have proved that *every complex symmetric matrix Q can be transformed to exactly one of the forms* $\begin{bmatrix} I_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ *by the transformations of the form $Q \mapsto C^t Q C$ defined by invertible complex matrices C.* As it follows from the Rank Theorem, here $r = \operatorname{rk} Q$, the rank of the coefficient matrix of the quadratic form. This guarantees that the normal forms with different values of r are pairwise non-equivalent, and thus completes the proof. □

To establish the geometrical meaning of r, consider a more general situation.

Given a quadratic form Q on a $\mathbb{K}$-vector space $\mathcal{V}$, its **kernel** is defined as the subspace of $\mathcal{V}$ consisting of all vectors which are Q-orthogonal to all vectors from $\mathcal{V}$:

$$\operatorname{Ker} Q := \{\mathbf{z} \in \mathcal{V} | Q(\mathbf{z}, \mathbf{v}) = 0 \text{ for all } \mathbf{v} \in \mathcal{V}.\}$$

Note that the values $Q(\mathbf{x}, \mathbf{y})$ do not change when a vector from the kernel is added to either of $\mathbf{x}$ and $\mathbf{y}$.[4]

The **rank** of a quadratic form Q on $\mathbb{K}^n$ is defined as the codimension of $\operatorname{Ker} Q$. For example, the quadratic form $z_1^2 + \dots + z_r^2$ on $\mathbb{K}^n$ corresponds to the symmetric bilinear form $x_1y_1 + \dots + x_ry_r$, and has the kernel of codimension r defined by the equations $z_1 = \dots = z_r = 0$.

[4] As a result, the symmetric bilinear form Q descends to the *quotient space* $\mathcal{V}/\operatorname{Ker} Q$ (see Supplement D).

Conics

The set of all solutions to one polynomial equation in $\mathbb{K}^n$:

$$F(x_1, \dots, x_n) = 0$$

is called a **hypersurface**. When the polynomial F does not depend on one of the variables (say, x_n), the equation $F(x_1, \dots, x_{n-1}) = 0$ defines a hypersurface in $\mathbb{K}^{n-1}$. Then the solution set in $\mathbb{K}^n$ is called a **cylinder**, since it is the Cartesian product of the hypersurface in $\mathbb{K}^{n-1}$ and the line of arbitrary values of x_n.

Hypersurfaces defined by polynomial equations of degree 2 are often referred to as **conics** — a name reminiscent of conic sections, which are "hypersurfaces" in $\mathbb{K}^2$. The following application of the Inertia Theorem allows one to classify all conics in $\mathbb{R}^n$ up to **equivalence** defined by compositions of translations with invertible linear transformations.

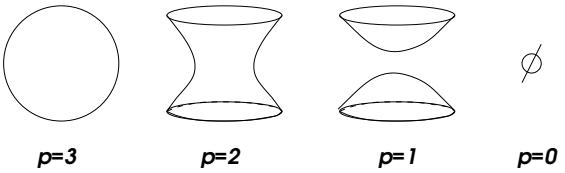

Figure 38

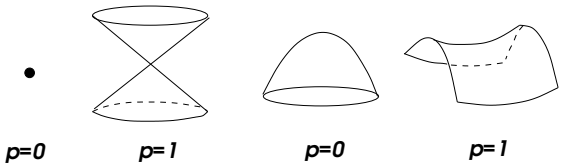

Figure 39

Theorem. ***Every conic in $\mathbb{R}^n$ is equivalent to either the cylinder over a conic in $\mathbb{R}^{n-1}$, or to one of the conics:***

$$x_1^2 + \dots + x_p^2 - x_{p+1}^2 - \dots - x_n^2 = 1, \quad 0 \le p \le n,$$
$$x_1^2 + \dots + x_p^2 = x_{p+1}^2 + \dots + x_n^2, \quad 0 \le p \le n/2,$$
$$x_n = x_1^2 + \dots + x_p^2 - x_{p+1}^2 - \dots - x_{n-1}^2, \quad 0 \le p \le (n-1)/2,$$

known as hyperboloids, cones, and paraboloids respectively.

For $n = 3$, all types of "hyperboloids" (of which the first type contains spheres and ellipsoids) are shown in Figure 38, and cones and paraboloids in Figure 39.

`Proof.` Given a degree 2 polynomial $F = Q(\mathbf{x}) + \mathbf{a}(\mathbf{x}) + c$, where Q is a non-zero quadratic form, $\mathbf{a}$ a linear form, and c a constant, we can apply a linear change of coordinates to transform Q to the form $\pm x_1^2 \pm \cdots \pm x_r^2$, where $r \leq n$, and then use the completion of squares in the variables $x_1, \ldots, x_r$ to make the remaining linear form independent of $x_1, \ldots, x_r$. When $r = n$, the resulting equations $\pm x_1^2 \pm \cdots \pm x_n^2 = C$ (where C is a new constant) define hyperboloids (when $C \neq 0$), or cones (when $C = 0$). When $r < n$, we can take the remaining *linear* part of the function F (together with the constant) for a new, $r+1$-st coordinate, provided that this linear part is non-constant. When $r = n-1$, we obtain the equations of paraboloids. When $r < n-1$, or if $r = n-1$, but the linear function was constant, the function F, written in new coordinates, does not depend on the last of them, and this defines the cylinder over a conic in $\mathbb{R}^{n-1}$. □

Classification of conics in $\mathbb{C}^n$ is obtained in the same way, but the answer looks simpler, since there are no signs $\pm$ in the normal forms of quadratic forms over $\mathbb{C}$.

`Theorem.` ***Every conic in $\mathbb{C}^n$ is equivalent to either the cylinder over a conic in $\mathbb{C}^{n-1}$, or to one of the three conics:***

$$z_1^2 + \cdots + z_n^2 = 1, \;\; z_1^2 + \cdots + z_n^2 = 0, \;\; z_n = z_1^2 + \cdots + z_{n-1}^2.$$

`Example.` Let Q be a non-degenerate quadratic form with *real* coefficients in 3 variables. According to the previous (real) classification theorem, the conic $Q(x_1, x_2, x_3) = 1$ can be transformed by a real change of coordinates into one of the 4 normal forms shown on Figure 38. The same real change of coordinates identifies the set of *complex* solutions to the equation $Q(z_1, z_2, z_3) = 1$ with that of the normal form: $\pm z_1^2 \pm z_2^2 + \pm z_3^2 = 1$. However, $-z^2$ becomes z^2 after the change $z \mapsto \sqrt{-1}z$, which identifies the set of complex solutions with the **complex sphere** in $\mathbb{C}^3$, given by the equation $z_1^2 + z_2^2 + z_3^2 = 1$. Thus, various complex conics equivalent to the complex sphere and given by equations with real coefficients, "expose" themselves in $\mathbb{R}^3$ by various *real forms*: real spheres or ellipsoids, hyperboloids of one or two sheets (as shown on Figure 38), or even remain invisible (when the set of real points is empty).

`Remark.` The same holds true in general: various hyperboloids (as well as cones or paraboloids) of the real classification theorem are real forms of complex conics defined by the same equations. They become equivalent when complex changes of coordinates are allowed. In this sense, the three normal forms of the last theorem represent hyperboloids, cones and paraboloids of the previous one.

EXERCISES

278. Find the place of surfaces $x_1x_2+x_2x_3 = \pm 1$ and $x_1x_2+x_2x_3+x_3x_1 = \pm 1$ in the classification of conics in $\mathbb{R}^3$.

279. Examine normal forms of hyperboloids in $\mathbb{R}^4$ and find out how many connected components ("sheets") each of them has. ✓

280. Find explicitly a $\mathbb{C}$-linear transformation that identifies the sets of complex solutions to the equations $xy = 1$ and $x^2 + y^2 = 1$.

281. Find the rank of the quadratic form $z_1^2 + 2iz_1z_2 - z_2^2$.

282. Define the **kernel** of an *anti*-symmetric bilinear form A on a vector space $\mathcal{V}$ as the subspace $\operatorname{Ker} A := \{\mathbf{z} \in \mathcal{V} \mid A(\mathbf{z}, \mathbf{x}) = 0 \text{ for all } \mathbf{x} \in \mathcal{V}\}$, and prove that the form descends to the quotient space $\mathcal{V}/\operatorname{Ker} A$.

283. Classify conics in $\mathbb{C}^2$ up to linear inhomogeneous transformations. ✓

284. Find the place of the complex conic $z_1^2 - 2iz_1z_2 - z_2^2 = iz_1 + z_2$ in the classification of conics in $\mathbb{C}^2$. ✓

285. Classify all conics in $\mathbb{C}^3$ up to linear inhomogeneous transformations.

286. Prove that there are $3n - 1$ equivalence classes of conics in $\mathbb{C}^n$.

Hermitian and Anti-Hermitian Forms

In Chapter 2, at the end of section *"Matrices"*, we established one-to-one correspondences between Hermitian, anti-Hermitian, Hermitian quadratic and anti-Hermitian quadratic forms on a *complex* vector space $\mathcal{V}$.

To recall, a **sesquilinear form** is a function $\mathcal{V} \times \mathcal{V} \to \mathbb{C}$, which is $\mathbb{C}$-linear in the 2nd argument, and anti-linear in the 1st. Such a form H is called **Hermitian-symmetric** if $H(\mathbf{w}, \mathbf{z}) = \overline{H(\mathbf{z}, \mathbf{w})}$ for all $\mathbf{z}, \mathbf{w} \in \mathcal{V}$. The corresponding **Hermitian quadratic form** is $H(\mathbf{z}) := H(\mathbf{z}, \mathbf{z})$. (It is denoted by the same letter, but takes in one vector argument, and assumes real values.) In a coordinate system $\mathbf{z} = z_1\mathbf{e}_1 + \cdots + z_n\mathbf{e}_n$ on $\mathcal{V}$, an Hermitian quadratic form is given by the formula

$$H(\mathbf{z}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \overline{z_i}\, h_{ij}\, z_j,$$

where the coefficient matrix $H = [h_{ij}]$ is Hermitian-symmetric: $H^\dagger = H$, i.e. $\bar{h}_{ij} = h_{ji}$. The corresponding Hermitian-symmetric form has the coordinate expression

$$H(\mathbf{z}, \mathbf{w}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \overline{z}_i \, h_{ij} w_j.$$

An **anti-Hermitian form**, by definition, is a sesquilinear form, Q, satisfying $Q(\mathbf{w}, \mathbf{z}) = -\overline{Q(\mathbf{z}, \mathbf{w})}$ for all $\mathbf{z}, \mathbf{w} \in \mathcal{V}$. Every such form (and the corresponding **anti-Hermitian quadratic form** $Q(\mathbf{z}) := Q(\mathbf{z}, \mathbf{z})$) is obtained by multiplication by $\sqrt{-1}$ from an Hermitian-symmetric (respectively Hermitian quadratic) form, and *vice versa*.

Theorem. ***Every Hermitian quadratic form H in $\mathbb{C}^n$ can be transformed by a $\mathbb{C}$-linear change of coordinates to exactly one of the normal forms***

$$|z_1|^2 + \cdots + |z_p|^2 - |z_{p+1}|^2 - \cdots - |z_{p+q}|^2, \quad 0 \le p + q \le n.$$

Proof. It is the same as in the case of the Inertia Theorem for real quadratic forms. We pick a vector $\mathbf{f}_1$ such that $H(\mathbf{f}_1) = \pm 1$, and consider the subspace $\mathcal{V}_1$ consisting of all vectors H-orthogonal to $\mathbf{f}_1$: $\mathcal{V}_1 = \{\mathbf{z} \mid H(\mathbf{f}_1, \mathbf{z}) = 0\}$. It does not contain $\mathbf{f}_1$ (since $H(\mathbf{f}_1, \mathbf{f}_1) = H(\mathbf{f}_1) \neq 0$), and has therefore complex codimension 1. We consider the Hermitian form obtained by restricting H to $\mathcal{V}_1$ and proceed the same way, i.e. pick a vector $\mathbf{f}_2 \in \mathcal{V}_1$ such that $H(\mathbf{f}_2) = \pm 1$, and pass to the subspace $\mathcal{V}_2$ consisting of all vectors of $\mathcal{V}_1$ which are H-orthogonal to $\mathbf{f}_2$. The process stops when we reach a subspace $\mathcal{V}_r$ of codimension r in $\mathbb{C}^n$ such that the restriction of the form H to $\mathcal{V}_r$ vanishes identically. Then we pick any basis $\{\mathbf{f}_{r+1}, \ldots, \mathbf{f}_n\}$ in $\mathcal{V}_r$. The vectors $\mathbf{f}_1, \ldots, \mathbf{f}_n$ form a basis in $\mathbb{C}^n$ which is H-orthogonal (since $H(\mathbf{f}_i, \mathbf{f}_j) = 0$ for all $i < j$ by construction), and $H(\mathbf{f}_i, \mathbf{f}_i) = \pm 1$ (for $i \le r$) or $= 0$ for $i > r$. Reordering the vectors $\mathbf{f}_1, \ldots, \mathbf{f}_r$ so that those with the values $+1$ come first, we obtain the required normal form for H, where $p + q = r$.

To prove that the normal forms with different pairs of values of p and q are non-equivalent to each other, we show (the same way as in the case of real quadratic forms) that ***the number p (q) of positive (respectively negative) squares in the normal form is equal to the maximal dimension of a subspace where the Hermitian form H (respectively $-H$) is positive definite.*** $\square$.

Corollary 1. ***An anti-Hermitian quadratic form Q in $\mathbb{C}^n$ can be transformed by a $\mathbb{C}$-linear change of coordinates to exactly one of the normal forms***

$$i|z_1|^2 + \cdots + i|z_p|^2 - i|z_{p+1}|^2 - \cdots - i|z_{p+q}|^2, \quad 0 \le p+q \le n.$$

Using matrix notation, one expresses a sesquilinear form with the coefficient matrix T by the matrix product formula $\mathbf{z}^\dagger T\mathbf{w}$, where $\mathbf{w}$ is a column, and $\mathbf{z}^\dagger$ is the row Hermitian-adjoint to the column $\mathbf{z}$, i.e. obtained from it by transposition and complex conjugation. Applying a $\mathbb{C}$-linear change of variables $\mathbf{z} = C\mathbf{z}', \mathbf{w} = C\mathbf{w}'$, we find

$$(\mathbf{z}')^\dagger T'\mathbf{w}' = \mathbf{z}^\dagger T\mathbf{w} = \mathbf{z}^\dagger C^\dagger TC\mathbf{w}, \quad \text{i.e. } T' = C^\dagger TC.$$

Corollary 2. ***Any Hermitian (anti-Hermitian) matrix can be transformed to exactly one of the normal forms***

$$\begin{bmatrix} I_p & 0 & 0 \\ 0 & -I_q & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \left(\textit{respectively} \quad \begin{bmatrix} iI_p & 0 & 0 \\ 0 & -iI_q & 0 \\ 0 & 0 & 0 \end{bmatrix} \right).$$

by transformations of the form $T \mapsto C^\dagger TC$ defined by invertible complex matrices C.

It follows that $p+q$ is equal to the rank of the coefficient matrix of the (anti-)Hermitian form.

EXERCISES

287. Check that $T^\dagger = T^t$ if and only if T is real.

288. Show that diagonal entries of an Hermitian matrix are real, and of anti-Hermitian imaginary.

289. Find all complex matrices which are symmetric and anti-Hermitian simultaneously. ✓

290.* Prove that a sesquilinear form T of $\mathbf{z}, \mathbf{w} \in \mathcal{V}$ can be expressed in terms of its values at $\mathbf{z} = \mathbf{w}$, and find such an expression. ϟ✓

291. Define sesquilinear forms $T : \mathbb{C}^m \times \mathbb{C}^n \to \mathbb{C}$ of pairs of vectors $(\mathbf{z}, \mathbf{w})$ taken from two different spaces, and prove that $T(\mathbf{z}, \mathbf{w}) = \langle \mathbf{z}, T\mathbf{w} \rangle$, where T is the $m \times n$-matrix of coefficients of the form, and $\langle \cdot, \cdot \rangle$ is the standard Hermitian dot-product in $\mathbb{C}^m$. ϟ

292. Prove that under changes of variables $\mathbf{v} = D\mathbf{v}'$, $\mathbf{w} = C\mathbf{w}'$ the coefficient matrices of sesquilinear forms are transformed as $P \mapsto D^\dagger PC$.

293. Prove that $\langle A\mathbf{z}, \mathbf{w} \rangle = \langle \mathbf{z}, B\mathbf{w} \rangle$ for all $\mathbf{z} \in \mathbb{C}^m$, $\mathbf{w} \in \mathbb{C}^n$ if and only if $A = B^\dagger$. Here $\langle \cdot, \cdot \rangle$ denote *Hermitian* dot-products in $\mathbb{C}^n$ or $\mathbb{C}^m$. ϟ

294. Prove that $(AB)^\dagger = B^\dagger A^\dagger$. ϟ

295. Prove that for (anti-)Hermitian matrices A and B, the **commutator** matrix $AB - BA$ is (anti-)Hermitian.

296. Find out which of the following forms are Hermitian or anti-Hermitian and transform them to the appropriate normal forms: ϟ

$$\bar{z}_1 z_2 - \bar{z}_2 z_1, \quad \bar{z}_1 z_2 + \bar{z}_2 z_1, \quad \bar{z}_1 z_1 + i\bar{z}_2 z_1 - i\bar{z}_1 z_2 - \bar{z}_2 z_2.$$

Sylvester's Rule

Let H be a Hermitian $n \times n$-matrix. Denote by $\Delta_0 = 1$, $\Delta_1 = h_{11}$, $\Delta_2 = h_{11}h_{22} - h_{12}h_{21}$, $\ldots$, $\Delta_n = \det H$ the minors formed by the intersection of the first k rows and columns of H, $k = 1, 2, \ldots, n$ (Figure 40). They are called **leading minors** of the matrix H. Note that $\det H = \det H^t = \det \bar{H} = \overline{\det H}$ is real, and the same is true for each Δ_k, since it is the determinant of an Hermitian $k \times k$-matrix. The following result is due to the English mathematician James **Sylvester** (1814–1897).

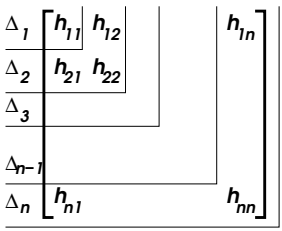

Figure 40

Theorem. ***Suppose that an Hermitian $n \times n$-matrix H has non-zero leading minors. Then the negative inertia index of the corresponding Hermitian form is equal to the number of sign changes in the sequence $\Delta_0, \Delta_1, \ldots, \Delta_n$.***

Remark. The hypothesis that $\det H \neq 0$ means that the Hermitian form is **non-degenerate**, or equivalently, that its kernel is trivial. In other words, for each non-zero vector $\mathbf{x}$ there exists $\mathbf{y}$ such that $H(\mathbf{x}, \mathbf{y}) \neq 0$. Respectively, the assumption that all leading minors are non-zero means that *restrictions of the Hermitian forms to all spaces of the standard coordinate flag*

$$\mathrm{Span}(\mathbf{e}_1) \subset \mathrm{Span}(\mathbf{e}_1, \mathbf{e}_2) \subset \cdots \subset \mathrm{Span}(\mathbf{e}_1, \ldots, \mathbf{e}_k) \subset \ldots$$

are non-degenerate. The proof of the theorem consists in classifying such Hermitian forms up to linear changes of coordinates that *preserve* the flag.

`Proof.` As before, we inductively construct an H-orthogonal basis $\{\mathbf{f}_1, \dots, \mathbf{f}_n\}$ and normalize the vectors so that $H(\mathbf{f}_i) = \pm 1$, requiring however that each $\mathbf{f}_k \in \mathrm{Span}(\mathbf{e}_1, \dots, \mathbf{e}_k)$. When such vectors $\mathbf{f}_1, \dots, \mathbf{f}_{k-1}$ are already found, the vector $\mathbf{f}_k$, H-orthogonal to them, can be found (by the Rank Theorem) in the k-dimensional space of the flag, and can be assumed to satisfy $H(\mathbf{f}_k) = \pm 1$, since the Hermitian form on this space is non-degenerate. Thus, ***an Hermitian form non-degenerate on each space of the standard coordinate flag can be transformed to one*** (and in fact exactly one) ***of the*** 2^n ***normal forms*** $\pm|z_1|^2 \pm \dots \pm |z_n|^2$ ***by a linear change of coordinates preserving the flag.***

In matrix form, this means that there exists an invertible *upper triangular matrix* C such that $D = C^\dagger H C$ is diagonal with all diagonal entries equal to ± 1. Note that transformations of the form $H \mapsto C^\dagger H C$ may change the determinant but preserve its sign:

$$\det(C^\dagger H C) = (\det C^\dagger)(\det H)(\det C) = \det H |\det C|^2.$$

When C is upper triangular, the same holds true for all leading minors, i.e. each Δ_k has the same sign as the leading $k \times k$-minor of the diagonal matrix D with the diagonal entries $d_1, \dots, d_n$ equal ± 1. The latter minors form the sequence $1, d_1, d_1 d_2, \dots, d_1 \dots d_k, \dots$, where the sign is changed each time as $d_k = -1$. Thus the total number of sign changes is equal to the number of negative squares in the normal form. □

When the form H is positive definite, its restrictions to any subspace is positive definite and hence non-degenerate automatically. We obtain the following corollaries.

`Corollary 1.` ***Any positive definite Hermitian form in*** $\mathbb{C}^n$ ***can be transformed into*** $|z_1|^2 + \dots + |z_n|^2$ ***by a linear change of coordinates preserving a given complete flag.***

`Corollary 2.` ***A Hermitian form in*** $\mathbb{C}^n$ ***is positive definite if and only if all of its leading minors are positive.***

Note that the standard basis of $\mathbb{C}^n$ is **orthonormal** with respect to the Hermitian dot product $\langle \mathbf{x}, \mathbf{y} \rangle = \sum \bar{x}_i y_i$, i.e. $\langle \mathbf{e}_i, \mathbf{e}_j \rangle = 0$ for $i \neq i$, and $\langle \mathbf{e}_i, \mathbf{e}_i \rangle = 1$.

`Corollary 3.` ***Every positive definite Hermitian form in*** $\mathbb{C}^n$ ***has an orthonormal basis*** $\{\mathbf{f}_1, \dots, \mathbf{f}_n\}$ ***such that*** $\mathbf{f}_k \in \mathrm{Span}(\mathbf{e}_1, \dots, \mathbf{e}_k)$***.***

`Remarks.` (1) The process of replacing a given basis $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ with a new basis, orthonormal with respect to a given positive definite Hermitian form and such that each $\mathbf{f}_k$ is a linear combination of $\mathbf{e}_1, \dots, \mathbf{e}_k$, is called **Gram–Schmidt orthogonalization**.

(2) Results of this subsection hold true for quadratic forms in $\mathbb{R}^n$. Namely, our reasoning can be easily adjusted to this case. Note also that every real symmetric matrix is Hermitian.

EXERCISES

297. Prove that for every symmetric matrix Q all of whose leading minors are non-zero there exists a *unipotent* upper triangular matrix C such that $D = C^t Q C$ is diagonal, and express the diagonal entries of D in terms of the leading minors. ✓

298. Use Sylvester's rule to find inertia indices of quadratic forms: ✓

$$x_1^2 + 2x_1x_2 + 2x_2x_3 + 2x_1x_4, \quad x_1x_2 - x_2^2 + x_3^2 + 2x_2x_4 + x_4^2.$$

299. Compute determinants and inertia indices of quadratic forms:

$$x_1^2 - x_1x_2 + x_2^2, \quad x_1^2 + x_2^2 + x_3^2 - x_1x_2 - x_2x_3.$$

300. Prove positivity of the quadratic form $\sum_{i=1}^n x_i^2 - \sum_{1 \le i < n} x_i x_{i+1}$.

301.* Prove that when the square of a linear form is added to a positive quadratic form, the determinant of the coefficient matrix increases. ϟ

4 The Minkowski–Hasse Theorem

Here we discuss the problem of clasification of quadratic forms over $\mathbb{K} = \mathbb{Z}_p$ and $\mathbb{K} = \mathbb{Q}$.

Finite Fields

Consider the case $\mathbb{K} = \mathbb{Z}_p$, the field of integers modulo a prime number $p \neq 2$. It consists of p elements corresponding to possible remainders $0, 1, 2, \dots, p-1$ of integers divided by p. It is indeed a field due to some facts of elementary number theory. Namely, when an integer a is not divisible by p, it follows from the Euclidean algorithm, that the greatest common divisor of a and p, which is 1, can be represented as their linear combination: $1 = ma + np$. Modulo p, this means that m is inverse to a. Thus every non-zero element of $\mathbb{Z}_p$ is invertible.

As we have seen, in classification of quadratic forms, it is important to know which scalars are complete squares.

`Examples.` (1) Let $Q = ax^2$ and $Q' = a'x^2$ be two non-zero quadratic forms on $\mathbb{K}^1$ (i.e. $a, a' \in \mathbb{K} - \{0\}$). Rescaling x to cx, where c can be any element from $\mathbb{K} - \{0\}$, transforms Q into ac^2x^2. Thus, the quadratic forms in $\mathbb{K}^1$ are equivalent if and only if $a' = ac^2$ for some non-zero c. i.e. if the ratio a'/a is a complete square in $\mathbb{K} - \{0\}$.

(2) When $\mathbb{K} = \mathbb{C}$, every element is a complete square, so there is only one equivalence class of *non-zero* quadratic forms in $\mathbb{C}^1$. When $\mathbb{K} = \mathbb{R}$, there are two such classes according to the sign of the coefficient (because complete squares are exactly the positive reals).

(3) When $\mathbb{K} = \mathbb{Z}_p$, there are $p-1$ non-zero quadratic forms which are divided into two equivalence classes. One class can be represented by the normal form x^2 and consists of those quadratic forms whose coefficient is a complete square, $a = c^2 \neq 0$. There are $(p-1)/2$ such forms, i.e. a half of all non-zero ones, since each complete square $a = c^2$ has exactly two different square roots: c and $-c$. Let $\varepsilon \neq 0$ be any non-square. Then, when c^2 runs all squares, εc^2 runs all the $(p-1)/2$ non-squares. Thus $Q = \varepsilon x^2$ can be taken for the normal form in the other equivalence class.

(4) In $\mathbb{Z}_{13}$, there are 12 non-zero elements represented by the integers $\pm 1, \dots, \pm 6$, their squares are $1, 4, -4, 3, -1, -3$ respectively, and the non-squares are $\pm 2, \pm 5, \pm 6$. Any of them (e.g. 2) can be taken for ε. Thus, every non-zero quadratic form on $\mathbb{Z}_{13}^1$ is equivalent to either x^2 or $2x^2$. This example suggests that there may be no choice of the normal form εx^2 good for all $\mathbb{Z}_p$ at once.

`Theorem.` ***Every non-zero quadratic form on $\mathbb{Z}_p^n$, $p \neq 2$, is equivalent to exactly one of the forms***

$$x_1^2 + x_2^2 + \cdots + x_r^2, \quad \varepsilon x_1^2 + x_2^2 + \cdots + x_r^2, \quad 1 \leq r \leq n.$$

Proof. First note that both normal forms have rank r. Since the rank of the coefficient matrix Q of a quadratic form does not change under the transformations C^tQC defined by invertible matrices C, it suffices to prove that a quadratic form of a fixed rank $r > 0$ is equivalent to exactly one of the two normal forms.

Next, the symmetric bilinear form corresponding to the quadratic form Q of rank r has kernel $\operatorname{Ker}(Q)$ (non-trivial when $r < n$) and defines a *non-degenerate* symmetric bilinear form on the quotient space $\mathbb{K}^n/\operatorname{Ker}(Q)$ of dimension r. Thus, it suffices to prove that a non-degenerate quadratic form on $\mathbb{K}^r = \mathbb{Z}_p^r$ is equivalent to exactly one of the two normal forms.

The normal forms, considered as non-degenerate quadratic forms on $\mathbb{Z}_p^r$, are not equivalent to each other. Indeed, they have diagonal coefficient matrices with determinants equal 1 and ε respectively, of which the first one is a square in $\mathbb{Z}_p$, and the second is not. But for equivalent non-degenerate quadratic forms, the ratio of the determinants is a complete square: $\det(C^tQC)/(\det Q) = (\det C)^2$.

To transform a non-degenerate quadratic form Q on $\mathbb{Z}_p^r$ to one of the normal forms, we can construct a Q-orthogonal basis $\{\mathbf{f}_1, \dots, \mathbf{f}_r\}$ and thus reduce Q to the form $a_1x_1^2 + \dots + a_rx_r^2$. Here $a_i = Q(\mathbf{f}_i) \neq 0$. We would like to show that a better choice of a basis can be made, such that $Q(\mathbf{f}_i) = 1$ for all $i > 1$. Let us begin with the case $r = 2$.

Lemma. ***Given non-zero*** $a, b \in \mathbb{Z}_p$***, there exist*** $(x, y) \in \mathbb{Z}_p^2$ ***such that*** $ax^2 + by^2 = 1$***.***

Indeed, when each of x and y runs all p possible values (including 0) each of ax^2 and $1 - by^2$ takes on $(p-1)/2 + 1$ different values. Since the total number exceeds p, we must have $ax^2 = 1 - by^2$ for some x and y. □

Thus, given a non-degenerate quadratic form $P = ax^2 + by^2$ in $\mathbb{Z}_p^2$, there exists $\mathbf{f} \in \mathbb{Z}_p^2$, such that $P(\mathbf{f}) = 1$. Taking a second vector P-orthogonal to $\mathbf{f}$, we obtain a new basis in which P takes on the form $a'x^2 + b'y^2$ with $a' = 1$ and $b' \neq 0$.

We can apply this trick $r - 1$ times to the quadratic form $Q = a_1x_1^2 + \dots + a_rx_r^2$ using two of the variables at a time, and end up with the form where $a_r = a_{r-1} = \dots = a_2 = 1$. Finally, rescaling x_1 as in Example 3, we can make a_1 equal either 1 or ε.

Remark. Readers comfortable with arbitrary finite fields can easily check that our proof and the theorem remain true over any finite field $\mathbb{K} \supset \mathbb{Z}_p$, $p \neq 2$, with any non-square in $\mathbb{K}$ taken in the role of ε.

EXERCISES

302. In $\mathbb{Z}_{11}$, compute multiplicative inverses of all non-zero elements, find all non-square elements, and find out if any of the quadratic forms x_1x_2, $x_1^2 + x_1x_2 + 3x_2^2$, $2x_1^2 + x_1x_2 - 2x_2^2$ are equivalent to each other in $\mathbb{Z}_{11}^2$.

303. Prove that when p is a prime of the form $4k-1$, then every non-degenerate quadratic form in $\mathbb{Z}_p^n$ is equivalent to one of the two normal forms $\pm x_1^2 + x_2^2 + \cdots + x_n^2$. ϟ

304. Prove that in a sutable coordinate system (u,v,w) in the space $\mathbb{Z}_p^3$ of symmetric 2×2-matrices $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$ over $\mathbb{Z}_p$, the determinant $ac-b^2$ takes on the form $u^2+v^2+w^2$.

The Case of $\mathbb{K} = \mathbb{Z}_2$.

This is a peculiar world where $2=0$, $-1=1$, and where therefore the usual one-to-one correspondence between quadratic and symmetric bilinear forms is broken, and the distinction between symmetric and anti-symmetric forms lost.

Yet, consider a symmetric bilinear form Q on $\mathbb{Z}_2^n$:

$$Q(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i q_{ij} y_j, \text{ where } q_{ij} = q_{ji} = 0 \text{ or } 1 \text{ for all } i,j.$$

The corresponding quadratic form $Q(\mathbf{x}) := Q(\mathbf{x},\mathbf{x})$ still exists, but satisfies $Q(\mathbf{x}+\mathbf{y}) = Q(\mathbf{x}) + 2Q(\mathbf{x},\mathbf{y}) + Q(\mathbf{y}) = Q(\mathbf{x}) + Q(\mathbf{y})$ and hence defines a *linear* function $\mathbb{Z}_2^n \to \mathbb{Z}_2$. This linear function can be identically zero, i.e. $Q(\mathbf{x},\mathbf{x}) = 0$ for all $\mathbf{x}$, in which case the bilinear form Q is called **even**. This happens exactly when all the diagonal entries of the coefficient matrix vanish: $q_{ii} = Q(\mathbf{e}_i,\mathbf{e}_i) = 0$ for all i. Otherwise the bilinear form Q is called **odd**. We begin with classifying *even non-degenerate* forms. As it is implied by the following theorem, such forms exist only in $\mathbb{Z}_2$-spaces of even dimension $n = 2k$.

`Theorem.` ***Every even non-degenerate symmetric bilinear form Q on $\mathbb{Z}_2^n$ in a suitable coordinate system is given by the formula:***

$$Q(\mathbf{x},\mathbf{y}) = x_1y_2 + x_2y_1 + \cdots + x_{2k-1}y_{2k} + x_{2k}y_{2k-1}. \qquad \text{(i)}$$

`Proof.` Pick any $\mathbf{f}_1 \neq \mathbf{0}$ and find $\mathbf{f}_2$ such that $Q(\mathbf{f}_1,\mathbf{f}_2) = 1$. Such $\mathbf{f}_2$ must exist since Q is non-degenerate. In $\mathrm{Span}(\mathbf{f}_1,\mathbf{f}_2)$, we have: $Q(x_1\mathbf{f}_1 + x_2\mathbf{f}_2, y_1\mathbf{f}_1 + y_2\mathbf{f}_2) = x_1y_2 + x_2y_1$, since Q is even, i.e. $Q(\mathbf{f}_1,\mathbf{f}_1) = Q(\mathbf{f}_2,\mathbf{f}_2) = 0$.

Let $\mathcal{V}$ denote the space of all vectors Q-orthogonal to $\mathrm{Span}(\mathbf{f}_1,\mathbf{f}_2)$. It is given by two linear equations: $Q(\mathbf{f}_1,\mathbf{x}) = 0$, $Q(\mathbf{f}_2,\mathbf{x}) = 0$, which are independent (since $\mathbf{x} = \mathbf{f}_1$ satisfies the first one but not the second, and $\mathbf{x} = \mathbf{f}_2$ the other way around). Therefore $\operatorname{codim}\mathcal{V} = 2$. If $\mathbf{v}\in\mathcal{V}$ is Q-orthogonal to all vectors from $\mathcal{V}$, then being Q-orthogonal to $\mathbf{f}_1$ and $\mathbf{f}_2$, it lies in $\operatorname{Ker} Q$, which is trivial. This shows that the restriction of the bilinear form Q to $\mathcal{V}$ is non-degenerate. We can continue our constriction inductively, i.e. find $\mathbf{f}_3,\mathbf{f}_4 \in \mathcal{V}$ such that $Q(\mathbf{f}_3,\mathbf{f}_4) = 1$, take their Q-orthogonal complement in

$\mathcal{V}$, and so on. At the end we obtain a basis $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{2k-1}, \mathbf{f}_{2k}$ such that $Q(\mathbf{f}_{2i-1}, \mathbf{f}_{2i}) = 1 = Q(\mathbf{f}_{2i}, \mathbf{f}_{2i-1})$ for $i = 1, \ldots, k$, and $Q(\mathbf{f}_i, \mathbf{f}_j) = 0$ for all other pairs of indices. In the coordinate system corresponding to this basis, the form Q is given by (i). □

Whenever Q is given by the formula (i), let us call the basis a **Darboux basis**[5] of Q.

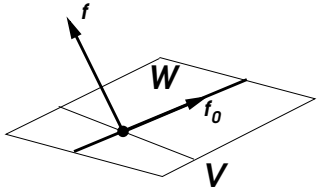

Figure 41

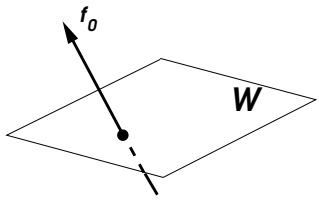

Figure 42

Consider now the case of *odd non-degenerate* forms. Let $\mathcal{W} \subset \mathbb{Z}_2^n$ denote the subspace given by one linear equation $Q(\mathbf{x}) = 0$. It has dimension $n-1$. The restriction to it of the bilinear form Q is even, but possibly degenerate. Consider vectors $\mathbf{y}$ Q-orthogonal to all vectors from $\mathcal{W}$. They are given by the system of $n-1$ linear equations: $Q(\mathbf{w}_1, \mathbf{y}) = \cdots = Q(\mathbf{w}_{n-1}, \mathbf{y}) = 0$, where $\mathbf{w}_1, \ldots, \mathbf{w}_{n-1}$ is any basis of $\mathcal{W}$. Since Q is non-degenerate, and $\mathbf{w}_i$ are linearly independent, these linear equations are also independent, and hence the solution space has *dimension* 1. Let $\mathbf{f}_0$ be the non-zero solution vector, i.e. $\mathbf{f}_0 \neq \mathbf{0}$, and $Q(\mathbf{f}_0, \mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{W}$. There are two cases: $\mathbf{f}_0 \in \mathcal{W}$ (Figure 41) and $\mathbf{f}_0 \notin \mathcal{W}$ (Figure 42).

In the first case, $\mathbf{f}_0$ spans the kernel of the form Q restricted to $\mathcal{W}$. We pick $\mathbf{f}$ such that $Q(\mathbf{f}, \mathbf{f}_0) = 1$. Such $\mathbf{f}$ exists (since Q is non-degenerate), but $\mathbf{f} \notin \mathcal{W}$, i.e. $Q(\mathbf{f}) = 1$. Let $\mathcal{V}$ consist of all vectors of $\mathcal{W}$ which are Q-orthogonal to $\mathbf{f}$. It is a subspace of codimension 1 in $\mathcal{W}$, which does not contain $\mathbf{f}_0$. Therefore the restriction of Q to $\mathcal{V}$ is non-degenerate (and even). Let $\{\mathbf{f}_1, \ldots, \mathbf{f}_{2k}\}$ (with $2k = n-2$) be a Darboux basis in $\mathcal{V}$. Then $\mathbf{f}, \mathbf{f}_0, \mathbf{f}_1, \ldots, \mathbf{f}_{2k}$ form a basis in $\mathbb{Z}_2^n$ such that in the corresponding coordinate system:

$$Q = xy + xy_0 + x_0y + \sum_{i=1}^{k} (x_{2i-1}y_{2i} + x_{2i}y_{2i-1}). \qquad \text{(ii)}$$

In the second case, $Q(\mathbf{f}_0, \mathbf{f}_0) = 1$, so that the restriction of Q to $\mathcal{W}$ is non-degenerate (and even). Let $\{\mathbf{f}_1, \ldots, \mathbf{f}_{2k}\}$ (with $2k = n-1$) be a Darboux basis in $\mathcal{W}$. Then $\mathbf{f}_0, \mathbf{f}_1, \ldots, \mathbf{f}_{2k}$ form a basis in $\mathbb{Z}_2^n$ such that in the corresponding coordinate system:

$$Q = x_0y_0 + \sum_{i=1}^{k} (x_{2i-1}y_{2i} + x_{2i}y_{2i-1}). \qquad \text{(iii)}$$

[5] After a French mathematician Jean-Gaston **Darboux** (1842–1917).

Corollary. ***A non-degenerate symmetric bilinear form in*** $\mathbb{Z}^n$ ***is equivalent to (iii) when*** n ***is odd, and to one of the forms (i) or (ii) when*** n ***is even.***

EXERCISES

305. For $\mathbf{x} \in \mathbb{Z}_2^n$, show that $\sum a_i x_i^2 = \sum a_i x_i$.

306. Show that on $\mathbb{Z}_2^2$, there are 4 non-degenerate symmetric bilinear forms, and find how they are divided into 2 equivalence classes. ϟ

307. Let Q be a quadratic form in n variables $x_1, \dots, x_n$ over $\mathbb{Z}_2$, i.e. a sum of monomials $x_i x_j$. Associate to it a function of $\mathbf{x}, \mathbf{y} \in \mathbb{Z}_2^n$ given by the formula: $B_Q(\mathbf{x}, \mathbf{y}) = Q(\mathbf{x} + \mathbf{y}) + Q(\mathbf{x}) + Q(\mathbf{y})$. Prove that B_Q is an even bilinear form.

308.** Let $\mathcal{Q}$ and $\mathcal{B}$ denote vector $\mathbb{Z}_2$-spaces of quadratic and symmetric bilinear forms on $\mathbb{Z}_2^n$ respectively. Denote by $p : \mathcal{Q} \to \mathcal{B}$ the linear map $Q \mapsto B_Q$, and by $q : \mathcal{B} \to \mathcal{Q}$ the linear map that associates to a symmetric bilinear form B the quadratic form $Q(\mathbf{x}) = B(\mathbf{x}, \mathbf{x})$. Prove that the range (resp. kernel) of p (of q) coincides with the kernel (range) of q (of p).

The Case of $\mathbb{K} = \mathbb{Q}$

All previous problems of this section belong to Linear Algebra, which is Geometry, and hence are relatively easy. Classification of *rational* quadratic forms belongs to Arithmetic and is therefore much harder. Here we can only hope to whet reader's appetite for the theory which is one of the pinnacles of classical Number Theory, and refer to [5] for a serious introduction.

Of course, every non-degenerate quadratic form Q on $\mathbb{Q}^n$ has a Q-orthogonal basis and hence can be written as

$$Q = a_1 x_1^2 + \dots + a_n x_n^2,$$

where $a_1, \dots, a_n$ are non-zero rational numbers. Furthermore, by rescaling x_i we can make each a_i integer and *square free* (i.e. expressed as a signed product $\pm p_1 \dots p_k$ of *distinct* primes). The problem is that such quadratic forms with different sets of coefficients can sometimes be transformed into each other by transformations mixing up the variables, and it is not obvious how to determine if this is the case. A necessary condition is that the inertia indices of equivalent rational forms must be the same, since such forms are equivalent over $\mathbb{R}$. However, there are many other requirements.

To describe them, let us start with writing integers and fractions using the *binary number system*, e.g.:

$$2009_{(10)} = 11111011001_{(2)}, \ -\frac{1}{3} = -.010101\ldots_{(2)}.$$

We usually learn in school that every rational (and even real) number can be represented by binary sequences, which are either finite or *infinite to the right*. What we usually don't learn in school is that rational numbers can also be represented by binary sequences *infinite to the left*. For instance,

$$-\frac{1}{3} = \frac{1}{1-2^2} = 1 + 2^2 + 2^4 + 2^6 + \cdots_{(10)} = \ldots 1010101._{(2)}$$

For this, one should postulate that powers 2^k of the base become smaller (!) as k increases, and moreover: $\lim 2^k = 0$ as $k \to +\infty$. Just as the *standard algorithms* for the addition and multiplication of finite binary fractions can be extend to binary fractions infinite to the right, they can be extended to such fractions infinite to the left. While the former possibility leads to completing the field $\mathbb{Q}$ into $\mathbb{R}$, the latter one gives rise to another completion, denoted $\mathbb{Q}_{(2)}$. In fact the same construction can be repeated with any *prime base* p each time leading to a different completion, $\mathbb{Q}_{(p)}$, called the **field of p-adic numbers**.

If two quadratic forms with rational coefficients are equivalent over $\mathbb{Q}$ they must be equivalent not only over $\mathbb{R}$ (denoted in this context by $\mathbb{Q}_{(\infty)}$), but also over $\mathbb{Q}_{(p)}$ for each $p = 2, 3, 5, 7, 11, \ldots$.

Classification of quadratic forms over each $\mathbb{Q}_{(p)}$ is relatively tame. For instance, it can be shown that over $\mathbb{Q}_{(2)}$, there are 16 (respectively 15 and 8) equivalence classes of quadratic forms of rank r when $r > 2$ (respectively $r = 2$ and $r = 1$). However, the classification of quadratic forms over $\mathbb{Q}$ is most concisely described by the following celebrated theorem.[6]

`Theorem.` ***Two quadratic forms with rational coefficients are equivalent over $\mathbb{Q}$ if and only if they are equivalent over each $\mathbb{Q}_{(p)}$, $p = 2, 3, 5, 7, \ldots, \infty$.***

These infinitely many equivalence conditions are not independent. Remarkably, *if all but any one of them are satisfied, then the last one is satisfied too*. It follows, for example, that if two rational quadratic forms are equivalent over every p-adic field, then they are equivalent over $\mathbb{R}$.

EXERCISES

309. Show that in $\mathbb{Q}_{(2)}$, the field of 2-adic numbers, $-1 = \ldots 11111$.

310.* Compute the (unsigned!) binary representation of $1/3$ in $\mathbb{Q}_{(2)}$. ϟ

311.* Prove that every non-zero 2-adic number is invertible in $\mathbb{Q}_{(2)}$. ϟ

312.* Prove that a 2-*adic unit* $\cdots * * * 1.$ (where $*$ is a wild card) is a square in $\mathbb{Q}_{(2)}$ if and only if it has the form $\cdots * * * 001$.

313.* Prove that over the field $\mathbb{Q}_{(2)}$, there are 8 equivalence classes of quadratic forms in one variable. ϟ

[6] This is essentially a special case of the **Minkowski–Hasse theorem** named after Hermann **Minkowski** (1864–1909) and Helmut **Hasse** (1898–1979).

314.* Let $\mathbb{K}$, $\mathbb{K}^\times$, $(\mathbb{K}^\times)^2$ and $\mathcal{V} := \mathbb{K}^\times/(\mathbb{K}^\times)^2$ stand for: any field, the set of non-zero elements in it, complete squares in $\mathbb{K}^\times$, and equivalence classes of all non-zero elements modulo complete squares. Show that $\mathcal{V}$, equipped with the operation, induced by multiplication in $\mathbb{K}^\times$, is a $\mathbb{Z}_2$-vector space. Show that when $\mathbb{K} = \mathbb{C}, \mathbb{R}$, or $\mathbb{Q}_{(2)}$, $\dim \mathcal{V} = 0, 1$ and 3 respectively.

315. For quadratic form Q and Q' in n variables with coefficients from $\mathbb{Z}$, prove that if they can be transformed into each other by linear changes of variables with coefficients in $\mathbb{Z}$, then $\det Q = \det Q'$. (Thus, $\det Q$, called the **discriminant** of Q, depends only on the equivalence class of Q.)