What is DCT?

High-School Trigonometry

And Undergraduate Linear Algebra

In Digital World

ALEXANDER KHEYFITS

Module Description Information

- **Title:**
  What is DCT? High-School Trigonometry and Undergraduate Linear Algebra in Digital World

- **Author:**
  Alexander I. Kheyfits, Department of Mathematics and Computer Science, Bronx Community College of the City University of New York

- **Abstract:**
  Our world has become digital. In this educational module we discuss a few basic concepts of Linear Algebra, which together with high-school trigonometry and some more advanced mathematics let us to digitize the enormous amounts of audio and visual information and to transmit it worldwide and beyond without noticeable distortion and delay.

- **Informal Description:**
  "Digital" is one of the current buzz-words, but what is its real meaning and why do we need it? In this module we describe at non-technical level, why people have to digitize the information and how this works. In short, when we transmit information by making use of the classical (=*analogous*) technology, say radio waves, we must transmit infinitely many values of the signal (the function), that is infinitely many real numbers. Even after rounding off, we still have infinitely many numbers to transmit. These procedures inevitably introduce errors, distortions ("noise"), which limit the quality of transmission.

  However in any particular application, the signals we must transmit, occupy only a limited range of values. For instance, our human ears can hear only sounds with the frequencies between about 12 and 20000 cycles per second. Therefore, it is enough to transmit only the sounds within this range of frequencies. What is more, still a century ago mathematicians discovered certain mathematical tools, which allow us to recover such signals (of course under certain assumptions, since there is no universal theorem!) from only finitely many values. Converting the real numbers into binary system, we see why we have to use the digital technology - it is enough to transmit only *finitely many* digits!

  Essentially, we *compress information*, that is, use smaller number of digits to transfer the "same" information. Certainly, we loose some information, but at some acceptable level.

Thus, digital technology allows us to drastically decrease the amount of information to be transmitted and improve the quality of transmission.

In the module, we explain more specifically, but without going into deep technicalities, within the power of the undergraduate and even well-motivated high-school students, some mathematics used to digitize and compress information before transmission and recover the information transmitted. Many problems with and without solutions and numerical examples are considered, which can be used by the instructors and students to get some hands-on experience.

- **Target Audience:** The undergraduate and motivated high-school students. The module is aimed at the students in an Introductory Linear Algebra class. It can be used as an addition to any standard Linear Algebra textbook. It can be also used by mathematically inclined students in engineering and communication theory, and motivated high-school students studying trigonometry.

- **Prerequisites:**
  The formal prerequisites include only the high-school algebra and basic trigonometry; all the necessary definitions are given and explained in the module.

- **Mathematical Fields:**
  Linear algebra and its applications; Fourier analysis; Interpolation and sampling

• **Application Areas:**
Communication theory; Data compression; Data transmission

• **Mathematics Subject Classification:**
MSC (2010) 15-01; 42A38; 94A08

• **Key Words:**
Linear Space, Basis, Interpolation and Sampling; Discrete Cosine Transform

• **Contact Information:**
Alexander I. Kheyfits
Department of Mathematics and Computer Science
Bronx Community Science (CUNY)
718-289-5616
akheyfits@gc.cuny.edu; alexander.kheyfits@bcc.cuny.edu

• **Other modules related to this module:**
Donna Beers and Catherine Crawford, "*Connecting Forensics and Linear Algebra*", CCICADA educational module, 2015.
There may be others of which we are not yet aware.

Acknowledgements:

• This text was started during the *Reconnect*-2014 workshop at the Massachusetts Maritime Academy in June of 2014. The author wants to express his gratitude to the staff of DIMACS Center at Rutgers University and to the MMA, and especially to Professor Margaret Cozzens for perfect organization of the workshop and interesting stimulating lectures.

• The author is thankful to Dr. Quanlei Fang for careful reading of the manuscript and thoughtful remarks and to Professors Donna Beers and Catherine Crawford for useful discussions at the initial stage of this project.

• My sincerest thanks go to Dr. Noah Heller for inviting me to present this topic at the Math for America workshop and to participants of this workshop for many interesting questions.

Contents

1. Introduction

When I speak with my students in the classroom, we easily hear each other. The *vocal folds*, (thin membranes, also called vocal cords or voice reeds) vibrate and push the air in our throats, creating waves of acoustical pressure, which reach ears of the people nearby and force their *tympanic membranes* to oscillate, thus creating certain signals going through the nerves, which our brains can recognize and interpret.

But if I am in New York City and want to talk to my friend in San Francisco, I physically unable to create acoustical waves powerful enough to reach California; and if I could, it would be a disaster in New York City. However, telephone and radio invented more than a century

ago, can carry sounds over big distances. These devices first change acoustical waves to electromagnetic waves, which cannot be heard but can go very far away, even around Earth and farther on, and then the electromagnetic waves are transformed back into acoustic waves, which human beings can hear.

Any such transformation inevitably introduces certain distortions, noise, which together with electro-magnetic fields in atmosphere destroy, partially or completely the information (voice, music, etc.) carried by waves. That is why we hear hissing and crackling in the old-fashioned telephone, or noise when we listen to the old gramophone records. Digital technology, together with certain other inventions, drastically decreases the level of noise in the transmitted signals. The technology is based on beautiful mathematical notions and results. The goal of this text, aimed at the high-school and undergraduate students and teachers, is to explain some mathematics, from the high-school trigonometry to college linear algebra and complex analysis, hidden behind and inside the digital technology.

The exposition is inevitably sketchy. The reader wishing to study these issues in more detail, can read, for example, a comprehensive treatise, of more than 1300 pages, by Salomon and Motta [7], and the references therein.

2. Linearity in Non-Linear World

Our world is non-linear; indeed, nobody expects that after consuming twice the amount of food, we will be able to lift double weight or to run twice faster. However, the tremendous successes of classical mathematics in sciences and technology during the previous four centuries are based on a simple *linear* mathematical concept, introduced by Isaac Newton and Gottfried Leibniz, called the differential of a function, which is the *linear part* of the increment of a function. Thus, we are led to describe the *linearity* in mathematical terms.

We begin by reviewing a few basic concepts of Linear Algebra; the latter is a part of mathematics studying linear structures, linear transformations, and their properties. For the reader's convenience and to make the text more self-contained, we remind some important definitions and include a few brief mathematical intermezzos. The reader can find more details in any *Linear Algebra* textbook. If the reader is familiar or is bored with this material, she can skip it.

2.1. **Crash Introduction to Linear Spaces.**

The classical wisdom claims that one cannot "add apples and pears". Indeed, we cannot say that 2 apples and 3 pears together make 5 apples, or 5 pears.

However, we can combine them in a more general category and say that 3 pears and 2 apples make 5 fruits. A mathematical framework for this procedure is a *Linear Space*.

Consider a set $\mathbf{V}$ of certain entities, called hereafter *vectors*, and another set $\mathbf{S}$, whose elements are called *scalars*. We assume that $\mathbf{S}$ is a *field*, that is, we can perform with the scalars the four basic arithmetic operations, *addition, subtraction, multiplication, and division*, and these operations verify certain standard properties, in particular, the addition and multiplication are *commutative, associative*, connected by the *distributive* rule, etc. The common examples are the fields of real and complex numbers.

Problem 1. *Find in your textbook or online the exact definition of the field.*

Less familiar example is the *two-element* field $Z_2 = \{0, 1\}$; which is also the simplest *Boolean Algebra*. The addition and multiplication in Z_2 are given by the *disjunction* $aVb = \max(a, b)$ and the *conjunction* $a\Lambda b = \min(a, b)$.

Problem 2. *Verify the field axioms for* Z_2. *What are the* inverse operations *, that is, the* subtraction *and* division *in* Z_2*?*

In our examples $\mathbf{S}$ is always the field of real numbers $\mathbb{R}$.

Definition 1. *The set* $\mathbf{V}$ *is called a* linear space *(or a* vector space*, which is the same) over the field of scalars* $\mathbf{S}$*, if the following* axioms *are valid.*

Vectors can be added, that is, for any two vectors $\mathbf{v}_1 \in \mathbf{V}$ *and* $\mathbf{v}_2 \in \mathbf{V}$*, there exists the unique vector* $\mathbf{v} \in \mathbf{V}$*, called their* sum *and denoted as* $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$*. The sum does not have to be the arithmetic sum of two numbers; we just use the old term (sum) in a new meaning, because the properties are similar. Thus, the nature of these objects (apples, pears, numbers, or anything else) is immaterial. We assume that the sum possesses the following properties.*

1) It is commutative*, that is, for any* $\mathbf{v}_1, \mathbf{v}_2 \in \mathbf{V}$*,* $\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{v}_2 + \mathbf{v}_1$.
2) It is associative*, that is, for any* $\mathbf{v}_1, \mathbf{v}_2$*, and* $\mathbf{v}_3 \in \mathbf{V}$*,*

$$(\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3 = \mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3).$$

3) There exists the unique neutral element*,* $\mathbf{0} \in \mathbf{V}$*, called the* zero *or the* origin *of the space* $\mathbf{V}$*, such that* $\mathbf{0} + \mathbf{v} = \mathbf{v} + \mathbf{0} = \mathbf{v}$ *for every vector* $\mathbf{v} \in \mathbf{V}$.

4) Each vector $\mathbf{v} \in \mathbf{V}$ *has the unique* opposite *vector* $-\mathbf{v} \in \mathbf{V}$*, such that*

$$\mathbf{v} + (-\mathbf{v}) = \mathbf{0}.$$

5) We suppose also that vectors can be multiplied by scalars, that is, to any vector $\mathbf{v} \in \mathbf{V}$ *and to each scalar* $\mathbf{s} \in \mathbf{S}$*, there corresponds the unique vector* $\mathbf{sv} \in \mathbf{V}$*, called their* product*. For any two scalars* $\mathbf{s}, \mathbf{t} \in \mathbf{S}$ *and any vector* $\mathbf{v}$ *the product satisfies* $\mathbf{s}(\mathbf{tv}) = (\mathbf{st})\mathbf{v}$*, and* $\mathbf{1v} = \mathbf{v}$*, where* $\mathbf{1}$ *is the unit element of the field* $\mathbf{S}$*.*

6) The multiplication of a vector by a scalar *is* distributive *with respect to vectors and with respect to scalars, that is,*

$$\mathbf{s}(\mathbf{v}_1 + \mathbf{v}_2) = \mathbf{sv}_1 + \mathbf{sv}_2$$

and

$$(\mathbf{s}_1 + \mathbf{s}_2)\mathbf{v} = \mathbf{s}_1\mathbf{v} + \mathbf{s_2v}.$$

Definition 2. ***1)*** *The vector*

$$\mathbf{v} = \mathbf{s}_1\mathbf{v}_1 + \mathbf{s_2}\mathbf{v}_2 + \cdots + \mathbf{s_k}\mathbf{v}_k$$

is called the linear combination of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ *with the (scalar) coefficients* $\mathbf{s}_1, \ldots, \mathbf{s}_k \in \mathbf{S}$*.*

2) *Vectors* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ *are called linearly independent, if any their linear combination is not zero unless all the coefficients are zero. Otherwise, that is, if there exist* k *scalars, not all zero, such that the corresponding linear combination of these vectors vanishes, the vectors are called linearly dependent.*

3) *The system of vectors* $\mathbf{A} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ *in a linear space* $\mathbf{V}$ *is called minimal or linearly independent, if none of its terms can be written as a linear combination of the other vectors of the system.*

4) *The set of vectors* $\mathbf{A} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ *in a linear space* $\mathbf{V}$ *is called a spanning system of the space* $\mathbf{V}$*, if every vector of the space can be written as a linear combination of the vectors of* $\mathbf{A}$*.*

5) *A minimal spanning system of vectors in a linear space is called a basis (of this space).*

6) *If a linear space has a basis consisting of finitely many vectors, the space is called finitely dimensional.*

Theorem 1. *All bases in a finitely dimensional space* $\mathbf{V}$ *consist of the same number of vectors. This number is called the dimension of the space* $\mathbf{V}$*.*

The simplest and most familiar example of a linear space is the family of geometric vectors in $\mathbb{R}^2$. Fix a plane, which represents all the Euclidean planes in the world, and two perpendicular lines in this plane, a horizontal line, which we call the $\mathcal{X}$ coordinate axis or the *abscissa*

axis, and a vertical line, called the $\mathcal{Y}$, or the *ordinate* axis. Their crossing point $\mathcal{O}$ is the *origin* of the system of coordinates. Now, if we pick any point $\mathcal{P}$ in the plane and consider its *orthogonal projections*[1] onto the coordinate axes, we find the two numbers, $x_{\mathcal{P}}$ and $y_{\mathcal{P}}$, called the coordinates of the point $\mathcal{P}$.

The pair

$$(x_{\mathcal{P}},\ y_{\mathcal{P}})$$

is called an *ordered pair*, since the pair $(x_{\mathcal{P}},\ y_{\mathcal{P}})$ represents another point (unless $(x_{\mathcal{P}} = y_{\mathcal{P}})$).

Given a coordinate system, any point in the plane can be represented as an *ordered pair* of real numbers $\mathcal{P} = (x_{\mathcal{P}}, y_{\mathcal{P}})$. And vice versa, for *every* ordered pair of real numbers $(x,\ y)$, we can find the *unique* point in the plane, whose coordinates are these numbers.

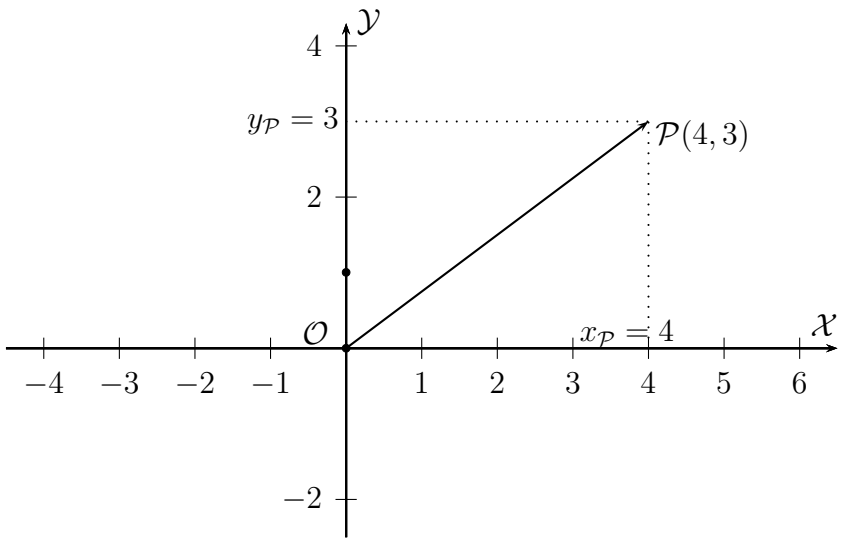

FIGURE 1. Geometric vectors and their projections in $\mathbb{R}^2$.

The vectors $\imath = (1, 0)$ and $\jmath = (0, 1)$ make the *standard basis* in $\mathbb{R}^2$. Any vector $\mathbf{v} = (x, y)$ can be written as a linear combination of the basis vectors $\mathbf{v} = x\imath + y\jmath)$, and this representation for each vector is unique.

Problem 3. *Prove that the pairs* $\imath, -\jmath$ *and* $\imath, \imath + \jmath$ *make other bases in* $\mathbb{R}^2$. *Give two more examples of bases in* $\mathbb{R}^2$.

[1]It is useful in many problems to dispense with the orthogonality, but the orthogonal coordinate systems are easier to work with.

The same procedure works in any $n-$dimensional vector space, where each vector can be written as an ordered $n-$tuple of its coordinates, that is, of its projections onto coordinate vectors.

Problem 4. *A triple of vectors* $\{i = (1, 0, 0); j = (0, 1, 0); k = (0, 0, 1)\}$ *is called the standard basis in* $\mathbb{R}^3$. *Prove that it is indeed a basis.*

Problem 5. *The Treasury issues the following coins: 1 penny, 1 nickel, 1 dime, 1 quarter, 1 half-dollar, and 1 dollar. Can we describe the monetary system of this example as a linear space? Which sub-systems of the system above make spanning systems in the set of all possible amounts of money? Which sub-systems are minimal? Are there bases among these sub-systems?*

Problem 6. *Answer the same questions with regard to the paper bills in existence.*

2.2. **Examples - $n-$Dimensional Vectors and Matrices.** We have considered in the previous section the familiar objects – two-dimensional and three-dimensional vectors, making the classical Euclidean spaces $\mathbb{R}^2$ and $\mathbb{R}^3$, which were studied in school geometry. Quite similarly, it is possible to consider *ordered* collections of n items, often called $n-$tuples or $n-$vectors. If the items are real numbers, the set of all these $n-$vectors is called the $n-$dimensional *Euclidean* space and denoted as $\mathbb{E}^n$ or[2] $\mathbb{R}^n$; we will use the latter symbol.

Problem 7. *Prove that for any* $n = 1, 2, 3, \ldots$, $\mathbb{R}^n$ *is an* $n-$*dimensional vector space. Its standard basis consists of unit vectors (called* orts*)* $(e_1, e_2, \ldots, e_n)$, *where the vector* e_k, $k = 1, 2, \ldots, n$ *has a* 1 *at the* k^{th} *place, and all its other components are* 0.

Problem 8. *Give other examples of bases in* $\mathbb{R}^n$.

As the next example, we consider matrices *with real entries. We remind that a* $k \times l$ *matrix is a rectangular array consisting of* k *rows and* l *columns,* $k \cdot l$ *elements in total. Denote the set of all* $k \times l$ *matrices with real elements as* $\mathbb{M}^{k \times l}$.

Problem 9. *Prove that the set* $\mathbb{M}^{k \times l}$ *is* $k \cdot l-$*dimensional vector space. Find its dimension and the standard basis. Find two other bases in this space.*

Prove that the space $\mathbb{M}^{k \times l}$ *of* $k \cdot l-$*matrices can be put in a one-to-one correspondence with the Euclidean space of vectors* $\mathbb{R}^n$ *of dimension* $n = k \cdot l$.

[2] $\mathbb{E}$ stands for Euclides and $\mathbb{R}$ for the real numbers.

2.3. **Two-pixel Example.** This very simplistic example will be useful when later on we study how images are appeared on TV or computer screens. Consider a rectangle consisting of two unit squares with exactly one common side, see Fig. 2, where each square can be either black (B) or white (W).

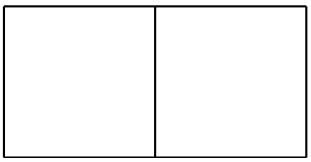

FIGURE 2. Two-pixel Example.

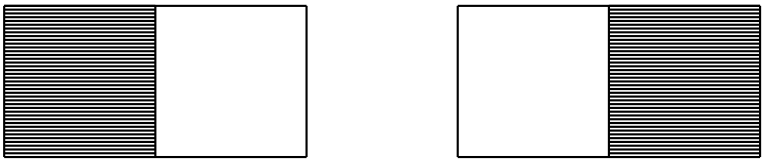

FIGURE 3. The Basis Vectors $(B-W)$ on the Left and $(W-B)$ on the Right for the Two-pixel Example.

Therefore, there are $2 \times 2 = 4$ possible configurations, $(B-B), (B-W), (W-B)$, and $(W-W)$. This finite, 4-element set can be made into a two-dimensional vector space over Z_2 (this set was defined in the paragraph between Problems 1 and 2) as the field of scalars, with the basis vectors $(B-W)$ and $(W-B)$ shown in Fig. 3, and operations induced from Z_2. For example, $(B-B) = 1\Lambda(W-B)V1\Lambda(B_W)$.

Problem 10. *Verify the other properties of the vector spaces in this example.*

This example shows that if we split any picture into small parts (pixels) we can treat any image as a vector in a certain finitely-dimensional space of simple images.

2.4. **Example - RGB.** Consider another less familiar example. Isaac Newton observed in 1672, that a triangular glass prism can separate white light into 7 rainbow colors,

Red, Orange, Yellow, Green, Blue, Indigo, Violet

Therefore, these colors can be viewed as spanning vectors in the set of all colors in our world. It turned out, however, that we do not need all these 7 colors, so that this set is not minimal. It is enough to have only the three basic colors, *Red, Green,* and *Blue*, see Fig. 3; this system is called **RGB** additive system of colors [3]. For example, *Yellow* color is generated by superimposing the *Red*, and *Green*, while combining all the three basic colors we will get *White*.

However, no two of the **Red, Green, Blue** triple can make the whole rainbow, thus, these three colors make a minimal spanning system in the space of colors.

Problem 11. *Is the set of colors and their hues as a vector space?*

A caveat is that we know how to mix only colors with positive coefficients. however, the set of positive real numbers is not a field. Therefore, we cannot 'subtract' colors [4]. Whence, it is not a linear space; this algebraic system is called a *cone*.

The bottom of the human eye, the *retina*, contains special cells, called *rods* and *cones*. The rods measure the *luminance*, that is, the *intensity of light* seen by the eye. The cones measure the *chrominance*, that is, the *intensity of colors*, comprising the incoming light. Apparently, our mother-nature has known the **RGB**-theory for millennia, since there are three kinds of cones, each measuring separately the intensity of **Red**, **Green**, and **Blue** colors.

2.5. **Example - Algebraic Polynomials.** Polynomials are functions that can be represented by a finite sum of whole powers of a certain indeterminate, say, t, as

$$P(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_k t^k + \cdots + a_{p-1} t^{p-1} + a_p t^p,$$

where $a_0, a_1, \ldots, a_p$ are numerical, real or complex coefficients, and p is a nonnegative integer number, called the *degree* of polynomial $P(t)$; we assume that $a_p \neq 0$, that is, the degree is exactly p. In this text we consider only polynomials with *real coefficients* and assume that t can change in the infinite interval $-\infty < t < \infty$.

[3] There are many applets online demonstrating addition and subtraction of colors. Just type 'color addition' 'color space', and Google brings them up. You can experiment with

[4] However, together with the additive systems like **RGB**, there are *subtractive* color systems.

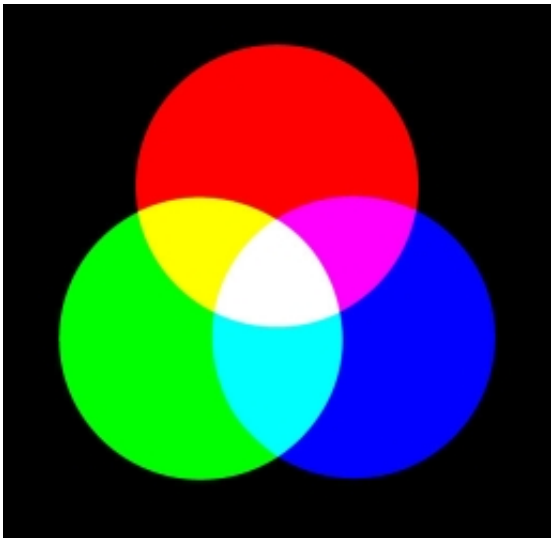

Figure 3. RGB color diagram

(http://en.wikipedia.org/wiki/*RGB_color*_model, accessed on 01/15/2015)

The family of the polynomials of degree *exactly* p is not a linear space, since, for example, the sum of two quadratic polynomials

$$(x^2 + x + 1) + (-x^2 + 1) = x + 2$$

is a linear rather than quadratic polynomial.

Problem 12. *1) Prove that the set of polynomials of degree at most n with real coefficients is a linear space of the dimension $n+1$; we denote this space as $\mathfrak{P}_n$.*

2) Prove that the system of powers $\{1, t, t^2, \dots, t^n\}$ is a minimal spanning system in $\mathfrak{P}_n$, that is, a basis. Give another example of a basis in $\mathfrak{P}_n$.

3) Is the system $\{1, t^2, t^3, \ldots, t^n\}$ *minimal in the space* $\mathfrak{P}_n$*? A spanning system in* $\mathfrak{P}_n$*? A basis?*

Problem 13. *Prove that the set of polynomials* $\mathfrak{P}_n$ *can be put in a one-to-one correspondence with the Euclidean space* $\mathbb{R}^{n+1}$*.*

By the definition of a basis, any element of a linear space can be written as a linear combination of the basis vectors; moreover, due to the minimality this combination is unique. Thus, any algebraic polynomial of degree at most n is the unique sum of the powers

$$1 = t^0, t, t^2, \ldots, t^{n-1}, t^n$$

with certain coefficients depending on the polynomial. However, if a family of vectors is a spanning but not minimal system, there may be several linear combinations representing the same vector. The coefficient of x^k is the magnitude (positive, negative, or zero) of the projection of the polynomial onto the x^k-axis, that is, onto the subspace of the monomials proportional to x^k.

Problem 14. *Check that the system* $\{1, t, 1 + t\}$ *is spanning but not minimal in* $\mathfrak{P}_1$*. thus, it is not a basis in* $\mathfrak{P}_1$*. For example, show that the* 0 *has at least two different representations through these vectors.*

On the other hand, the system $\{1, t^2\}$ *is minimal but not spanning in* $\mathfrak{P}_2$*.*

A basis of a vector space cannot *precisely* represent elements of a larger, ambient space.

Problem 15. *Prove that no linear combination of the three vectors (monomials)* $1, t, t^2$ *can be equal to* t^3 *identically for all real* t*.*

Otherwise, the cubic polynomial $at^3 + bt^2 + ct + d$ would have more than three roots, which contradicts to the Fundamental Theorem of Algebra, which implies that a non-trivial polynomial of n^{th} degree has at most n different real or complex roots.. It is sufficient also to observe that the ratio $t^3 \div t^2$ tends to infinity with t.

However, if we restrict ourselves to a bounded set of $t-$values, say, $-1 \leq t \leq 1$, we can give an estimation of the worst discrepancy between t^3 and quadratic polynomials. For instance, if we want to approximate the simplest first-degree polynomial $P(t) = t$ by zero-degree polynomials, that is, by constants $f(t) = a$, over $-1 \leq t \leq 1$, then the maximum error is $\max_{-1 \leq t \leq 1} |t - a| = |a| + 1$, and its smallest value over all the constants a is 1, attained for $a = 0$.

Problem 16. *Prove this statement.*

Using more advanced techniques, the estimate can be extended onto polynomials of arbitrary degree [5].

The algebraic polynomials are not periodic, thus they are not convenient when we deal with oscillating functions. In such problems, *trigonometric polynomials* are more suitable.

2.6. **Trigonometric Polynomials.** The algebraic polynomials are linear combinations of the monomials $t^j,\ j = 0, 01, 2, \ldots$. Similarly, any finite linear combination of the simplest trigonometric functions

$$1 = \cos(0 \times \theta),\ \sin\theta = \sin(1 \times \theta),\ \cos\theta,\ \sin(2\theta),\ \cos(2\theta), \tag{1}$$

$$\sin(3\theta),\ \cos(3\theta), \ldots, \sin(n\theta), \cos(n\theta), \ldots,$$

(we consider only the combinations with *real coefficients*) is called a *trigonometric polynomial*. The *degree* of a trigonometric polynomial is the largest among the coefficients of θ. For example,

$$5 - 6\sin 3\theta + 4\cos 2\theta \tag{2}$$

is the trigonometric polynomial of degree 3. The trigonometric function

$$\sin\theta - 2\sin^2(\theta)$$

is not a trigonometric polynomial as is, because the definition requires the exponents to be 1, but it can be transformed as

$$\sin\theta - (1 - \cos(2\theta)) = \cos 2\theta + \sin\theta - 1, \tag{3}$$

therefore, it can be written as a trigonometric polynomial of degree 2. The trigonometric function

$$\sin(\frac{1}{2})\theta \tag{4}$$

also is not a trigonometric polynomial, however, it becomes a trigonometric polynomial after a linear substitution $\theta = 2\phi$, and its smallest positive period is $2\pi \div \frac{1}{2} = 4\pi$.

Trigonometric polynomials are periodic functions; for example, the smallest positive period of polynomial (3) is 2π, which is the least common multiple of π and 2π.

Problem 17. *Find the smallest positive period of (2) and that of the general trigonometric polynomial (1) of degree n.*

[5] See, for example, [2] for the detailed exposition of these problems.

Problem 18. *Prove that if $\lambda > 0$ is a constant, then the smallest positive period of either of the functions*

$$S_\lambda(\theta) = \sin(\lambda\theta) \quad \textit{and} \quad C_\lambda(\theta) = \cos(\lambda\theta)$$

is $\frac{2\pi}{\lambda}$.

Due to this periodicity, in what follows we restrict ourselves to the interval $(-\pi, \pi)$. Introduce the set $\mathfrak{P}_m^{\sin}$, consisting of the linear combinations of the sines of degree at most m,

$$\{\sin\theta,\ \sin 2\theta,\ \sin 3\theta, \dots, \sin m\theta\}$$

and the set $\mathfrak{P}_m^{\cos}$, consisting of the linear combinations of cosines of degree at most m,

$$\{1,\ \cos\theta,\ \cos 2\theta,\ \cos 3\theta, \dots, \cos m\theta\}.$$

Problem 19. *1) Prove that the set $\mathfrak{P}_m^{\sin}$ is a linear space of the dimension m consisting of odd trigonometric polynomials, and the set $\mathfrak{P}_m^{\cos}$ is a linear space of the dimension $m+1$ consisting of even trigonometric polynomials.*

2) The function $f(x) = x$ is odd; does it belong to $\mathfrak{P}_m^{\sin}$?

3) The function $g(x) = |x|$ is even; does it belong to $\mathfrak{P}_m^{\cos}$?

4) Prove that for $0 < \theta < \pi$ the system

$$\{\sin\theta, \dots, \sin(m\theta)\}$$

is a minimal spanning system, that is, a basis in $\mathfrak{P}_m^{\sin}$, while the system

$$\{1,\ \cos\theta, \dots, \cos(m\theta)\}$$

is a basis in $\mathfrak{P}_m^{\cos}$. Give other examples of bases in these spaces.

2.7. **A Special System of Trigonometric Functions.** To develop the Discrete Cosine Transform (DCT), we need a special basis in the 8-dimensional Euclidean space $\mathbb{R}^8$.

For a positive integer n, let us consider n trigonometric functions

$$B_k(t) = \cos\left(\frac{k\pi}{n}t\right),\ k = 0, 1, \dots, n-1.$$

Since $\cos 0 = 1$, $B_k(0) = 1$ for any k and n, the function B_0 is constant, $B_0 \equiv 1$, and (see Problem 14?) every function $B_k(t),\ k \neq 0$, is periodic with the smallest positive period $2\pi \div \frac{k\pi}{n} = \frac{2n}{k}$. The functions B_k are not trigonometric polynomials, for the coefficients of t are not, in general, integer numbers, but in applications in this module they behave like the polynomials, since they can be made into polynomials by a linear change of variable – see an example after equation (4?).

We are mostly concerned with the case $n = 8$; the corresponding functions $B_0(t) - B_7(t)$ for $0 \le t \le 8$ are shown in Fig. 3 - Fig. 6. It is worth mentioning that each function $B_k,\ k = 0, 1, \ldots, 7$, has k zeros within the range $0 \le t \le 8$.

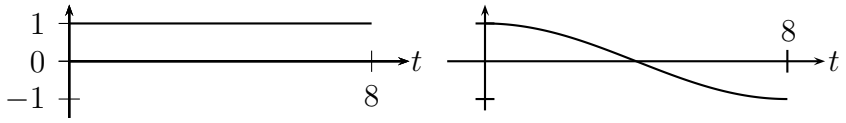

FIGURE 4. Trigonometric functions $B_0(t) = 1$ and $B_1(t) = \cos\left(\frac{\pi}{8}t\right),\ 0 \le t \le 8$.

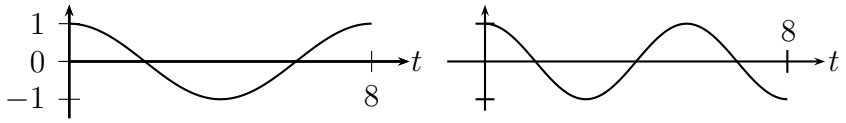

FIGURE 5. Trigonometric functions $B_2(t) = \cos\left(\frac{2\pi}{8}t\right)$ and $B_3(t) = \cos\left(\frac{3\pi}{8}t\right),\ 0 \le t \le 8$.

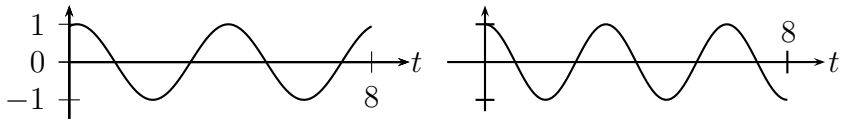

FIGURE 6. Trigonometric functions $B_4(t) = \cos\left(\frac{4\pi}{8}t\right)$ and $B_5(t) = \cos\left(\frac{5\pi}{8}t\right),\ 0 \le t \le 8$.

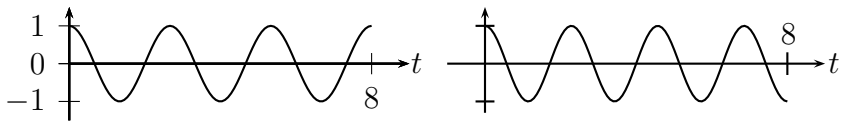

FIGURE 7. Trigonometric functions $B_6(t) = \cos\left(\frac{6\pi}{8}t\right)$ and $B_7(t) = \cos\left(\frac{7\pi}{8}t\right),\ 0 \le t \le 8$.

We need also the shifted functions $C_k(t) = B_k(t + 1/2)$, that is,

$$C_k(t) = \cos\left(\frac{k\pi}{8}(t + 1/2)\right) = \cos\left(\frac{k\pi}{8}t + \frac{k\pi}{16}\right),\ k = 0, 1, \ldots, 7, \tag{5}$$

graphed in Fig. 7 - Fig. 10. The elementary trigonometric identities

$$\cos(x \pm y) = \cos x \cos y \mp \sin x \sin y \tag{6}$$

show that these functions are linear combinations of the *stretched* trigonometric polynomials $\cos\left(\frac{k\pi}{8}t\right)$ and $\sin\left(\frac{k\pi}{8}t\right)$, $k = 1, 2, \ldots, 7$. Similarly to B_k, every function C_k, $k = 0, 1, \ldots, 7$, also has k zeros within the range $0 \le t \le 8$, or within $-1/2 \le t \le 15/2$.

Problem 20. *Prove trigonometric identities (6?)-(7?).*

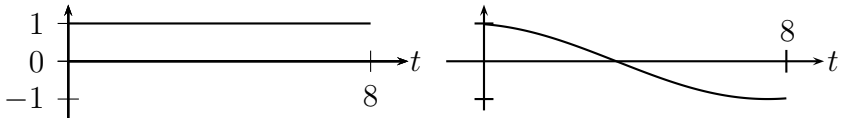

FIGURE 8. Trigonometric functions $C_0(t) = 1$ and $C_1(t) = \cos\left(\frac{t+1/2}{8}\pi\right)$, $0 \le t \le 8$.

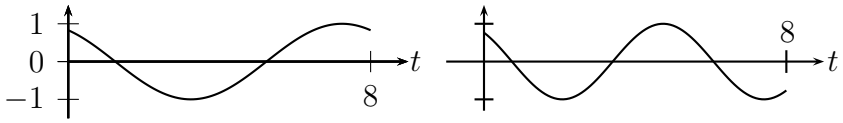

FIGURE 9. Trigonometric functions $C_2(t) = \cos\left(\frac{2(t+1/2)}{8}\pi\right)$ and $C_3(t) = \cos\left(\frac{3(t+1/2)}{8}\pi\right)$, $0 \le t \le 8$.

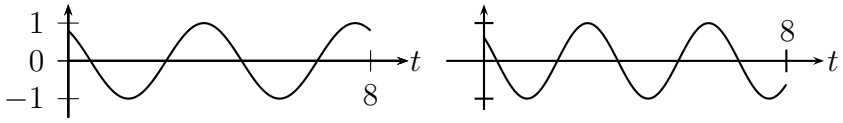

FIGURE 10. Trigonometric functions $C_4(t) = \cos\left(\frac{4(t+1/2)}{8}\pi\right)$ and $C_5(t) = \cos\left(\frac{5(t+1/2)}{8}\pi\right)$, $0 \le t \le 8$.

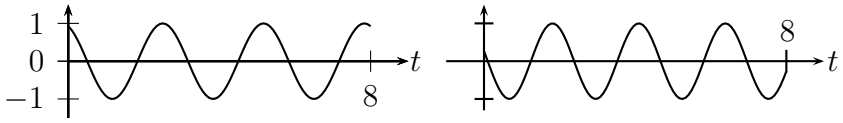

FIGURE 11. Trigonometric functions $C_6(t) = \cos\left(\frac{6(t+1/2)}{8}\pi\right)$ and $C_7(t) = \cos\left(\frac{7(t+1/2)}{8}\pi\right)$, $0 \le t \le 8$.

For non-integer α, the trigonometric function $y = \cos(\alpha t)$ is not a trigonometric polynomial, however, it preserves their many important

properties. In particular, it is periodic with the smallest positive period of $\frac{2\pi}{\alpha}$. It is even for any real α. However, $C_k(t),\ k > 0$, are not even functions; because of the shift in the argument, $t + 1/2$ instead of t. They are "even" with respect to the vertical line $t = -1/2$, that is, their graphs are symmetrical with respect to this vertical line.

One can prove in different ways that the functions $C_k,\ k = 0, 1, \ldots, 7$, are linearly independent over the interval $0 \le t \le 8$. The proof is shortest if we notice that these functions are eigenfunctions of certain ordinary differential operator [8]. An elementary, but a bit cumbersome direct proof is also possible, for there are enough convenient points between 0 and π, where the 8 cosine functions, taken at any combination, vanish.

Problem 21. *Prove directly that the system of functions* $\{C_k(t)\},\ k = 0, 1, \ldots, n-1$, *is linearly independent for* $-1/2 \le t \le 15/2$.

Problem 22. *Prove that the set of all linear combinations with real coefficients of the functions* $C_0(t) - C_7(t)$, *where* $0 \le t \le 8$, *or, what is equivalent, for* $-1/2 \le t \le 15/2$, *is a linear space; we denote it as* $\mathfrak{L}_8$.

We know which geometrical vectors in $\mathbb{R}^2$ or $\mathbb{R}^3$ are *orthogonal* (or *perpendicular*), this relationship is understood in the sense of classical Euclidean geometry. In general vector spaces, like $\mathfrak{L}_8$ though, where we do not have our natural visual intuition, the notion of *perpendicular vectors* must be explicitly defined. It turns out that the two vectors, $f(t)$ and $g(t),\ a < t < b$, must be *defined orthogonal* if the integral of their product is zero, that is,

$$\int_a^b f(t)\ g(t)dt = 0.$$

In particular, two functions $f, g \in \mathfrak{L}_8$ are orthogonal, if

$$\int_{-1/2}^{15/2} f(t)\ g(t)dt = 0. \tag{7}$$

Problem 23. *1) Prove that the two functions* B_k *and* B_l *are orthogonal if* $k \ne l$, *and are not orthogonal if* $k = l$.

2) Similarly, the functions C_k *and* C_l *with different* k *and* l *are orthogonal, and are not orthogonal if* $k = l$.

3) Hence, $C_k,\ k = 0, 1, \ldots, 7$ *make a basis in* $\mathfrak{L}_8$.

4) Are the functions B_k *and* C_l *orthogonal?*

Problem 24. *However,* $\cos(2m\pi t),\ m = \pm 1, \pm 2, \ldots$, *is orthogonal to every* C_k, *so that we cannot expand this function against the basis of* C_k *in* $\mathfrak{L}_8$. *Therefore,* $\cos(2m\pi t) \notin \mathfrak{L}_8$

3. More Mathematics Relevant to Transmitting Information. Cardinality and Transformations

Let us go back to the beginning of this module. The acoustical waves carry information and bring it to the listener's ears. Acoustical waves are oscillations of the air pressure, they evolve in time, and we can graph them in Cartesian coordinates as pressure vs. time. Thus, the acoustical waves are familiar *mathematical functions*, functions of time. However, they also carry information, and because of that, engineers call them *signals*; we will use these two terms, functions and signals, interchangeably, as synonyms.

3.1. **Intermezzo on the Cardinality of Infinite Sets.** It is a convenient place here for another mathematical intermezzo, about the number of elements of a set. We have no problem with the *finite sets*, say, the number of students in the classroom – we can just count them. However, time intervals, even as short as one second, contain *infinitely* many points, that is, real numbers marking time moments. But how to measure infinity?

We are in a situation, mathematicians have encountered many times before; indeed, we have a method for solving certain problems (find the number of elements in a finite set), but the method does not work for more general problems (for infinite sets). One way to deal with this problem is to try to invent another method, equivalent to the known one in an old, simpler setting, but working in more general cases. The procedure that works in our current problem, is called *a one-to-one correspondence*.

We can compare finite sets by the number of their elements. Say, our left hand has five fingers, and our right hand has five fingers; therefore, both hands have an equal number of fingers, or these two sets of fingers have the same number of elements. However, two sets can be compared without direct enumerating their elements, just by putting them in a *one-to-one correspondence*. Indeed, we do not have to count fingers to conclude that we have the same amount of fingers on our left hand as on the right hand; it suffices to place one hand on the other. This method works for the infinite sets as well. We can conclude, for example, that there are exactly as many even (or odd) integer numbers, as all the integer numbers, even if this may initially look counter-intuitive.

Two sets are said to have the same *cardinality*, if they can be put into a one-to-one correspondence with one another. Using this approach, we can classify *infinite sets* and compare them with regard to the "number", the cardinality of their elements. It turns out that infinite sets may have different cardinalities. Paraphrasing G. Orwell [4], "All the

infinities were created equal, but some of them are more infinite than others".

Any set, which has the same cardinality as the set of natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$, that is, again, can be put in a one-to-one correspondence with $\mathbb{N}$, is called a *countable* set. This is the smallest cardinality of an infinite set.

Problem 25. *1) Prove that the sets of even or odd integer numbers, prime numbers, perfect squares, complex numbers with integer components, integer-valued points in $\mathbb{R}^n$ are countable.*

In more advanced mathematical textbooks you can find a proof that it is impossible to put any countable set in a one-to-one correspondence with the set of real numbers, or even with the set of points comprising the interval $[0, 1]$. It is said that any non-empty interval of real numbers, like $(0, 1)$, in particular, a time interval, like *one second*, has the *cardinality of continuum*, which is incomparably more than the cardinality of the set of natural numbers. The sets of natural, or integer, or rational numbers are *countable* sets. The intervals of the real line have bigger cardinality than countable sets.

Problem 26. *What is the cardinality of the set of complex numbers? Of the set of points of the three-dimensional space? Of the polynomials with real coefficients? Of the polynomials with rational coefficients?*

3.2. **Analogous and Digital Signals.** When acoustical signals are transmitted by way of the classical radio or telephone technology (inductors-resistors-capacitors connected by wires), it is said to be an *analogous signal*. This technology has its limitations. We discussed in the previous section that any time interval contains continuum-many points, thus if we want to transmit an analogous signal *precisely*, without any distortion, we must transmit continuum-many numbers, which likely requires infinite time. Therefore, it is unfeasible to transmit the analogous signals precisely using their representation as functions of time.

There exists another representation of waves, though. Any wave, in particular, acoustical wave has also another description. Waves are functions of time, but also they are *carriers of energy*. The energy is carried through by the components, sub-waves of the signal having different *frequencies*. In musical terms, there is the basic tone and there are overtones. The bridge between the two representations is known as the *Fourier transform*, called after French physicist and mathematician Joseph Fourier. This transform represents the signals, that is, functions of time, as combinations of simpler signals having different *frequencies*. The prism of Newton that we described above, performs the Fourier

transform and shows the frequencies composing the white light. The frequencies make the *spectrum* of a signal. When we enjoy music, we hear these sounds-frequencies together; many people with good hearing can distinguish, to some extent, these frequencies. *Spectroscopes*, like Newton's prism, do it precisely.

It is known that in the first, *linear approximation* the waves of different frequencies are *linearly independent*. They, so to say, "don't see" one another, and go through one another independently, without any interference. This fundamental property of linear independence of the elementary frequencies is called the *Principle of Superposition*. And again, to use the Principle of Superposition, analogous signals must be digitized.

By *digitizing* (visual) information we mean changing its representation as continuous functions (of time, space, etc.) to *discrete* representation, that is, to strings of digits. In doing that we proceed in two steps, first, by *sampling* the original information, and second, by *quantizing* the numbers created at the first step. These steps are discussed in more detail in the following sections. *Digital technology*, which exploits these discrete representations, can transmit huge amounts of information faster and with far less distortion than analogous devices. If we would have endless resources, we could achieve perfect transmission.

3.3. **Fourier and Cosine Transform.** Mathematically, the Fourier and similar transforms are integrals of the signals as functions of time or frequencies. The kernels of these integrals contain trigonometric, that is, periodical functions. Acoustical and electromagnetic waves are also superpositions of periodical oscillations, hence it is quite natural to use these transformations to represent, to process, and to transmit acoustical and visual information.

Under some conditions, the Fourier integrals become trigonometric series, such as, for instance, the cosine-series

$$f(t) = \sum_{k=0}^{\infty} a_k \cos(kt).$$

Here only the coefficients a_k are specific for the signal $f(t)$, the functions $\cos(kt)$ are universal, linearly independent vectors in the linear space of signals.

However, the Fourier transform may be inconvenient in certain problems, for it is *complex-valued*, while physical quantities are real-valued. In many problems it is preferable to employ the real-valued "half" of the Fourier transform, that is, either the *cosine transform* or the *sine*

transform. With regard to transmitting the information, the former has definite advantages that will be discussed later. That is why many current standards for transmission and compression of information, like JPEG (Joint Photographic Experts Group), MPEG (Moving Picture Experts Group), GIF (Graphic Interchange Format), etc., are based on the discrete version of the classical cosine transform, called DCT (Discrete Cosine Transform). We first discuss the classical (continuous) cosine transform.

Given a *signal*, that is, a function $f(t)$ of a variable $t,\ 0 \le t < \infty$, its (Fourier) *cosine transform* is the function $F(\omega)$ of the frequency variable ω, given by the integral

$$F(\omega) = F[f](\omega) = \left(\frac{2}{\pi}\right)^{1/2} \int_0^\infty f(t)\cos(\omega t)dt. \tag{8}$$

For the improper integral (6) to exist, the function f must decay fast enough, otherwise we have to use distributions and other mathematical tools beyond the level of this module.

The integral transforms, like the Laplace transform, the Fourier transform, the cosine transform and others are valuable in many problems because they simplify certain advanced mathematical operations. For example, linear ordinary differential equations with constant coefficients, being Laplace transformed, become algebraic (polynomial) equations, which can be solved much easier than the differential equations. However, after solving the derived algebraic equation, we want to return to the original functions, that is, we need the *inversion formulas*. The advantage of the cosine transform (8) is that its inverse transform (9) looks exactly as the direct one given by,

$$f(t) = \left(\frac{2}{\pi}\right)^{1/2} \int_0^\infty F(\omega)\cos(\omega t)d\omega. \tag{9}$$

The factor $(2/\pi)^{1/2}$ is a normalization constant, it is chosen so that to make formulas (8)-(9) symmetric.

Problem 27. *Let f be a (fast enough) decaying function, for example, $f(t) = 1/(1+t^2)$, or $f(t) = e^{-\alpha t}$, or even a discontinuous function $f(t) = 0$ if $a < t < \infty$, and $f(t) = 1$ if $t < a < \infty$, and $f(a) = 1/2$, where a is a positive parameter. Verify formulas (8)-(9) by a direct calculation.*

The choice of the value $f(a) = 1/2$, the proof of the fact that formula (9) indeed inverts the cosine transform (8) and vice versa, and the formulation of the precise conditions of the validity of inversion (8)-(9) is beyond the scope of this text.

So far so good, but where is the digital technology, information, etc.? We are close. The reader has definitely noticed that both equations (8)-(9) involve functions defined over the infinite interval $[0, \infty)$, thus, to use the Fourier-cosine transformation, we must know continuum-many values of a function in the integrand, which is problematic. To deal with this obstacle, we firstly notice that human beings normally can hear frequencies from about 12 Hz to 20 000 Hz. One Hertz (Hz) is the frequency of one complete oscillation per second. Thus, from the point of view of human perception, we can safely cut away, just remove both 'tails', below 12 Hz and above 20 000 Hz. But there is more to the story.

3.4. **Interpolation and Sampling.** The interval, $(12, 20\ 000)$ has, as any other interval, the cardinality of continuum. Thus, we seem to have the same problem as before, namely, we have to process continuum many function values. However, mathematicians had discovered the solution still centuries back, well before the digital age, when the need to transmit huge amounts of information was not imminent. Let $x_1, x_2, \ldots, x_n, x_{n+1}$ be any pair-wise distinct real or complex numbers. Still in 1779, British mathematician Waring proved that any polynomial of degree n

$$P(x) = \sum_{k=0}^{n} a_k x^k$$

can be represented through its values at the points $x_i,\ i = 1, \ldots, n, n+1$, called *sampling* or *interpolation points* or *nodes*, by the formula

$$P(x) = P(x_1)f_1(x) + P(x_2)f_2(x) + \cdots + P(x_{n+1})f_{n+1}(x), \tag{10}$$

where

$$f(x) = a(x - x_1)(x - x_2)\cdots(x - x_n)(x - x_{n+1})$$

is a polynomial of degree $n + 1$ with roots at the nodes, $a \neq 0$ is a constant factor, and for $j = 1, 2, \ldots, n,\ n + 1$,

$$f_j(x) = \frac{f(x)}{(x - x_j)f'(x_j)} =$$

$$= \frac{(x - x_1)\cdots(x - x_{j-1})(x - x_{j+1})\cdots(x - x_{n+1})}{(x_j - x_1)\cdots(x_j - x_{j-1})(x_j - x_{j+1})\cdots(x_j - x_{n+1})}.$$

Polynomials f_j are called the *fundamental interpolation polynomials* with nodes x_j, because they depend only on the nodes x_j, but not upon any polynomial $P(x)$. Later, formula (10) was rediscovered by Euler and Lagrange (1795) independently, and is called now the Lagrange's interpolation formula.

We remark that a polynomial of n^{th} degree has $n+1$ coefficients $a_0, a_1, \ldots, a_n$, so that to find them, we need $n+1$ conditions, given by the values of $P(x)$ at the $n+1$ nodes of interpolation.

Let us consider a couple of simple special cases of (10). If $n = 1$, then the sum in (10) contains only two terms,

$$P(x) = \frac{x - x_2}{x_1 - x_2} P(x_1) + \frac{x - x_1}{x_2 - x_1} P(x_2),$$

and can be recognized as the equation of the line through the points $(x_1, P(x_1))$ and $(x_2, P(x_2))$. We see that given two points

$$(x_1, P(x_1)), (x_2, P(x_2)),$$

this formula *interpolates*, that is, recovers the values of the linear polynomial $P(x)$ at all other, even complex points x.

Let be next $n = 2$, thus, we are given 3 points, $(x_1, P(x_1)), (x_2, P(x_2))$ and $(x_3, P(x_3))$. Since the coefficients are quadratic polynomials, the formula now recovers the unique parabola, going through the given three points in the plain.

Problem 28. *If $n = 2$, there is an exceptional case, when the quadratic trinomial does not exist; more precisely, it degenerates into a simpler function. Study this case separately.*

Problem 29. *What is the meaning of the formula when $n = 0$?*

Problem 30. *Consider the case $n = 3$.*

Formula (10) evaluates the value of the polynomial $P(x)$ at any point x, through the values of P at the points $x_k,\ k = 1, 2, \ldots, n+1$, that is why it is called the *interpolating* formula. We reiterate that the rational functions in the formula are universal, in the sense that they do not depend on any polynomial but only upon the nodes of interpolation $\{x_i,\ 1 \le i \le n+1\}$. Solely the factors $P(x_i)$ depend upon the polynomial P. Thus, if the signal to be transmitted is a polynomial of $n-$th degree and the nodes $x_1, x_2, \ldots, x_{n+1}$ are known in advance, we do not have to transmit continuum, nor even countably many numbers, we are to transmit only the $n+1$ numbers $P(x_1), P(x_2), \ldots, P(x_{n+1})$, since the coefficients

$$f_1(x) = \frac{(x - x_2)(x - x_3) \cdots (x - x_{n+1})}{(x_1 - x_2)(x_1 - x_3) \cdots (x_1 - x_{n+1})},$$

$f_2(x), \ldots, f_n(x)$ are the same for any $n-$th degree polynomial and can be computed and stored in advance.

This process is called the (algebraic) Lagrange interpolation, since it involves the algebraic polynomials. However, both acoustical and optical signals are carried by waves, that is, by periodic processes; thus, we are interested in interpolation of signals by trigonometric polynomials, discussed in Subsection 5.4.

3.5. **Intermezzo about the WKS Theorem.** Formulas, similar to (8), are valid for functions, which are much more general than polynomials. The following result, stated here without some details, is often called the Whittaker-Kotel'nikov-Shannon (WKS) sampling theorem, after the three mathematicians, who independently discovered it in 1915, 1933, and 1949 years, respectively. It is also called Nyquist or Ogura formula.

Theorem 2. *If f is a square-integrable function such that the Fourier transform $(F[f])(\omega) = 0$ whenever*[6] *$|\omega| > L > 0$, then for every real or complex number x*

$$f(x) = \sum_{-\infty}^{\infty} f\left(\frac{n\pi}{L}\right) \frac{\sin(Lx - n\pi)}{Lx - n\pi}. \tag{11}$$

This result states that under the conditions above, which verify in many important applications, a continuous signal f, defined over a finite or infinite interval, can be *precisely* recovered by making use of the interpolation series (11) (called also *cardinal* series). The coefficients

$$\frac{\sin(Lx - n\pi)}{Lx - n\pi},$$

similarly to formula (10), do not depend on the function f, they can be stored in hardware at the receiver's end in advance. Only the signal values at the sampling nodes $\frac{\pi}{L}n$ do matter. Thus, if we want to transmit the signal f, we have to send only a countable set (not a continuum!) of the function values $f(n\pi/L)$, $-\infty < n < \infty$.

What is more, if this is an acoustical signal, human beings do not hear the higher frequencies anyway, so that we have to deal only with frequencies limited to finite interval, and after the sampling we get a *finite* set of values of the signal. We have to transmit only a finite set of sampled function values. And even if the information is distributed over the whole number line, the information carried by higher frequencies is often either completely negligible as in the case of speech, or can be neglected without a noticeable distortion of the result.

[6]Such functions are called "band-limited functions." A function $f(x)$ is called 'square-integrable' on the real axis, if the improper integral $\int_{-\infty}^{\infty} |f(x)|^2 dx$ is convergent.

An important corollary of the WKS-Nyquist theorem claims that if the transmitted signal is composed of signals with different frequencies, and the largest of these frequencies is f_{max}, then to insure perfect (lossless) reconstruction, we can *sample* the signal with any frequency bigger than $2f_{max}$, called the *Nyquist frequency*.

We presented here only a simplest version of the WKS-theorem. It can be extended onto more general classes of functions, than the band-limited functions; the sampling points can repeat and do not have to make an arithmetic progression, and other restrictions can be relaxed. The electromagnetic waves, can be treated by similar techniques. In the sequel sections we are concerned with transmission of visual information. We show in more detail how certain mathematical ideas, in particular, linearity, trigonometry, and interpolation work in the image transmission.

4. Transmitting Optical Images

Our world is filled with many colors, however, as was explained above, to make any color, it is enough to have just three basic ones – **R**ed, **G**reen, and **B**lue, abbreviated as **RGB**. It is clearly seen when one prints color pictures by making use of a modern photo-printer - paper goes through the printer three times. After the first pass we see a one-colored (monochromatic) image, where it is difficult to recognize the picture is being printed. After the second pass, the picture is more recognizable, but the colors still look distorted. After the third pass the good picture, close to the original one, appears on paper.

Hence, to get a full-colored image, it is enough to decompose a picture into three monochromatic images, corresponding to the three basic colors, then to transmit each of these three components separately, and finally to assemble these three images into one color picture. Because of that, from now on we consider the transmission of monochromatic images. Moreover, we can safely assume that we transmit a black-and-white picture, where every point has a certain degree (or shade) of *grayness*.

However, as we discussed above, the analogous technology cannot deliver the signals of very high quality and cannot transmit very large amounts of the information with big speed, which is required by modern life. It turns out though that the digital technology is able to resolve these issues. Hence, the question is, how to 'digitize' the image, so that we are to transmit not analogous signals, but the *digits*, the strings of zeros 0 and ones 1. A very few words about technology are in order.

4.1. **Pixels.** The screen of any modern device (camera, computer monitor, TV set, cell phone, etc.) physically consists of many very small elements, like dots, called *pixels* (=picture elements). When we shot a picture of an object, the light reflected by the object, falls on the screen of the camera, and every pixel receives its portion of the light, that is, a certain amount of energy.

The electronics behind (or inside) the screen measures the *luminance*, which is a technical term for the intensity of grayness of every pixel measured on scale from 0 to 1, where 0 and 1 denote the extreme cases; 1 usually corresponds to the absolutely black pixel, because it did not get any energy, and 0 to the absolutely white pixel, but it can be another way. This scale is divided into $256 = 2^8$ levels (shades) of grey. The electronics measures the luminance of every pixel as a whole number between 0 and 255 inclusive. Therefore, the system creates the array (matrix) of whole numbers between 0 and 255. Each pixel on the screen is represented by an entry of the matrix. For example, if the intensity of a pixel is measured as a half of the maximal intensity, the corresponding entry of the matrix is $\frac{1}{2}256 - 1 = 127$. This matrix shows the luminance of the screen image.

The entries of the matrix are written in the binary place-value system, that is, as 8-element strings of 0s and 1s, from $0_{10} = 00000000_2$ through $255_{10} = 11111111_2$, where the subscript indicates the base, decimal $_{10}$ or binary $_2$, of the number system used. If the image is colored, electronics creates three matrices, representing the intensity (*chrominance*) of the three basic colors (**RGB**) for each pixel of the screen. Thus, from mathematical point of view, every pixel is an $8-$digit, or $8-$bit binary integer number.

Problem 31. *Prove that there are exactly* $2^8 = 256$ *different* $8-$*digit binary numbers.*

If you have a TV with the resolution of 1920×1080, which means 1920 pixels per row and 1080 pixels per column, the screen contains 2 073 600 pixels, and it receives 50-60 frames a second. So that, every second the system receives and has to process about 80 billions of bits of information.

We arrive finely to our main question: how to transmit this huge amount of information preserving the good quality and in limited time? To achieve this goal, modern communication systems *compress* the information by making use of its *redundancy*, meaning that certain elements of most images can be safely discarded before transmission, either because these elements do not affect our human perception, or

because they can be recovered from the remaining elements. As we know, in the case of audio transmission people do not hear the frequencies beyond the interval 12 Hz - 20000 Hz, and in the case of visual transmission people cannot distinguish (resolve) details, which are too close to each other.

What is more, the neighboring pixels are usually *correlated*, that is, both their luminance and chrominance have close values, and can be satisfactory predicted if we know nearby values. It should be remarked here, that in most cases grayness does not change abruptly, there is significant correlation between the shades of neighboring pixels. The procedure of removing the redundancy of an image is called *de-correlation*. The Discrete Cosine Transform (DCT) is widely used, because it can satisfactory de-correlate images.

5. Discrete Cosine Transform

5.1. **Technological Intermezzo.** A few mathematical topics, we need in our discussion, have been already introduced, and now we show how they work together in the main topic of this module: reliable and fast transmission of very large amounts of information, either with no distortion at all, this is called the *lossless* transmission, or with an acceptable distortion, which is called the *lossy* transmission. From now on, we concentrate on the transmission of visual information. But before talking mathematical issues, very few words are in order about technology.

Say, you want to send out an image, for instance, a picture taken by your phone. The software divides the entire image into squares of size 8×8 pixels, called *data units* or *macro-blocks*; this is the most common current format. For every macro-block, the electronics produces an 8×8 matrix of luminance values for each of its 64 pixels, with the entries being the whole numbers from 0 to 255 inclusive. Since $256 = 2^8$, even in the case of monochromatic image, every pixel generates 8 bits = 1 byte of information, thus, one macro-block produces 64 bytes, and every instance the whole screen sends millions and millions of bytes of information.

However, since the system must transmit the information from all the macro-blocks simultaneously, we often do not have enough resources, say memory, to have both the high quality and high speed of transmission. The size of a macro-block, 8×8, was chosen after experiments as a compromise, the trade-off, which is always present in applications. On the one hand, it allows to essentially compress the information to be transmitted, and on the other hand, it preserves the quality of the

images transmitted, so that our eyes cannot notice that the picture on screen is slightly differs from the original one. Of course, it is possible to increase the transmission speed by sacrificing the quality of the image, or vice versa, to improve the quality by reducing the speed.

5.2. **DCT.** A mathematical tool for compressing the information to be transmitted was developed by Ahmed, Natarajan and Rao in 1974 [1], it is called *Discrete Cosine Transform* (DCT). DCT is used for compressing information by many current standards, including JPEG, MPEG, GIF, etc., it compresses the data from each macro-block into a smaller array of numbers, carrying essentially the same important information. DCT is a matrix transformation. It takes on a matrix of intensities, described in Sect. 4.1, and transforms it *before transmitting* into another matrix. The system then simplifies the new matrix by replacing some its entries with zeros, and transmits only this compressed matrix. At the receiving end, the system applies the *inverse* DCT, producing an image we see on the screen.

Matrix transformations, like DCT, are realized by multiplying the given matrix on both sides by some known universal matrices, which we do not have to transmit.

Before defining DCT and studying its properties, we consider the following extremely simplified example, the 8×8 integer-valued matrix of intensities of some model block,

$$X = \begin{pmatrix} 25 & 25 & 25 & 25 & 25 & 25 & 25 & 25 \\ 25 & 25 & 25 & 25 & 25 & 25 & 25 & 25 \\ 25 & 25 & 25 & 25 & 25 & 25 & 25 & 25 \\ 25 & 25 & 25 & 25 & 25 & 25 & 25 & 25 \\ 25 & 25 & 25 & 25 & 25 & 25 & 25 & 25 \\ 25 & 25 & 25 & 25 & 25 & 25 & 25 & 25 \\ 25 & 25 & 25 & 25 & 25 & 25 & 25 & 25 \\ 25 & 25 & 25 & 25 & 25 & 25 & 25 & 25 \end{pmatrix}.$$

The entries of this matrix express energetic content of the 64 pixels, comprising a model macro-block, such that all of its pixels have the same luminance. To see why DCT is useful, we apply the DCT of size 8 to this matrix, first without any explanation, which will follow. The DCT transform of this matrix is

$$DCT(X) = \begin{pmatrix} 200 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We know that any entry of the original matrix represents, in certain units, the amount of energy, received by the pixel corresponding to this entry. Moreover, the *total energy* of a micro-block is represented by the *sum of squares* of the entries of the intensity matrix. Now we observe that except for the just one entry on the top-left of the matrix, all the other elements of the transformed matrix are zeros. It is not by chance. This property of DCT is called the *energy compaction* or energy concentration. Since the entries of this matrix represent the energy contents of the pixels of the data unit, we see that in this example DCT concentrated the energy of the whole block into just one entry.

Certainly, in real examples we cannot expect that all but one elements vanish, but in practice most of the elements of this matrix carry little energy; in other words, they carry very little information. Hence, they can be neglected, *replaced with zeros* without significantly distorting the result.

The reason for that is, in particular, the redundancy of real images - neighboring pixels are often strongly *correlated*: if we know the intensity of a pixel, we can predict, with the good probability, the intensity of its neighbors. DCT reduces this redundancy, it *de-correlates* the image.

The DCT matrix is *orthogonal*, it satisfies the equation, called the Parseval relation. Namely, it preserves the sum of squares of the entries of a matrix subject to an orthogonal transformation; this sum of squares represents the energy (or the *variance*) of the system before and after transformation. For instance, in the simple example above, the sums of the squares of the entries of the given matrix X and the transformed matrix $DCT(X)$ both are equal $25^2 \times 64 = 200^2$. Hence, DCT *redistributes* the energy between different frequencies without any loss, compacting the energy closer to the low frequencies.

To compress the image even more, we can notice, that our vision is more sensitive to changes in lower frequencies, thus, the transformed signal emphasizes more important information, while giving less stress

to less important, lower frequency details.

To see intuitively, how DCT concentrates the energy, let us look at the following Fig. 12. This is a modified Fig. 1, where the new coordinate axes $\mathcal{X}' - \mathcal{Y}'$ are the standard axes, rotated in positive direction, counter-clock wise.

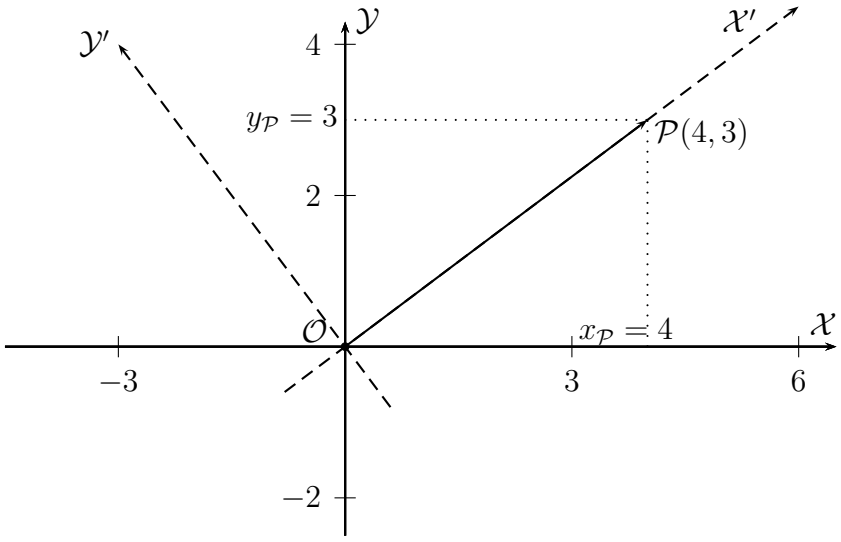

FIGURE 12. Geometric vectors and their projections in $\mathbb{R}^2$. The dashed lines make the rotated coordinate system $\mathcal{X}' - \mathcal{Y}'$.

In the standard coordinates, the point $\mathcal{P}$ has coordinates $(4, 3)$ comparable in magnitude with its distance to the origin, which is 5. In the rotated coordinates $\mathcal{X}' - \mathcal{Y}'$, though, the $\mathcal{Y}'$-coordinate vanishes. This is the fundamental property of the DCT – it treats the original $8 \times 8-$matrix as a vector in the $64-$dimensional space (Recall Problem 9) and rotates the coordinate axes so that the most of components of the vector become small. Later on, we discuss how the DCT computes the basis vectors of the rotated axes. Whence, the DCT solves one of the basic problems of the Linear Algebra, it computes the transfer matrix from one basis to another one.

In a sense, the energy concentration feature of DCT follows from an elementary property of the $\cos-$function: for $0 \leq x \leq \pi/2$ it is monotonically decreasing from $1 = \cos(0)$ to $0 = \cos(\pi/2)$.

Another elementary property of the cosine function, which makes it so useful in signal processing, is its*evenness*, that is, $\cos(-x) = \cos x$

for all real (and also complex) x. Indeed, after computing the DCT of a signal in the frequency domain, we want to return back to the representation of the signal as a function of time. One of the fastest ways to numerically invert the DCT is by making use of the well-developed algorithms of the Fast Fourier Transform (FFT). Since the function $\cos x$ is even, this can be easily done by extending the DCT from an interval $(0, a)$ onto the symmetric interval $(-a, 0)$. There may be other considerations, leading to symmetry with respect to another vertical line, say $x = -1/2$, instead of $x = 0$.

The squares are 2-dimensional images, so that in practice we have to deal with 2-dimensional transformations. However, the squares are *direct (Cartesian) products* of two *linear* segments, therefore, the 2-dimensional DCT is *separable*, it reduces to two consecutive 1-dimensional DCTs. To compute the DCT of a matrix, we can first find the 1-dimensional DCT of every row of the matrix, and then compute the 1-dimensional DCT of columns of the resulting matrix. Because of that, we study now in more detail the mathematics of the 1-dimensional DCT.

5.3. **One-Dimensional DCT.** The 1-dimensional DCT applies to an $8-$vector and recomputes it into another $8-$vector. As a matter of fact, there are 8 discrete cosine transforms, DCT-I through DCT-VIII, having slightly different properties [8]. The transform we are dealing with, is DCT-II introduced in [1]. It is often called *the DCT*, because it is, probably, the most widely used version of DCT.

We now concentrate on our major topic – the mathematics of the DCT. So that, we can forget about pixels, luminance, etc., and work with a given $8-$vector, whose eight elements are whole numbers between 0 and 255 inclusive. If the picture is polychromatic, there are three matrices, the separate intensities (chrominance) of **R**ed, **G**reen, and **B**lue colors.

We define the transformation of $8-$dimensional signals, that is, the DCT of length 8 only, but the construction can be straightforwardly extended to any natural number $n \geq 2$ [7]. By tradition, the indices go from 0 through 7, rather than from 1 through 8.

In the following formulas $\mathbf{A}^T$ stands for the *transposed* vector (or the transposed matrix) of a vector (a matrix) $\mathbf{A}$; for a row $n-$vector

$$\mathbf{A} = (a_1, a_2, \ldots, a_n),$$

its transpose $\mathbf{A}^T$ is a column $n-$vector

$$\mathbf{A}^T = \begin{pmatrix} a_1 \\ a_2 \\ ... \\ a_n \end{pmatrix},$$

and vice versa.

The "T" in DCT stands for "Transform", DCT is a transformation of real, that is with real components, n vectors into real n vectors as follows. If $n = 8$, the definition is as follows.

Definition 3. *The DCT of a real-valued column vector of length* 8*,*

$$\mathbf{U} = (u_0, u_1, \ldots, u_7)^T,$$

is the real-valued column vector, also of length 8,

$$\mathbf{V} = DCT(\mathbf{U}) = (v_0, \ldots, v_7)^T, \tag{12}$$

with the components

$$v_0 = \frac{1}{2\sqrt{2}} \sum_{i=0}^{7} u_i$$

and for $k = 1, 2, \ldots, 7$,

$$v_k = \frac{1}{2} \sum_{i=0}^{7} u_i \cos\left(\frac{k(i+1/2)}{8}\pi\right) = \frac{1}{2} \sum_{i=0}^{7} u_i C_k(i),$$

where the weights C_k, defined by equation (6?) in Sect. 2.6, are evaluated at point i. The numbers v_k are called the *DCT transform coefficients*.

Thus, the DCT $\mathbf{V} = \mathbf{V}(\mathbf{U})$ of an $8-$vector $\mathbf{U}$ is a weighted linear combination of cosine-functions having different frequencies, with the coefficients u_i being the amplitudes of the component signals with those frequencies. Another way, the components of image-vector $\mathbf{V}$ are "trigonometrically weighted" components of the data vector $\mathbf{U}$. The component v_0 is called the DC average of the signal, the other components are called AC [7] components.

[7] After Direct Current and Alternating Current.

The definition in 8-dimensional case can be conveniently written down by making use of the matrix

$$\mathbf{C} = \frac{1}{2}\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdots & \frac{1}{\sqrt{2}} \\ \cos\frac{\pi}{16} & \cos\frac{3\pi}{16} & \cdots & \cos\frac{15\pi}{16} \\ \cos\frac{2\pi}{16} & \cos\frac{6\pi}{16} & \cdots & \cos\frac{30\pi}{16} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cos\frac{7\pi}{16} & \cos\frac{21\pi}{16} & \cdots & \cos\frac{105\pi}{16} \end{pmatrix},$$

whose entries are the trigonometric functions $C_k(t)$, sampled at the points $t_i = (i + 1/2)\pi$. Now the image-vector $\mathbf{V}$ (or the image-matrix if we consider all the 8 column-vectors for a macro-block) is computed as the matrix product

$$\mathbf{V}^T = \mathbf{C} \cdot \mathbf{U}^T. \tag{13}$$

Definition 4. *The matrix* $\mathbf{M}$ *is called* orthogonal *if it satisfies the equation*

$$\mathbf{M} \cdot \mathbf{M}^T = \mathbf{I},$$

$\mathbf{I}$ *being the identity matrix.*

This equation immediately shows that an orthogonal matrix is non-singular; moreover, its determinant is ± 1.

Problem 32. *Prove that the definition of an orthogonal matrix can be stated as* $\mathbf{M}^{-1} = \mathbf{M}^T$.

Problem 33. *Prove by a direct computation that the DCT matrix* $\mathbf{C}$ *above is orthogonal. A few known trigonometric identities can be of use, for instance, for* $k = 1, 2, \ldots,$

$$\cos\frac{k\pi}{16} + \cos\frac{3k\pi}{16} + \cos\frac{5k\pi}{16} + \cdots + \cos\frac{15k\pi}{16} = 0. \tag{14}$$

Problem 34. *Prove (14) and extend it for any* $n = 2, 3, \ldots.$

Problem 35. *Prove that a transformation given by the orthogonal matrix satisfies Parseval's relation, that is, the sum of the squares of the entries of the original matrix and the transformed matrix is the same.*

Let us consider examples.

Example 1. *Let be*

$$\mathbf{U} = (\alpha, \alpha, \alpha, \alpha, \alpha, \alpha, \alpha, \alpha),$$

where α is a real constant. By Definition 3, and due to Problem 33, we have

$$\mathbf{V}^T = \mathbf{C} \times \mathbf{U}^T = \alpha \left(2\sqrt{2}, 0, 0, 0, 0, 0, 0, 0\right)^T. \tag{15}$$

Example 2. *Let*

$$\mathbf{U} = (1, 0, 1, 0, 1, 0, 1, 0).$$

Now we have

$$\mathbf{C}\times\mathbf{U}^T = \frac{1}{2}\begin{pmatrix} 2\sqrt{2} \\ \cos(\pi/16) + \cos(5\pi/16) + \cos(9\pi/16) + \cos(13\pi/16) \\ 0 \\ \cos(3\pi/16) + \cos(15\pi/16) + \cos(27\pi/16) + \cos(39\pi/16) \\ 0 \\ \cos(5\pi/16) + \cos(25\pi/16) + \cos(45\pi/16) + \cos(65\pi/16) \\ 0 \\ \cos(7\pi/16) + \cos(35\pi/16) + \cos(63\pi/16) + \cos(91\pi/16) \end{pmatrix},$$

and rounding to the nearest thousandth,

$$\mathbf{C} \times \mathbf{U}^T = \left(\sqrt{2},\ 0.255,\ 0; 0.301,\ 0,\ 0.450,\ 0,\ 1.281\right)^T. \tag{16}$$

Example 3. *Let*

$$\mathbf{U} = (0, 1, 0, 1, 0, 1, 0, 1).$$

In this case

$$\mathbf{V} = \mathbf{C}\times\mathbf{U}^T = \frac{\alpha}{2}\begin{pmatrix} 2\sqrt{2} \\ \cos(3\pi/16) + \cos(7\pi/16) + \cos(11\pi/16) + \cos(15\pi/16) \\ 0 \\ \cos(9\pi/16) + \cos(21\pi/16) + \cos(33\pi/16) + \cos(45\pi/16) \\ 0 \\ \cos(15\pi/16) + \cos(35\pi/16) + \cos(55\pi/16) + \cos(75\pi/16) \\ 0 \\ \cos(21\pi/16) + \cos(49\pi/16) + \cos(77\pi/16) + \cos(105\pi/16) \end{pmatrix}$$

$$= \left(\sqrt{2}, -0.255, 0, -0.301, 0, -0.450, 0, -1.281\right)^T. \tag{17}$$

The data in these three examples are strongly correlated, that is, they follow a certain pattern; if you know only two elements of the given vector, you can precisely restore the whole vector. Thus, the results show good energy compaction. Now we consider an example with random, uncorrelated data.

Example 4. *Let* $\mathbf{U} = (1, 2, 15, 0, 5, 6, 7, 255)$. *Multiplying the matrices and rounding off the results to the integers; we compute*

$$v_0 = \frac{1}{\sqrt{8}} \sum_{k=0}^{7} u_k \approx 151,$$

$$v_1 = (1/2) \sum_{k=0}^{7} u_k \cos\left(\frac{1(k+1/2)}{8}\pi\right) \approx -87$$

$$v_2 = (1/2) \sum_{k=0}^{7} u_k \cos\left(\frac{2(k+1/2)}{8}\pi\right) \approx 88$$

$$v_3 = (1/2) \sum_{k=0}^{7} u_k \cos\left(\frac{3(k+1/2)}{8}\pi\right) \approx -174$$

$$v_4 = (1/2) \sum_{k=0}^{7} u_k \cos\left(\frac{4(k+1/2)}{8}\pi\right) \approx 34$$

$$v_5 = (1/2) \sum_{k=0}^{7} u_k \cos\left(\frac{5(k+1/2)}{8}\pi\right) \approx -56$$

$$v_6 = (1/2) \sum_{k=0}^{7} u_k \cos\left(\frac{6(k+1/2)}{8}\pi\right) \approx 116$$

$$v_7 = (1/2) \sum_{k=0}^{7} u_k \cos\left(\frac{7(k+1/2)}{8}\pi\right) \approx 39.$$

The computations are elementary though cumbersome, and intended only as a demonstration of the method. There are powerful software packages computing DCTs, e.g., in *MatLab* and in *Mathematica*; see also [3].

Problem 36. *Find the DCT of the vectors* $\mathbf{U} = (1, 1, 1, 1, 0, 0, 0, 0)$ *and* $\overline{\mathbf{U}} = (0, 0, 0, 0, 1, 1, 1, 1)$.

Let us compare the first and the fourth examples. In the first example, the components are equal, that is, they are again highly correlated, they can be predicted if we know only one of them, and the energy compaction is perfect. In the last example, the components of the vector are random, they are not correlated, thus the redistribution of energy is not obvious, but even in this example the sum of squares of the first four frequencies contains 78% of the sum of all the eight frequencies. These examples show again why the DCT is so useful and widely used – this transformation essentially decreases the redundancy of the image

and allows us to transmit much less digits, while preserving the good quality of the image.

Problem 37. *Compare the results in the Examples 1, 2, and 3; what is the relationship between equations (15), (16), and (17)? What about the results of Problem 36?*

The sum of the DCT vectors (16) and (17) in Examples 2 and 3 is the vector (15) of Example 1. This is not a coincidence, since the given vectors in these examples satisfy the same relation, and as it follows from equation (13), DCT is a *linear operator*, that is, for any vectors $\mathbf{U}', \mathbf{U}''$ and any constants α, β,

$$DCT(\alpha\mathbf{U}' + \beta\mathbf{U}'') = \alpha DCT(\mathbf{U}') + \beta DCT(\mathbf{U}'').$$

5.4. **Trigonometric Interpolation.** There is very visible geometry behind the ability of the DCT to compact energy. Since we want to use trigonometric functions, it is natural to consider the segment $[0, \pi]$ as the domain of the functions, and to divide this interval into 8 equal parts, because we consider the 8-vectors,

$$\left[0, \frac{\pi}{8}\right], \ \left[\frac{\pi}{8}, \frac{2\pi}{8}\right], \ \ldots, \left[\frac{7\pi}{8}, \pi\right].$$

Given an 8-vector of the luminance values, $(u_o, u_1, \ldots, u_7)$, we interpolate them by a trigonometric polynomial. The nodes of interpolation are still free, and different choices lead to different DCTs. In particular, if we choose the middle points of these intervals,

$$\frac{\pi}{16}, \frac{3\pi}{16}, \ldots, \frac{15\pi}{16},$$

we recognize here the arguments of the basis DCT vectors $\cos\left(\frac{2k+1}{16}\pi\right)$, which form the DCT-matrix $\mathbf{C}$, see p. 35. Thus, we arrive at *the* DCT, that is, DCT-II.

Define now the trigonometric functions

$$P_k(t) = \cos\left(\frac{k(t+1/2)}{8}\pi\right), \ k = 0, 1, \ldots, 7,$$

which are trigonometric polynomials up to the stretching/compressing and shifts in the arguments.

Every component v_i of the DCT-vector $\mathbf{V}$ in Eqn. (12) is a sample of the trigonometric polynomial of degree 8. Indeed, for $i, k = 0, 1, \ldots, 7$,

$$\cos\left(\frac{k(i+1/2)}{8}\pi\right) = \cos\left(\frac{k(t+1/2)}{8}\pi\right) \text{ if } t = i.$$

Due to the shift $t + 1/2$ in the argument, the trigonometric functions $P_k(t)$ are not even, rather they are symmetric with respect to the vertical line $t = -1/2$. We notice also that $P_k(-1/2) = \cos 0 = 1$. It turns out that the linear combination of the *fundamental* interpolating polynomials $P_k(t)$ is a polynomial, which interpolates the DCT-images v_i of the points u_i. In the following straightforward proof we follow [6, p. 287].

Theorem 3. *If the vector* $\mathbf{V} = (v_0, \dots, v_7)^T$ *is the DCT-transform of the vector* $\mathbf{U} = (u_0, u_1, \dots, u_7)^T$*, that is,* $\mathbf{V} = \mathbf{C} \cdot \mathbf{U}$*, where the* $\mathbf{C}$ *is the orthogonal matrix defined above, then the trigonometric function*

$$P(t) = \frac{1}{\sqrt{8}} u_0 + \frac{1}{2} \sum_{i=1}^{7} u_i \cos\left(\frac{i(2t+1)}{16}\pi\right)$$

satisfies $P(k) = v_k$*,* $k = 0, 1, \dots, 7$*. Therefore,* $P(t)$ *interpolates the data* $\{(0, u_0), (1, u_1), \dots, (7, u_7)\}$*.*

To prove the theorem, it is enough to rewrite the equations above as vector equations and use the orthogonality of the matrix $\mathbf{C}$.

Problem 38. *Write down the details of the proof.*

5.5. **IDCT and Quantization.** In applications, the integral transforms, including the DCT, are auxiliary tools. After computing the DCT, we want to return back to the original image, maybe with certain acceptable losses, so that, we need to have an *inverse transform*. In our problem it is easy, since the DCT matrix $\mathbf{C}$ is orthogonal, therefore, it is non-singular and can be inverted. Because of that, the Inverse DCT (IDCT, for short) exists and can easily be shown to coincide with DCT III. To compute the IDCT explicitly, we multiply the basic equation $\mathbf{V}^T = \mathbf{C} \cdot \mathbf{U}^T$ by $\mathbf{C}^{-1}$ on the left, resulting in

$$\mathbf{U}^T = \mathbf{C}^{-1} \cdot \mathbf{V}^T.$$

In coordinates, the IDCT is

$$u_i = \frac{1}{2} \sum_{k=0}^{7} C_k v_k \cos\left(\frac{(i + 1/2)k\pi}{8}\right), \; i = 0, 1, 2, \dots, 7,$$

where

$$C_0 = 1/\sqrt{2} \text{ and } C_1 = \cdots = C_7 = 1. \tag{18}$$

Problem 39. *Apply the latter formulas to the examples 1-4 above; for computations, set* $\alpha = 1$*. starting with the computed values* $\mathbf{V}$*, you should get numerical results close to the given values* $\mathbf{U}$*. Keep in mind*

that due to the inevitable rounding errors, the results can be slightly different from vectors $\mathbf{U}$.

An average image consists of millions of pixels, even in the monochromatic case we attribute to every pixel 8 bits of information, so that the number of bits, we have to transmit, is huge. To use the energy compaction efficiently, after the matrix of the DCT frequency coefficients is computed, the system often makes what is called *quantization*, which is a kind of *rounding off* the integers or decimals. While doing the quantization, we change the original numbers to the integers.

The idea is a simple one. Our vision has a limited resolution. If we change a few digits, especially pertinent to high-frequency harmonics, some information is *lost*, but the image will experience a very small, negligible distortion. Between, say, 0 and 1 there are 10 001 steps of the size 0.0001. If we know from experiments or theory, that our eyes can resolve two numbers, only if the difference is at least, for eaxxmple, 0.01, we can safely shift the numbers between 0 and 0.01 to the closest of these two decimals. Thus, instead of 10 000 steps we have just 100 steps and 100 numbers to transmit.

From experiments we know that the human eyes are less sensitive to changes in chrominance than in luminance. This allows us to apply stronger compression to the chrominance matrices than to the luminance one.

In practice, for example, in JPEG or similar standards, to quantize the results the system just multiplies the matrix by a special matrix, whose entries were chosen by experiments. Thus, we have many more zeros and essentially less other digits to transmit.

For instance, consider again Example 4 and round off the DCT matrix $\mathbf{V}$ to 5 places in binary system. Since $151_{10} = 10010112_2$, we round the latter to $10011000_2 = 152_{10}$, etc. Hence, we apply the IDCT to the vector $\widehat{\mathbf{V}} = (152, -88, 88, -176, 32, -56, 120, 40)$, that is, we compute the product $\widehat{\mathbf{U}^T} = \mathbf{C}^T \cdot \widehat{\cdot}$ After rounding, the result is the vector

$$\widehat{\mathbf{U}} = (.7, .8, 154, -1, 4, 8, 7, 257).$$

Except for the outlier 154, all the other coefficients are quite close to the given vector $\mathbf{U}$.

Problem 40. *Recalculate this example preserving 6, 4, 3, 2, 1 binary digits and compare the results.*

5.6. **Two-Dimensional DCT.** The one-dimensional DCT is useful, for instance, in audio compression. However, in compression of images, which are two-dimensional, we need the two-dimensional[8] DCT. It works similarly to the one-dimensional DCT. Due to the separability, the two-dimensional DCT uses the same matrix $\mathbf{C}$ and computes the DCT of the entire data block $\mathbf{U}$ of luminance or chrominance values of pixels as the matrix product,

$$\mathbf{V} = \mathbf{C} \cdot \mathbf{U}\mathbf{C}^T.$$

If we again consider the 8×8 blocks, the entries of this product can be explicitly written as

$$v_{i,j} = \frac{1}{4} C_i C_j \sum_{l=0}^{7} \sum_{k=0}^{7} u_{k,l} \cos\left(\frac{(2l+1)j\pi}{16}\right) \cos\left(\frac{(2k+1)i\pi}{16}\right),$$

where $i, j = 0, 1, \ldots, 7$, and the coefficients C_i were defined in (18). The double sum above can be rewritten as a repeated sum, therefore, the two-dimensional DCT is *separable*, it is computed by first applying the one-dimensional DCT to every row of the data matrix, and then to every column of the new transformed matrix.

The basis vectors of the two-dimensional DCT are $8 \times 8 = 64$ products of the 8 DCT basis functions $C_k(t)$ in Fig. 8-11. If we recompute these 64 functions in terms of degrees of grayness, we get the following 64 patterns, representing 64 basis patterns (basis vectors) of the 2-dimensional DCT - see Fig. 13. The image of the whole macro-block is composed of these 64 basis images exactly as in two-pixel example in section 2.3 – see fig. 2 and 3.

The inverse two-dimensional DCT is also separable, it is given by equations

$$u_{k,l} = \frac{1}{4} \sum_{i=0}^{7} \sum_{j=0}^{7} C_i C_j v_{i,j} \cos\left(\frac{(2l+1)j\pi}{16}\right) \cos\left(\frac{(2k+1)i\pi}{16}\right)$$

for $k, l = 0, 1, \ldots, 7$.

Problem 41. *Let $X = (x_{ij})$ be a square matrix of order n with real entries and $Y = (y_{ij})$ be its 2D-DCT. Let*

$$P_n(s,t) = \frac{2}{n} \sum_{k,l=0}^{n-1} a_k a_l y_{k,l} \cos\left(\frac{k(2s+1)\pi}{2n}\right) \times \cos\left(\frac{l(2t+1)\pi}{2n}\right),$$

where $a_0 = 1/\sqrt{2}$ and $a_k = 1$ for $k \geq 1$. Prove that

$$P_n(i,j) = x_{i,j},\ 0 \leq i, j \leq n-1.$$

[8]Three-dimensional DCT was also studied in the literature.

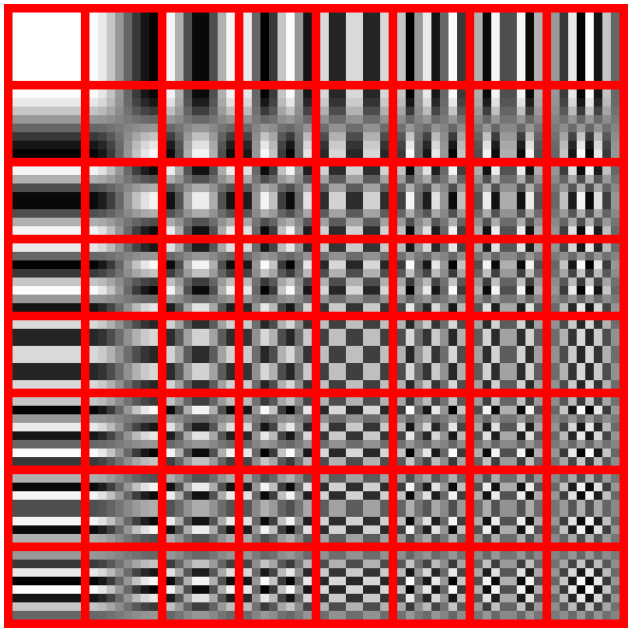

FIGURE 13. 2-dimensional DCT basis functions (basis elementary images).

Problem 42. *Finally, consider a two-dimensional example. Verify that the 2D-DCT of the* $8 \times 8-$*matrix*

$$\mathbf{U} = \begin{pmatrix} 6 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

is the following matrix [9]

$$\mathbf{V} = \begin{pmatrix} 3.0008 & 2.3261 & 1.6333 & 1.5560 & 1.5004 & 1.0398 & 0.6764 & 0.4625 \\ 2.6688 & 1.2503 & 0.5250 & 0.5710 & 0.6936 & 0.3816 & 0.2162 & 0.2486 \\ 2.2305 & 1.0387 & 0.6766 & 0.6556 & 0.6533 & 0.4381 & 0.2803 & 0.2065 \\ 1.6260 & 0.7478 & 0.8581 & 0.7485 & 0.5880 & 0.5002 & 0.3554 & 0.1482 \\ 1.0003 & 0.4484 & 0.9800 & 0.7958 & 0.5001 & 0.5318 & 0.4059 & 0.0891 \\ 0.4851 & 0.2048 & 0.9663 & 0.7502 & 0.3929 & 0.5014 & 0.4002 & 0.0407 \\ 0.1584 & 0.0549 & 0.7801 & 0.5896 & 0.2706 & 0.3941 & 0.3231 & 0.0109 \\ 0.0208 & -0.0014 & 0.4368 & 0.3254 & 0.1379 & 0.2175 & 0.1809 & -0.0003 \end{pmatrix}.$$

Here, the sum of the squares of the given matrix $\mathbf{U}$ is 54, while that sum for the transformed matrix $\mathbf{V}$, rounded to the tenth, is 53.8.

Instead of quantizing the binary version of the matrix $\mathbf{V}$, in this example we round the latter matrix off to the nearest decimal integer, which results in the following matrix $\widehat{\mathbf{V}}$ with many zeros in the right-bottom corner, $\mathbf{V}$

$$\widehat{\mathbf{V}} = \begin{pmatrix} 3 & 2 & 2 & 2 & 2 & 1 & 1 & 0 \\ 3 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Applying the IDCT to this matrix and rounding it off to the tenth, we restore the approximation $\widehat{\mathbf{U}}$ to the given matrix,

$$\widehat{\mathbf{U}} = \begin{pmatrix} 6.5 & 1.1 & -0.3 & 2.7 & 1.1 & 1.1 & 1.3 & 1.3 \\ 1.5 & 1.1 & 1.0 & .9 & 1.1 & 0.3 & -0.5 & 0.7 \\ 0.6 & 0.3 & -0.1 & -0.2 & 0.0 & 0.4 & -0.3 & 0.1 \\ 1.1 & -0.6 & -0.2 & 0.4 & -0.3 & 0.2 & 0.3 & 0.1 \\ 1.1 & -0.3 & 0.2 & 0.7 & 0.2 & 0.0 & -0.0 & 0.3 \\ 1.5 & 0.1 & -0.1 & 0.3 & -0.1 & 0.2 & -0.0 & -0.1 \\ 0.8 & -0.3 & 0.1 & 0.3 & -0.3 & -0.1 & -0.1 & -0.3 \\ 0.9 & -0.2 & -0.3 & 0.1 & 0.0 & 0.0 & -0.3 & 0.2 \end{pmatrix}.$$

Considering our very rough rounding (instead of quantization), this is a quite good approximation to the original matrix $\mathbf{U}$.

[9] To multiply the matrices in this example, we used the free online calculator at http://ww.bluebit.gr/matrix-calculator/ by BlueBit Software; accessed on 12/26/2014.

References

[1] N. Ahmed, T. Natarajan, K. R. Rao, Discrete Cosine Transform, *IEEE Trans. on Computers*, Vol. C-32, 90-93, 1974 .

[2] N. I. Akhiezer, *Theory of Approximation*. New York, F. Ungar Pub. Co. 1956.

[3] F. Chaplais, *Implementation of the DCT transform with application to the JPEG transform of test images*; April 16, 2012; http://cas.ensmp.fr/ chaplais/fr/resources/JPEGhandOn.pdf, accessed on 8.26/2014.

[4] G. Orwell, *Animal Farm*, Penguin in association with Martin Secker & Warburg Ltd, London, 2007.

[5] K. R. Rao, P. Yip, *Discrete Cosine Transform*, Acad. Press, 1990.

[6] J. Rebaza, *First Course in Applied Mathematics*, Wiley, 2012.

[7] D. Salomon, G. Motta, *Handbook of Data Compression*, 5th ed. Springer-Verlag, London Limited, 2010.

[8] G. Strang, The Discrete Cosine Transform, *SIAM Review*, Vol. 41, No. 1, pp. 135-147.